International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

INTELLIGENT HEART DISEASE PREDICTION SYSTEM WITH MONGODB

AKASH JARAD[1], ROHIT KATKAR[2], ABDUL REHAMAN SHAIKH[3], ANUP SALVE[4]

[1]Dept. of Computer Engineering, JSPM'S J.S.C.O.E, Pune, India

[2]Dept. of Computer Engineering, JSPM'S J.S.C.O.E, Pune, India

[3]Dept. of Computer Engineering, JSPM'S J.S.C.O.E, Pune, India

[4]Dept. of Computer Engineering ,JSPM'S J.S.C.O.E ,Pune, India

Abstract

The Healthcare and medical fields are rich in information but is not properly used to its potential .This results in the weak or bad decision making ability .This paper focuses on the aspect which is neglected in the process i.e. the data which is not mined. We use 14 attributes to predict the chances of heart disease prediction and thereby preventive measures can be taken to avoid it. The system is reliable, expandable and web-based and user-friendly. It also serves a purpose of training nurses and doctors newly introduced in the field related to heart disease.

Keywords :-Data Mining, K-Means, Naïve bayes, Heart disease prediction, Sensitivity and Specificity, MongoDB, filtration, NoSql.

1.INTRODUCTION

Recently WHO conducted a survey which shows, approximately 17.3 million deaths globally are due to CVD [Cardio Vascular Diseases], heart attacks and strokes. The deaths due to heart disease in countries are due to exertion, work overload, mental stress and so on. Treatment and Diagnosis is complicated and is an important task that needs to be executed accurately and efficiently. The diagnosis is often based on doctor's experience & knowledge. This leads in some cases as unwanted outcomes & excessive medical costs of treatments for patients. Therefore a medical diagnosis system is designed that takes advantage of collected data base and decision from the previous records [1].

Some hospitals have decision support systems, but they are limited. They can answer simple what–if queries like: - What is the total average of patients suffering from heart disease? Doubts which are probably lost in the data which is not mined like "Identify the important factors that will be or are the cause of Cardio Vascular Diseases?"And "For the given patient records, predict the probability of patients suffering from heart disease" are left unanswered. This system helps in diagnosing disease with less no of medical tests & more effective treatments [9].

2.HEART DISEASE

In conventional system diagnosis of Heart Disease is basically done by ECG [Electrocardiogram]. The diagnosis is by using combination of ECG and clinical symptoms. The confirmation of Heart Attack is later done which usually takes a few hours due to the rise in the level of CPK [creatinine phosphokinase]. It is released by the dying heart muscles into the blood circulation when they begin to dissolve [3].

3.DATA-SETS

The Data set used is obtained from Data mining repository of California University, Irvine (UCI). Data set from Cleveland, Hungary, Switzerland, long beach set are collected. Cleveland, Hungary, Switzerland and long beach data set contains 76 attributes totally. But we have considered only 14 attributes which are basically proven to be important. Among all those Cleveland data set is the most commonly used data set, it has less missing attributes than others which helps in better result. Figure 1 shows some sample of data set collected from the UCI repository. We store the data entry in MongoDB data base, it is more simplified to use and can easily adapt to future changes [13] [14].

Table 1: Attributes Used

No	Name	Description
1	Age	Age in Years
2	Sex	1=male, 0=female
3	Cp	Chest pain type (1 = typical angina, 2=atypical angina, 3 = non-angina pain, 4 = asymptomatic).
4	Trestbps	Resting blood sugar (in mm Hg on admission to hospital).
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1=true, 0=false).
7	Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy).
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest

11	Slope	Slope of the peak exercise ST segment (1=up sloping, 2=flat, 3= down sloping)
12	Ca	Number of major vessels colored by Fluoroscopy
13	Thal	3= normal, 6=fixed defect, 7= reversible defect
14	num	Class (0=healthy, 1=have heart disease).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0
16	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
17	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
18	48	1	2	110	229	0	0	168	0	1	3	0	7	1
19	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
20	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
21	49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
22	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
23	58	0	1	150	283	1	2	162	0	1	1	0	3	0
24	58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
25	58	1	3	132	224	0	2	173	0	3.2	1	2	7	3
26	60	1	4	130	206	0	2	132	1	2.4	2	2	7	4
27	50	0	3	120	219	0	0	158	0	1.6	2	0	3	0

Figure 1: Sample Data Set

4.DATA MINING TECHNIQUES PREFERRED

Data mining techniques such as Classification, Clustering and many more are used in extracting knowledge from database. Medical data is mined by using the techniques mentioned above and the diagnosis is carried out. Practical use of Data mining techniques in medical data is explained below:

5.Data Mining and Classification

Classification is done based on supervised machine learning Algorithm. K-nn, Decision List Algorithm, Naïve Bayes, performance is based on accuracy and the time taken to build the model [11]. Naïve bayes algorithm is commonly used and better from all since it takes only some to calculate the accuracy than other algorithm used and also it lead to lower error rates. Naïve Bayes algorithm [7] [4] gives 52.23% of accurate result. The below Table shows the performance study of the algorithm.

Table 2: Performance Results of Algorithm

Algorithm used	Accuracy	Time taken
Naïve Bayes	52.33%	609ms
Decision List	52%	719ms
KNN	45.67%	1000ms

Equations

a. K-means

Given a set of observations (**x**1, **x**2, …, $\mathbf{x_n}$), where each observation is a *d*-dimensional real vector, *k*-means clustering focuses to partition the *n* observations into *k* sets ($k \le n$) [8]

S = {*S*1, *S*2, …, S_k} so as to minimize the within-cluster sum of squares (WCSS):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

The algorithm is composed of the following steps:

1.) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2.) Assign each object to the group that has the closest centroids.

3.) When all objects have been assigned, recalculate the positions of the K centroids.

4.) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

b. Naïve Bayes

A conditional probability is of some conclusion, C, given some observation, E, where there is a dependence relationship between C and E. This probability is denoted as P(C |E)[3][4] where:

$$P(C|E) = P(E|C)P(C) / P(E)$$

Why our system uses Naive bayes algorithm ?

Naive Bayes or Bayes' Rule acts as the basis for many machine-learning and data mining methods. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data[4].

6.IMPLEMENTATION

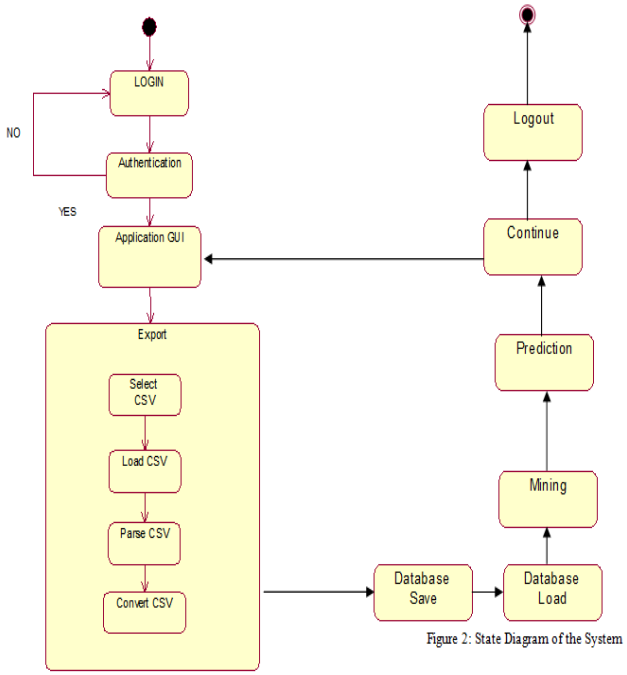


Figure 2: State Diagram of the System

An application in provided to the patient where he enters his details. Login is provided as it is unique to each patient from wherein authentication is done. After the patient is an authorized user he is given access to application GUI where he enters his symptoms. The symptoms are stored in the database and can be loaded, selected via CSV format [5] [6]. Information can be imported in the database via Excel files. Database is saved and loaded. Mining techniques are applied that is k-means and naïve bayes. K-means is applied at first. Clustering is done with respect to above parameters considering age as the primary parameter. After the age is clustered, various groups are formed. Naïve bayes is applied which gives the conditional probability of the patient who will suffer heart disease in the future with respect to rest of the parameters as declared above.

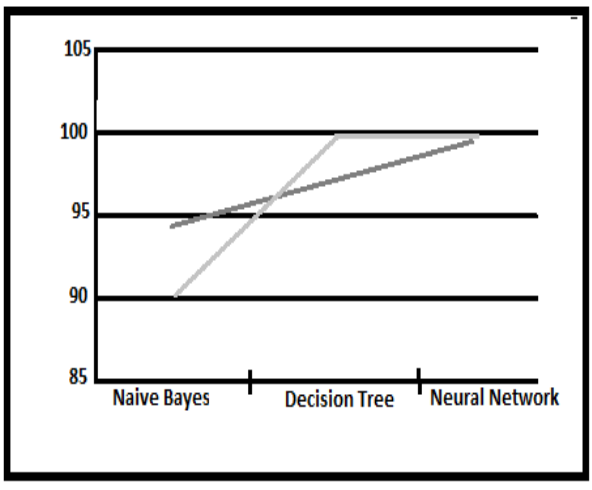


Figure 3: Graphical Accuracy of Methods [7]

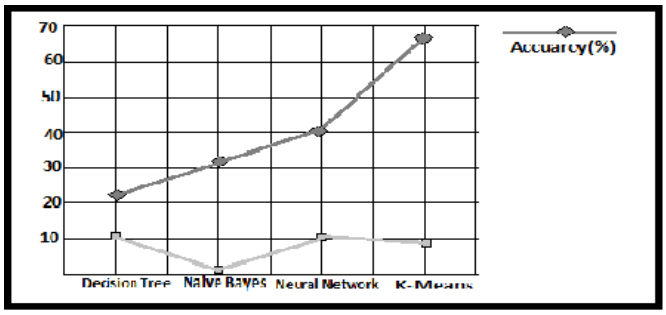


Figure 4: Graphical Accuracy of Methods [7]

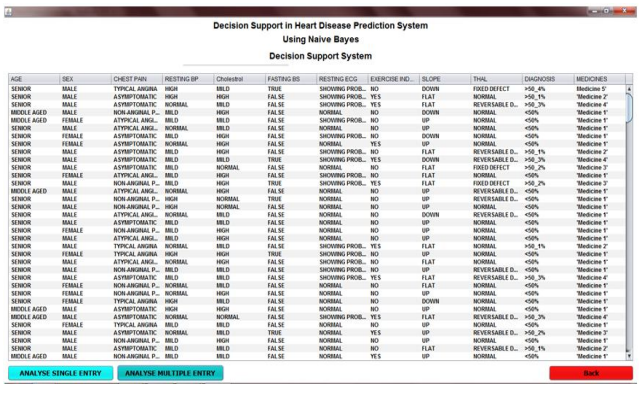
7.CONCLUSION

The overall objective of our work is to predict accurately with less number of tests and attributes the presence of heart disease. In this paper, fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span. Two data mining classification techniques were applied namely K-means & Naive Bayes. As shown above, it is clear that Naïve Bayes has better accuracy in less time than others. This system can be further expanded. Other data mining techniques can also be used for predication e.g. Neural Networks, Time series, Association rules.

FUTURE SCOPE

Our system is could be used as a training tool for Nurses and Doctors who are freshly introduced in the field related to heart diseases. The system also has a feedback from the experienced doctors who can give their views and opinions about certain medicines /practices done by the doctor on the patient. Thus, the patient can have a choice in choosing the medicines he should take in order to have a healthier life. Moreover, if implemented on a large scale it can be used in medical facilities like hospital, clinics where a patient wouldn't have to wait in long queues for treatment if he is feeling symptoms related to heart disease. This system can further be developed for other diseases also [10]. As we have used MongoDB for our backend database the feature changes can be easily made without affecting our working system [13].

Result Screenshots:-



References

[1] Liangxiao. J, Harry.Z, Zhihua.C and Jiang.S "One Dependency Augmented Naïve Bayes", ADMA,186-194, 2005.
[2] Asha Rajkumar and Mrs. Sophia Reena, " Diagnosis of Heart Disease using Data Mining Algorithms, Global Journal of Computer Science and Technology, vol 10(10), 2010, 38-43.
[3] Shadab Adam Pattekari and Asma Parveen "Prediction System For Heart Disease Using Naïve Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.
[4] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh M.Tech, Asst. Professor Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor, "Decision Support in Heart Disease Prediction System using Naïve Bayes" G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)2011.
[5] "CSV File Reading and Writing" (http:/ / docs. python. org/ library/csv. html).Retrieved July 24, 2011. "is no"CSV standard".
[6] Y. Shafranovich. "Common Format and MIME Type for Comma- Separated Values (CSV) Files" (http:/ / tools. ietf. org/ html/ rfc4180) Retrieved September 12, 2011.
[7] Sivagowry, Dr. Durairaj. M2 and Persia. "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease" 2013.
[8] Bala Sundar V, "Development of Data Clustering Algorithm for predicting Heart", IJCA, Vol 48(7), June 2012, pp 8-13.
[9] Chaltrali S. Dangare and Sulabha, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques",IJCA, Vol 47(10), pp 44-48, June 2012.
[10] Manjusha K. K, K. Sankaranarayanan, Seena P " Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification" 2014 International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 1, January 2014 ISSN: 2277 128X .
[11] K.R. Lakshmi , M.Veera Krishna and S.Prem Kumar Performance Comparisonof Data Mining Techniques for Predicting of Heart Disease Survivability International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 ISSN 2250-3153.
[12] 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.
[13] Rupali Arora, Rinkle Rani Aggarwal "Modeling and Querying Data in MongoDB "International Journal of Scientific & Engineering Research, Volume 4, Issue 7, July-2013 141 ISSN 2229-5518.
[14] White paper Big Data MongoDB vs Hadoop Big Solutions for Big Problems Written by: Deep Mistry, Open Software Integrators

AUTHORS

AKASH JARAD, ROHIT KATKAR, ABDUL REHAMAN SHAIKH, ANUP SALVE Pursuing degree course, Bachelor of Engineering in Computer Science at Dept. of Computer Engineering, JSPM'S J.S.C.O.E, Pune, India