Heart Disease Prediction using Naive Bayes Classification Technique

K.Naga Prasanthi[1], Vanama Vamsi Krishna[2], Chippada Sai Durga Prasad[3], Singu Tarun[4], ReddyVamsi Krishna[5]

[1]Sr.Asst.Professor, [2,3,4,5]Student

Dept of CSE, Laki Reddy Bali Reddy College of Engineering,Mylavaram

Abstract **-** In the modern growing field health is one of the most important factor to be considered. Heart is one of the most important thing in the human system. On seeing the day by day health conditions most of the people die by the sudden heart attack so the main aim of this project is to predict whether the person is having the heart disease or not for this we are using the data mining technique called as naive bayes algorithm. By using this algorithm we calculate the yes or no probability of heart disease of a person based on data which is given by the user and check whether the person has a chance of heart disease or not.

Keywords **-** Naive Bayes, Datamining

I. INTRODUCTION

The algorithm must focus on the user entered data. So the project may be useful to many hospitals, in order to check whether the patients have chance of heart disease or not. Based on the results they may take further steps in order to cure the disease. The system contains both user and admin modules, so the admin have chance to add his own data and predict the results accurately. The user also have flexibility to enter the present symptoms and check the present condition of the heart. At the end of the project in user module we provide a short report about the system and based on the results the admin may improve the efficiency of the algorithm by improving the better quality results.

II. LITERATURE SURVEY

The project mainly depends on the probability so to calculate the probability there are many ways in the data mining but out of those algorithms we use naive bayes algorithm in order to get better quality of results and we have done with all kinds of approaches and finally selects the bayes rules to predict the data for good quality results.

III. METHODOLOGY

In this project we are using the naive bayes algorithm to find out the probability. Firstly we calculate the total yes and no probability and then based on the user entered data we compare the selected data with the data set values and finally produce the total yes and total no probability and check which one is higher. Based on those results we finally conclude with results whether the person is having the heart disease or not.

A. Data Sources - The data is taken from the UCI machine learning repository. Cleveland data for heart disease is collected. As the data is pre processed, it can be directly utilized for data mining activity.

B. Data Pre-processing - Data pre-processing is the one of the most important step in the data mining why because the data in the real world is dirty so if we consider the dirty data then we can't produce the quality results. So after pre-processing stage only we proceed to further steps to fill data into relational data base.

C. Database - In this project we are using the oracle data base to be set relational data tables after completion of data pre-processing we import the data from the excel sheet to the relational data tables.

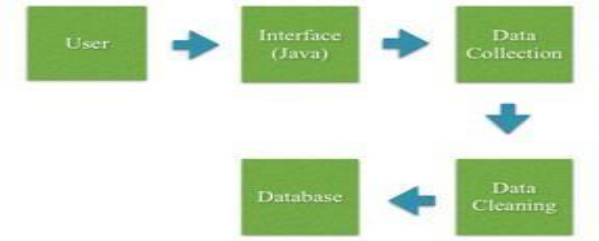

Figure 1: Data collection process

D. Naive Bayes - Bayes theorem is completely dependent on the dependency chances. It finds the probability of events occuring by knowing the probability of another event that might occur.

$$P(A/B) = (P(B/A)P(A))/P(B).$$

1. Basically we are trying to find probability of A,by given that event B is always true. Here event B is called as evidence.
2. P(A) is the Prior of A and P(B) is the evidence
3. P(A/B) is the posteriori probability of B.
   - P (A | B), the posterior, is the degree of belief having accounted for B.
   - P (B | A) / P(B) represents the support B provides for A [1].

E. Algorithm - The Naive Thomas Bayes rule relies on Bayesian theorem as given as follows:

1. Every information sample is diagrammatical by associate n dimensional feature vector, X = (x1, x2.. xn), portraying n measurements created on the sample from n attributes, severally A1, A2, An.
2. Suppose that there are a unit m categories, C1, C2,..Cm. Given associate unknown information sample, X (i.e., having no category label), the category unknown, classifier can predict that X belongs to the class having

the very best posterior chance, conditioned if and solely if: P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= I therefore we tend to maximize P(Ci|X). The category Ci that P(Ci|X) is maximized is called the most posteriori hypothesis. By Thomas Bayes theorem.

3. As P(X) is constant for all categories, solely P(X|Ci)P(Ci) want be maximized. If the category previous chances don't look like to be proverbial, then it's ordinarily considered that the categories area unit equally possible, i.e. P(C1) = P(C2) = ..= P(Cm),and we would thus maximize P(X|Ci). Otherwise, we tend to maximize P(X|Ci)P(Ci).

Note that the category previous chances could also be calculated by P(Ci) = Si/s ,where Si is that the range of coaching samples of sophistication Ci, and s is that the total range of coaching samples, On X. That is, the naive chance assigns associate unknown sample X to the category Ci.

IV. ATTRIBUTES

This database contains 13 attributes (which have been extracted from

A. Attribute Information -

1. Age
2. Gender
3. Blood Pressure
4. Chest Pain
5. Cholesterol
6. Blood sugar
7. Rest ECG
8. Angina
9. Max Heart Rate
10. Thal
11. Weight
12. Weight

V. SYSTEM ARCHITECTURE

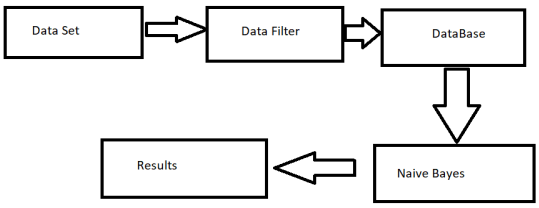

Figure 2; System Architecture

VI. PROPOSED SYSTEM

In this paper we are using the data mining technique namely –Naïve Bayes. We are using Naïve Bayes as a new method for heart disease prediction. By comparing it we can show that our proposed system is better than the other two techniques .The advantages of using Naïve Bayes method are given as follows: The Proposed system consists of the following modules:

Module-1: Registration and Data collection - In this project there are two modules one is user and another is admin module. We provide the login for the admin separately and user has a flexibility to register first and then login with his own credentials.

Module-2: Finding the Probability - After the collection of the data entered by the patient, the proposed system will generate the probability whether the patient is in risk of a heart disease or not.

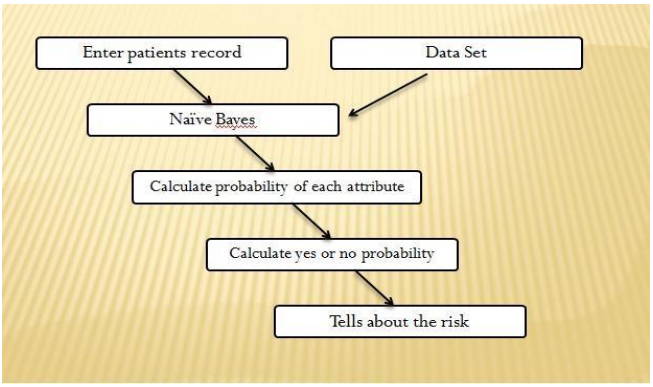

Figure 3: Working Algorithm

Module -3: Review and Analysis - Based on the review given by the user there may be chance of improving the algorithm and better results in future.

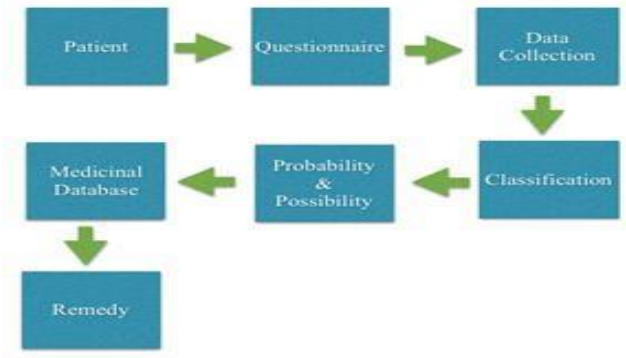

Figure 4: Flow of Algorithm

Overview of Project:

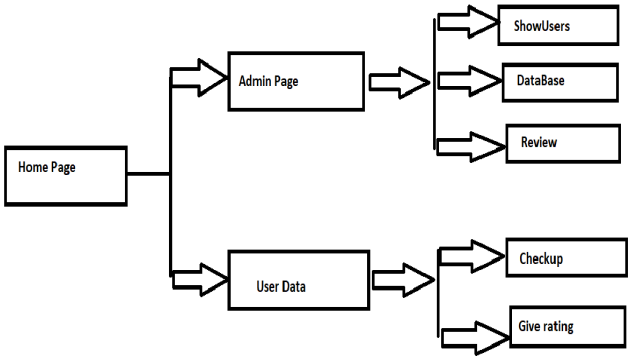

Figure 5: Overall Process

VII. IMPLEMENTATION

Module-1: Data Collection and Registration - Data Collection and Registration consists of registration page where the patient's data is collected and stored in the database.

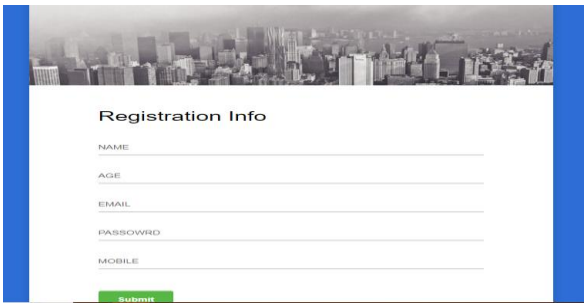

Figure 6:Registration page

Module-2: Dataset - The data set contain the values which are collected from the cleave land data sources.

DATA SET

AGE	GENDER	BP	PAIN	CHOLESTREOL	SUGAR	RESTECG	ANGINA	HEARTRATE	THAL
67	0	115	3	564	0	2	0	160	1
57	1	124	2	261	0	0	0	141	0
64	1	128	4	263	0	0	1	105	0
74	0	120	2	269	0	2	1	121	0
65	1	120	4	177	0	0	0	140	0
56	1	130	3	256	1	2	1	142	0
59	1	110	4	239	0	2	1	142	1
60	1	140	4	293	0	2	0	170	1
63	0	150	4	407	0	2	0	154	4
59	1	135	4	234	0	0	0	161	0

Figure 7: Dataset Details

In the above figure 5.2, the database details have been specified. Different attributes, their corresponding values etc have been specified.

Module-3: Probability Finder - In the Third module, i.e the probability finder, the patient will have to enter the values of the attributes which are specified in the form. The algorithm will accept the data from the patient; it will check in the database and then show the result.

Figure 8: User data entry page

Modules of User and Admin - In these modules the screens shows the user tasks.

Figure 9: User Modules page

Module-4: Review - In this module the user has a flexibility to submit the feedback of the system so that that may be useful to the admin to improve the quality of the results.

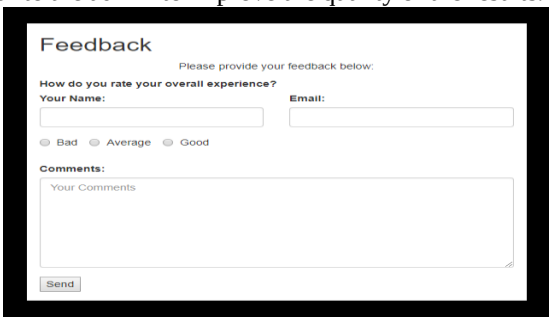

Figure 10: Feedback page

VIII. TIME COMPLEXITY

Analysis - On seeing the results of the project the analysis report as follows. The Project may predict the positive results nearly 80% and the following results are brief summary of reports.

	Total	Correct	Wrong
Yes	14	12	2
No	6	3	3
Overal	20	15/20	5/20

Figure 11: Analysis

IX. CONCLUSION

Finally the conclusion of this project is to calculate the chance of heart disease. The classifier proposed here works with 80% accuracy. We can just predict the chance of heart disease occurring. By enhancing the algorithm we can further improve by making an android application so that it may be more flexible to the user usage.

X. REFERENCES

[1]. Ankita dewan, Meghna sharma ,"Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classi_cation"(2014) 15, 13-24.

[2]. Hlaudi Masethe, Mosima Masethe, "Prediction of Heart Disease using Classi_cation Algorithms" (2009),11994 12000.

[3]. K. Thenmozhi, , P.Deepika, "Heart Disease Prediction Using Classi_cation with Di_erent Decision Tree Techniques"(2011), 2227-2235.

[4]. M.A.Nishara Banu1 , B Gomathy, "Disease Predicting System Using Data Mining Techniques" (2007) 177, 3799-3821.

[5]. Ozgur Depren, Murat Topallar, Emin Anarim, M. Kemal Ciliz, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks" (2005) ,713-722.