INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Health prediction system using Data Mining

Omkar Nevase
developer9216@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Prafull Bansode
prafullbansode3@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Rishikesh Nimbalkar
rishikesh.nimbalkar12@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Sanket Yeginwar
sankety24@gmail.com
Sinhgad College of Engineering, Pune, Maharashtra

Suvarna Pawar
sspawar.scoe@sinhgad.edu
Sinhgad College of Engineering, Pune, Maharashtra

ABSTRACT

"Health Prediction" system based on predictive modeling, which predicts the disease (probability of disease) of the user on the basis of the symptoms that the user provides as an input to the system. The system analyses the symptoms provided by the user as input and gives the probability of the disease as an output to the user. Disease Prediction is done by implementing the Naïve Bayes Classifier. Naïve Bayes Classifier calculates the probability of the disease. Therefore, the average accuracy of 75% is obtained for disease prediction. The system also provides suggestions of Doctor's to the user based on the symptoms analyzed. Along with it also provides suggestions of nearby Doctor's available in the area of Patient. The user can also share his/her medical reports online with the Doctor, based on which doctor can provide treatment to the user. Also, there is chatting facility available through which Doctor and User can interact with each other before taking an appointment or after the treatment also. The Doctor can also manage all the patients' records online.

Keywords— *Data Mining, Prediction, Classification, Naïve Bayes*

1. INTRODUCTION

At present, when one suffers from particular disease, then the person has to visit to doctor which is time consuming and costly too. Also if the user is out of reach of doctor and hospitals it may be difficult for the user as the disease cannot be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other Heart related Disease Prediction System using data mining techniques that analyses the risk level of the patient.

Disease Predictor is a web-based application that predicts the disease of the user with respect to the symptoms given by the user. Disease Prediction system has data sets collected from different health-related sites. With the help of Disease Predictor, the user will be able to know the probability of the disease with the given symptoms. As the use of the internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to the internet than hospitals and doctors. People do not have the immediate option when they suffer from a particular disease. So, this system can be helpful to the people as they have access to the internet 24 hours.

2. LITERATURE SURVEY

K.M. Al-Aidaroos, A.A. Bakar and Z. Othman has conducted research for the best medical diagnosis mining technique. For this author compared Naïve Bayes with five other classifiers i.e. Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN) and a simple rule-based algorithm (ZeroR). For this, 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007) were selected for evaluating the performance of all algorithms. In the experiment, it was found that NB outperforms the other algorithms in 8 out of 15 data sets so it was concluded that the predictive accuracy results in Naïve Bayes are better than other techniques.

Darcy A. Davis, Nitesh V. Chawla, Nicholas Blum, Nicholas Christakis, Albert-Laszlo Barabasi have found that global treatment of chronic disease is neither time or cost efficient. So the authors conducted this research to predict future disease risk. For this CARE was used (which relies only on a patient's medical history using ICD- 9-CM codes in order to predict future diseases risks). The CARE Framework combines collaborative filtering methods/techniques with clustering to predict each patient's greatest disease risks based on their own medical history and that of similar patients. Authors have also described an Iterative version, ICARE, which incorporates ensemble concepts for improved performance. These system does not require any specialized information and provide predictions for medical conditions of all kinds in a single run. The impressive future disease coverage of ICARE represents more accurate early warnings for thousands of diseases, some even years in advance. If applied to full potential, the CARE framework can also be used to explore broader disease histories, suggest previously unconsidered concerns, and facilitating discussion about early testing and prevention and much more. [1]

Table 1: Algorithm comparison

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echoeardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	84.97	78.45	83.18	78.21	81.81	54.76
Postooerative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	8\15	5\15	0\15	2\15	1\15	1\15

M. A. Nishara Banu, B Gomathy used medical data mining techniques like association rule mining, classification, clustering I to analyze the different kinds of heart-based problems. A decision tree is made to illustrate every possible outcome of a decision. Different rules are made to get the best outcome. In this research age, sex, smoking, overweight, alcohol intake, blood sugar, heart rate, blood pressure is the parameters used for making the decisions. The risk level for different parameters is stored with their id's ranging (1-8). ID lesser than 1 of weight contains the normal level of prediction and higher ID other than 1 comprise the higher risk levels .K-means clustering technique is used to study the pattern in the dataset. The algorithm clusters information's into k groups. Each point in the dataset is assigned to the closest cluster. Each cluster centre is recomputed as the average of the points in that cluster. [2]

3. RELATED WORK

3.1 Data mining

Data mining is a technique of selecting, discovering and modelling huge amounts of data. This process has become an increasingly insidious activity in all areas of medical science research. Use of Data mining has resulted in the discovery of useful hidden patterns from huge databases. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, and soft computing and data visualization; and includes classification and regression techniques. Few of the research works are done, but all of them are focusing on a few methods of analysis, diagnosis or prediction of this disease by using different tools and techniques and most of the system focusing only on few diseases, this work is focused on the early prediction of various diseases by using Naïve Bayes.

3.2 Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier that depends on Bayes' theorem with strong i.e. naive independence assumptions. It is also be called an "independent feature model". In general terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naive Bayes classifiers are trained to work in supervised learning. Naive Bayes classifier mainly pre-assumes the effect of a variable value on a predefined class that is not dependent on the value of another variable. This is called the property of class conditional independence. Naïve Bayesian is mainly used to form models with Predictive capabilities. Naïve Bayes ha prediction rate of above 60%.

Bayes theorem provides a way of manipulative the posterior probability,

Likelihood   Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability   Predictor Prior Probability

(a) P (c|x) is the posterior probability of class (target) given predictor (attribute).
(b) P (c) is the prior probability of given class.
(c) P (x|c) is the likelihood which is the probability of predictor given class.
(d) P (x) is the prior probability of the predictor.

3.2.1 Advantages of Naïve Bayes

(a) The easy handle of a large amount of data.
(b) It mainly requires a small amount of training set
(c) to estimate the parameters i.e. mean and
(d) Variance needed for classification.
(e) Fast to train and Fast to classify
(f) Not sensitive to irrelevant features
(g) Handles real and discrete data
(h) Handles streaming data well

3.2.2 Disadvantages of Naïve Bayes

(a) Loss of accuracy
(b) variables, but these dependencies are not
(c) Handled by the classifier.
(d) Assumes independence of features

4. PROPOSED METHODOLOGY

The purpose of the Health Prediction System is to provide the online system to the users to get instant guidance on their health issues. The patient passes the Symptoms related to their disease and the System (Naïve Bayes Classifier) processes the Symptoms and displays the result in the form of the information related to the diseases associated with the Symptoms. We will have the database with a number of diseases and their symptoms used to check symptoms. Above all, we hope to provide a comfortable user experience and best Health Prediction.

Along with the predicted disease by the system, the system also provides the doctor's recommendations along with the results. In case if the system fails to predict exact disease it will show similar disease matching to symptoms along with the doctor's recommendations. Chatting facility is also available where the patients can interact with the doctors online, ask queries, and make appointments, feedbacks on treatment procedure, etc. A doctor can manage all the patient appointment online also the doctor can view/update patient medical records. The administrator of the system plays an important role, it adds doctors, verifies doctors, adds new disease to the database along with its symptoms and performs another task of administrator.

4.1 Algorithm
Following is the algorithm used for disease prediction.

Equation 1:

$$P\ (Y|X1,\ldots X_n)=\frac{P(Y)P(X1,\ldots\ldots,Xn|Y)}{P(X1,\ldots\ldots\ ,Xn)}$$

Where,
Y is a class feature
X1, X2,….., Xn are dependent features
From equation 1 we get…

Equation 2:

$$P\ (Disease|\ symptom1,\ symptom\ 2,\ \ldots\ ,symptom_n)$$
$$=\frac{P(Disease)P(symptom1,\ldots\ldots,symptomn\ |Disease)}{P(symptom1,symptom2,\ldots\ldots,Symptom\ n)}$$

Using the naïve independe nce assumption:

$$P\ (symptom1,\ldots,Symptomn\mid Disease)$$
$$=\ P\ (Symptomi\mid Disease).Where\ i=1,2,3,\ldots\ ,n$$

Equation 3:

$$P\ (Disease\mid symptom1,\ symptom\ 2,\ldots,\ symptom_n)$$
$$=\frac{P(Disease)P(Symptom\ i\mid Disease)}{P(symptom1,symptom2,\ldots,\mathrm{Symptomn})}$$

So the relation becomes…

Equation 4:

$$P\ (Disease\mid symptom1,\ symptom\ 2,\ \ldots\ symptom\ n)$$
$$=\frac{P(Disease)P(Symptom\ i\mid Disease)}{P(symptom1,symptom2,\ldots,Sympom\ n)}$$

Since P (Disease | symptom1,.., Symptom n) is constant, we can use the following classification rule:

$$P\ (Disease\ |symptom1,symptom2,\ldots\ ,symptomn)$$
$$=P\ (Disease)\ \Pi ni$$
$$=1\ P\ (Symptomi\mid Disease)$$

$$P\ (Disease\ |symptom1,symtom2,\ldots\ ,symptomn)\alpha$$
$$P\ (Disease)\ \Pi ni=1\ P\ (Symptomi\mid Disease)$$

$$Y\ =ARG\ MAX\ P\ (Disease)\ \Pi ni$$
$$=1\ P\ (Symptomi\mid Disease)$$

The value P (Symptom i | Disease) can be calculated by using multinomial Naïve Bayes.

Which is given by:

$$P\ (symptomi\mid Disease)=\frac{Nyi\ +\ \alpha}{Nyi\ +\ \alpha\ n}$$

N_{yi} = Frequency of the same disease in the dataset
N_y = Total symptoms of the particular disease
n =total symptoms in the dataset
α= 1, Known as Laplace Smoothing

The value of P (Disease) can be calculated by using the Laplace Law of Succession which is
Given by:

$$P(Disease)=\frac{\mathrm{N(Disease)}\ +\ 1}{\mathrm{N}\ +\ 2}$$

5. DESIGN OF THE MAIN FRAMEWORK

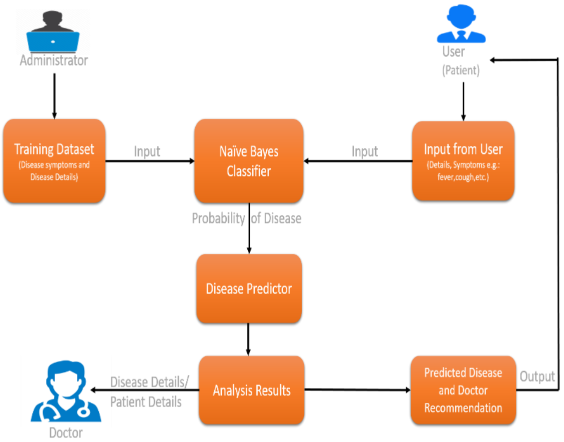


Fig. 1: System architecture

The application consists of 3 types of users.

5.1 Patient
Patients are one of the important users of the System. They give detailed information about their health issues in the form of the questions asked by the system.

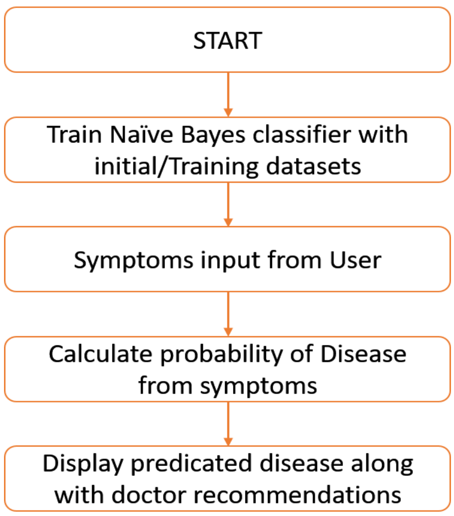
5.2 Doctor
Doctors use the system to get the patient's medical history generated by the system and to view the patient's profile.

5.3 Admin
Admin has all authorities regarding both the users such as patient and doctor. Also, Admin can add new Diseases to the System.

Here we have made use of software likes visual studio which is an integrated development environment for any programming language, .net MVC framework. The server type used is IIS server with relational database management system like MSSQL.

The tasks performed for disease prediction are as follows:

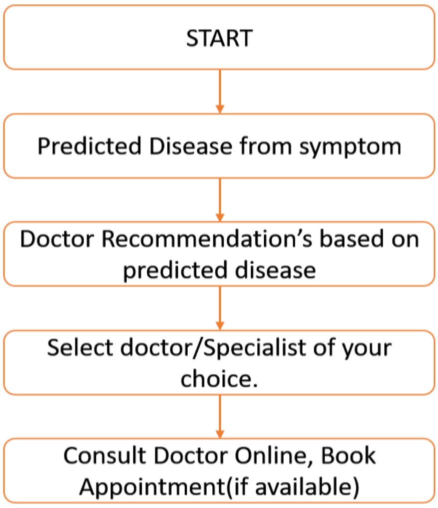
- Train Naïve Bayes Classifier with the dataset containing details of symptoms and their corresponding diseases.
- Accept Symptoms from user/patient.
- Pass the symptoms accepted from user to naïve Bayes classifier, test symptoms with trained data, and calculate Posterior probability.
- Display Predicated Disease and recommended doctors.



Doctor Recommendations once the disease is predicated:

- Display Predicated Disease from symptoms.

- Show the doctor's recommendation based on predicated disease.
- Select doctor/specialist as per your choice.
- Consult with a doctor either with online chatting option or book doctor's appointment or both.



6. RESULTS

6.1 Prediction

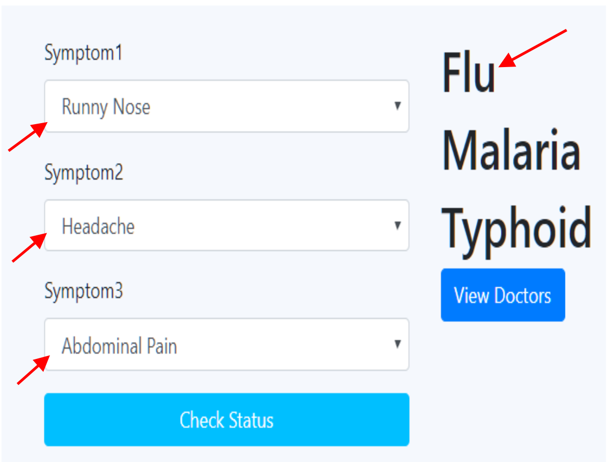


Fig. 2: Prediction

6.2 Doctor recommendation

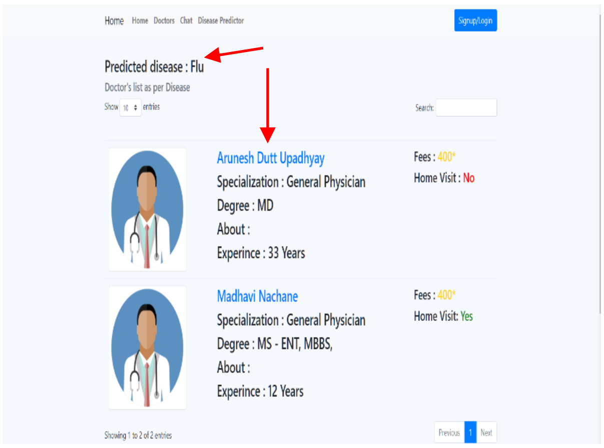


Fig. 3: Doctor Recommendation

7. FUTURE WORK

The system has a varied implication in the medical sector, it gets integrated with NLP i.e. Speech-to-Text and Text-to-Speech conversion module, it will be the great work. So that specialized users would also use this system. The system can effectively lessen human efforts of several visits of the humans in the clinic for appointments. Some queries etc. With the use of this system, user can reduce their effort.

8. CONCLUSION

The proposed system predicts the disease on the basis of the symptoms. The system is designed in such a way that it takes symptoms from the user as input and produces output i.e. predict disease. Average prediction accuracy probability of 75% is obtained. After the system predicts the disease from symptoms the system provides the list of Doctor's matching to the predicted disease and can also interact with the doctor through the application where he can clear his/her queries.

9. REFERENCES

[1] http://www.mayoclinic.org/diseases-conditions/heart-disease/basics/definition/con-20034056. http://www.idf.org/about-diabetes.

[2] Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus - A case study. ARPN Journal of Engineering and Applied Sciences, January, 10(1), 8-13.

[3] Gomathi, K. (2012). An empirical study on breast cancer using data mining techniques. *International Journal of Research in Computer Application & Management*, July, 2(7), 97-102.

[4] WEKA: Data Mining Software in Java. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/.

[5] Han J, Kamber M, Pei J. "Data mining: concepts and techniques: concepts and techniques" [M]. Elsevier,2011

[6] MSKESSON,http://www.mckesson.com/blog/changing-trends-in-mobile-health technology/(accessedon1Jun,2016).

[7] M.-J. Huang, M.-Y. Chen and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", Expert Systems with

[8] Applications, vol. 32, pp. 856-867, 2007.

[9] S. H. Ha and S. H. Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", International Journal of Computer and Information Engineering, vol. 4, no. 1, pp. 33-38, 2010.

[10] Cummings Jr D D. "All care health management system", U.S. Patent 5,301,105[P]. 1994.

[11] Samy GN, Ahmad R, Ismail Z. Security threats categories in healthcare information systems. Health Informatics Journal. 2010 Sep 1; 16(3):201–9.

[12] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics. 2008; 77(2):81–97.

