Disease Prediction with Naïve Bayes Classifier

Nitesh Kumar Verma, Shivangi Singh, Shardul Singh Chauhan, Dinesh Kumar*

KIET Ghaziabad, Uttar Pradesh, India

niteshrajm@gmail.com, singhshivangi676@gmail.com, shardulchauhan007@gmail.com, dineshvashist@gmail.com*

Received: *30.06.2020* ***Accepted:*** *18.07.2020*

ABSTRACT

The project "Disease Predictor" system which is based on predictive modelling technique predicts the disease of the user based on the symptoms that the user provides as an input to the system. The system observes the symptoms provided by the user as input and calculates the probability of the disease and as an output displays the disease with highest probability. Disease Prediction is done by implementing the Naïve Bayes Classifier. Naïve Bayes Classifier calculates the probability of the disease. The result so obtained shows an average prediction accuracy probability of about 60%.

Keywords - Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN), k-nearest neighbours (KNN), Predictive Modelling, Naïve Bayes Classifier.

1. INTRODUCTION

At present, when one suffers from a particular disease, then the person has to visit a doctor which is time consuming and costly too. Also, if the user is out of reach of doctors 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 exists other Disease Predictor Systems that makes use of data mining techniques to analyze 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. The Disease Predictor system has data sets collected from different health related sites. Disease Predictor aims to find probability of the disease with the given symptoms. This system can be helpful to the people as they have access to the internet 24 hours.

2. LITERATURE SURVEY

Here, we take a portion of the papers identified, with Disease Prediction, they are as underneath. K.M. Al-Aidaroos, A.A. Bakar and Z. Othman have led the exploration for the best clinical conclusion mining method [3]. For this creator contrasted Naïve Bayes and five different classifiers for example Calculated Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN) and a basic principle-based calculation (ZeroR). For this, 15 genuine clinical issues from the UCI AI storehouse (Asuncion and Newman, 2007) were chosen for

assessing the exhibition everything being equal. In the trial it was discovered that NB beats different calculations in 8 out of 15 informational indexes so it was presumed that the prescient precision results in Naïve Bayes is superior to different methods.

Table 1 Comparison Predictive Accuracy of Bayes and Other Technique

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
Lymphpraphy	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**

Davis et al. (2008) [1] have discovered that worldwide treatment of interminable illness is neither time or cost proficient. So, the creators directed this exploration to foresee future malady chance. For this CARE was utilized (which depends just on a patient's clinical history utilizing ICD-9-CM codes so as to foresee future sicknesses dangers). CARE joins synergistic sifting techniques with bunching to foresee every patient's most prominent ailment dangers dependent on their own clinical history and that of comparable patients. The tale frameworks require no particular data and give expectations to ailments of various types in a solitary run. The noteworthy future ailment inclusion of ICARE speaks to progressively precise early alerts for a great many infections, some even a long time ahead of time. Applied to maximum capacity, the CARE system can be utilized to investigate a more extensive malady, Disease Predictor 7 accounts, recommend already unconsidered concerns, and encourage conversation about early testing and counteraction.

A web-based application has been developed in [2] which answers certain predefined questions for disease identification. Here the data which is hidden is retrieved from a database with the help of which the entered

inputs are compared with the trained data set. This system is capable of answering more complex questions for diagnosing the heart diseases.

An exploration research paper [4] gives a review of momentum strategies of information revelation in databases utilizing information mining procedures that are being used in the present clinical examination especially in Heart Disease Prediction. Number of trials has been directed to think about the exhibition of prescient information mining procedure on the equivalent dataset and the result uncovers that Decision Tree outflanks and sometime Bayesian arrangement is having comparative exactness as of choice tree however other prescient strategies like KNN, Neural Networks, Classification dependent on grouping isn't performing admirably.

Utilizing the clinical information [5], mining methods like affiliation rule mining, grouping, bunching I to investigate the various types of heart-based issues. Choice tree is made to outline each conceivable result of a choice. Various standards are made to get the best result. In this examination age, sex, smoking, overweight, liquor admission, glucose, hear rate, pulse are the boundaries utilized for settling on the choices. Hazard levels for various boundaries are put away with their id's extending (1-8). ID lesser than 1 of weight contains the ordinary degree of expectation and higher ID other than 1 include the higher hazard levels. K-implies bunching procedure is utilized to consider the example in the dataset. The calculation bunches data into k gatherings. Each point in the dataset is doled out to the shut bunch.

3. PROBLEM STATEMENT

There are many tools related to disease predictor. But particularly heart related diseases have been analyzed and risk level is generated. But generally, there are no such tools that are used for prediction of general diseases. So, disease predictor helps in the prediction of general diseases.

4. PROPOSED SYSTEM

The proposed system acts as a decision support system and will prove to be an aid for physicians with the diagnosis. The algorithm used i.e.; Naïve Bayes produces the basis of probability of the symptoms.

The proposed system carries two basic modules:

A. Classification module

This module deals with the classification the disease according to the symptoms. The data is divided into two categories namely: training data set and testing data set. Training data set is used for training the system while testing data set is used for testing the system. The system so trained is checked for the predefined questions asked to the user. When user answers the questions, the responses so obtained are checked by the system and then the disease is classified according to the symptoms.

B. User module

This module takes input from the user in which the user has to answer certain predefined questions. Based upon the questions and answers selected by the user, the user will be informed about the disease and then the method to diagnose the disease.

5. WORKING

Disease Predictor is a web-based application that predicts the disease of the user with respect to the symptoms given by the user. The Disease Predictor 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. The user inputs all the symptoms that he has developed. On the basis of symptoms, the naive Bayes algorithm comes into effect. The disease with the most probability is shown as output.

5.1 Input

In the input design, user inputs the symptoms faced by him. The input may consist of a single symptom or multiple symptoms based on the user. After the user is done with inputting the symptoms, he may click on predict to get the most probable disease.

5.2 Output

In the output, the naive Bayes come into effect. The probability of the symptoms is calculated and the disease with highest probability is displayed.

5.3 State Diagram

The state diagram shown below describes the whole procedure the user needs to follow to use the DISEASE PREDICTOR web-based application to get the appropriate results and the work done by the application in return.

The steps are as follows:

Step 1.The user needs to open the disease predictor

Step 2.Input symptoms

Step 3.Click on the predict button

Step 4.The software then makes use of the naive Bayes algorithm to calculate the probability

Step 5.After the calculation the most probable disease is displayed

Step 6.If the user is interested to get into the details of the disease, he may get it

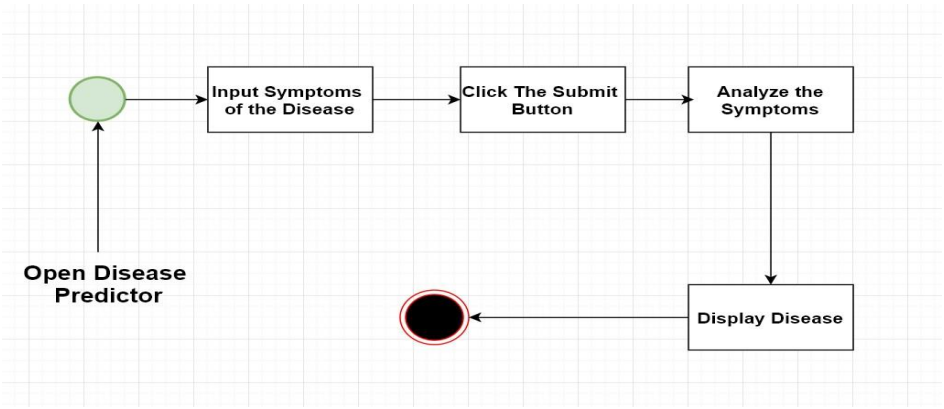

Figure 1: State Diagram

5.4 Naïve Bayes Classifier

This is a family of very simple type of probabilistic classifier which is based on application of Bayes' theorem with a strong and independent assumptions between various features.

Naïve Bayes classifier is highly scalable in nature. It requires various number of parameters that are linear in the number of variables (features/predictors) in a learning problem.

Bayes Rule: A conditional probability which leads to certain conclusion say C, given with some of the observations, O. Here a dependence relationship exists between C and O. This probability can be denoted by P(C/O):

$$P(C/O) = P(O/C).\ P(C)/P(O)\text{…….(i)}$$

Naive Bayesian Classification Algorithm: The Naive Bayesian classifier works as follows:

a) Suppose D be a training set of tuples. So, each record can be represented by an n dimensional attribute vector, Y= (y1, y2…, yn-1, yn), depicting n measurements made on the tuple from n attributes, i.e. B1 to Bn.

b) Suppose there are m number of classes that can be used for prediction, P1, P2… Pm. Given any random record, Z. Here the classifier will predict that Z belongs to the class which has the highest posterior probability, that is conditioned on Z. Therefore, Naïve Bayesian classifier predicts that the tuple Z belongs to the class Pk if and only if:

$$P\ (Pk|Z) > P\ (Pl|Z) \text{ for } 1 \leq l \leq m \text{ and } l \neq k\text{…….(ii)}$$

Thus, we maximize P(Pk|Z). The class Pk for which P(Pk|Z) is maximized is called the maximum posteriori hypothesis.

6. RESULT

We get the yield as an infection dependent on the most noteworthy likelihood according to the indications given by the client. The innocent Bayes calculation is put to use to compute the likelihood of the illnesses. Some of the results obtained so far for the above system is stated below:

Table 2 Results Obtained

Symptoms Entered	**Disease Diagnosed**	**Accuracy Rate**
Fever, fatigue, headache, pain, blood in urine	Urinary Infection	58%
Eye irritation, runny and stuffy nose, watery eyes, sneezing	Allergy	60%
Fever, headache, intense pain, fatigue, dry cough	Flu	61%
Itching, redness, tearing, burning sensation in eyes	Conjunctivitis	62%
Food intolerance, Menstrual cramps, stress, bacterial infection	Diarrhoea	57%
Average Accuracy Rate:		59.6%

6.1 INPUT

In the input design, the client enters any number of symptoms (as many as he faces)

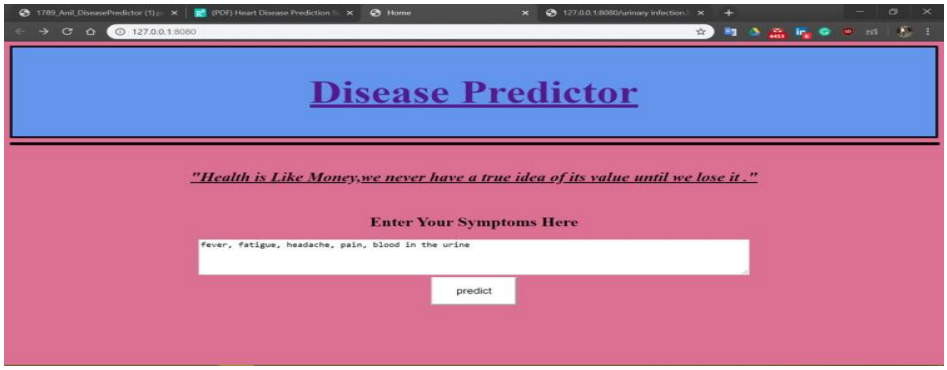

Figure 2: Input

6.2 OUTPUT

Output yields the disease which has the most probability as per the calculation.

Figure 3: Output

6.3 DETAILS OF DISEASE

If the User is interested in getting into the details of disease he may click on the "click here" button available beside the text "for more details" at the bottom of the page.

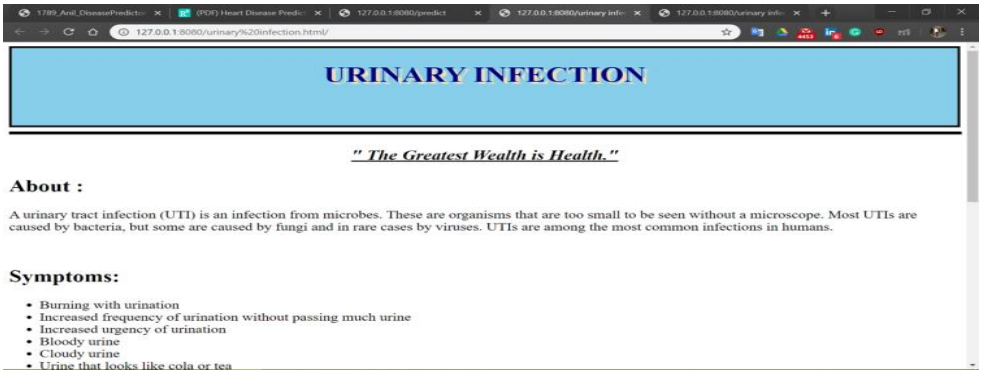

Figure 4: Details of Disease

7. FUTURE WORK

Every one of us would like to have a good and fast medical care system and physicians are expected to be medical experts and take good decisions all the time. But it's highly unlikely to memorize all the knowledge,

patient history, records needed for every situation. Although they have all the massive amounts of data and information; it's difficult to compare and analyse the symptoms of all the diseases and predict the outcome. The client need not run every time to the doctor for the consultation and prescription of minor diseases. Instead they can reach them online.

Now-a-days a lot of work is being done in this field in order to help people in their medication without event physically present in front of the doctor.

8. CONCLUSION

This venture intends to foresee the malady based on the symptoms. The task is planned so that the framework accepts manifestations from the client as info and produces yield for example foresee sickness. Normal forecast exactness likelihood of 55% is acquired.

A definitive objective is to encourage facilitated and very much educated social insurance frameworks fit for guaranteeing most extreme patient fulfilment. Patients can get the opportunity to get higher knowing and can get the opportunity to expect a ton of duty regarding his or her own consideration, on the off chance that they are to utilize the data inferred. Doctor jobs can presumably adjust to a great deal of a guide than head, who will exhort, caution and help singular patients. Doctors may see a great deal of satisfaction in applying as positive results increment and negative results decline. Maybe time with singular patients can increment and doctors will some other time have the opportunity to make positive and enduring associations with their patient.

9. REFERENCES

Davis, A. Chawla, N., Blumm, N., Christakis, N., Barbasi, A. L., Predicting Individual Disease Risk Based on Medical History, Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, 2008.

Adam, S., and Parveen, A., Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012.

Al-Aidaroos, K., Bakar, A., and Othman, Z., Medical Data Classification with Naive Bayes Approach, Information Technology Journal, 2012.

Soni, J., Ansari, U., Sharma, D., and Soni, S., Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications,2011. ·

Nishar Banu, MA Gomathy, B, Disease Predicting System Using Data Mining Techniques, International Journal of Technical Research and Applications,*2013.*