iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5002

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Prediction of Different Diseases and Development of a Clinical Decision Support System using Naïve Bayes Classifier

Fatema Tuz Zohra
Computer Science and Engineering, BAUST, Bangladesh

Abstract: This work is dedicated to patients specially, rural patients can get to know the early stage detection of diseases before laboratory tests reducing the unlimited waiting time and cost expenditure. Clinical Decision Support System (CDSS) can be used for analyzing diseases to predict almost accurate disease automatically and patient's query. This work has been done with the help of a doctor as a human expert. We collected 300 sample data from patients. We have made the dataset from our sample patient's information. Naïve Bayes classifier is used here to classify the diseases easily. The selected diseases are Malaria, Tuberculosis, Stroke, Fever, Diabetes, Heart disease. The prediction of a disease is measured with the prediction of a doctor before laboratory tests to get the system's accuracy. Here we got 100% accuracy on the trained dataset containing 180 cases.
Keywords: Clinical Decision Support System (CDSS), Classifier, Expert System, Naïve Bayes.

I. INTRODUCTION

Healthcare is one of the basic needs of survival of people. Now a days, people get easily contaminated with disease at early age due to lack of proper knowledge, improper food habit, physical inactivity which leads to death with so much cost for medical issues. Besides in rural areas, health care system is not developed yet successfully. Specialist doctors and diagnostic centres are not always available there. Common symptoms of diseases make people confused to know the right disease what they are facing. Sometimes people don't get understand that they should go to which doctor for his disease**.** Providing the right explanations regarding the situation of a patient at the right time is a key for improving the diagnostic process in health care system [1]. The health care system can be improved by utilizing the right explanations of patient's case at the right time.

There are many common diseases like heart disease, cancer, diabetes, malaria, fever, stroke, tuberculosis etc. Heart disease is the number one cause of death globally. 17.6 million deaths cause to heart disease in 2016 but within 2030 it can be more than about 23.6 million [2]. Diabetes is the most growing disease now-a-days. About 7.1 million people have diabetes and almost equal number with undetected diabetes. This will be double in 2025 [3]. To reduce this increasing situation, the main work will be raising consciousness among patients and not to delay with their diseases. That is why we develop a model so that patients can get information about their diseases after analysing the given symptoms. They can also enquiry about any specific features of hospital information to the system. This model can also help new physicians to get a decision. CDSS which acts like a part of a doctor isn't like a human whom need to sleep or eat. CDSS do not have to waste time for enjoyment or sickness. As a result, it can active always and provide its function. It can cut the costs for any hospital from paying full time receptionists to provide directions for patients [4]. Moreover the patient who are in rural area can not go to hospital for determining disease and to find the right doctor, find solution and know available doctor's name instant. Our aim is to build a system that will help people who wants to save their precious time rather than waiting and gets a right way to doctor after disease prediction.

II. BACKGROUND

A. Expert System

Expert systems are an application of machine learning. Expert systems solve real problems itself which normally would require a specialized human expert. Expert systems can be used broadly in healthcare, education, business, finance, manufacturing. They do not get bored or tired. This technology is based on the premise that what makes a person an expert is years of experience that enables him to recognize certain patterns in a problem as being similar to patterns he has seen previously. It had been implemented as a software which not only synchronized and personalized the emails but also generate automatic reply either by sending fixed reply or the response send earlier [5]. They personalized emails according to the user and helped in automatically replying to the known queries which saved a lot of time.

B. Expert System for Healthcare

CDSS can be defined as also "A computer system that uses two or more patient data to generate case specific or encounter specific advice" [6]. The main idea of it to raise awareness among patients. "Model-based set of procedures for processing data and judgements to assit a manager in his decision making" [7]. Case based reasoning is developed and diagnosis only for palliative care [8].

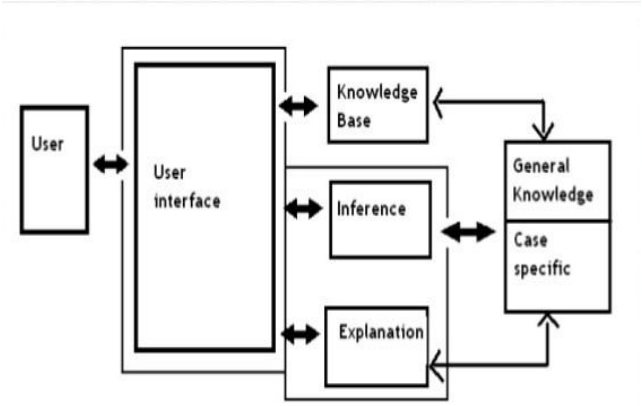


Fig 1: Expert system architecture

For the explanation of a situation of a patient needs data to characterize the problems. So data can be collected from patients which will be used as a source of information. [15] Users can have a user id. They can interface through user interface. They can ask general knowledge about the hospital. They can also ask for suggestion depending on their physical problems that are facing. Then the case will be analysed depending its key symptoms and case specific disease result will be provided with a probability. If any doctor present then he will be preferred for that disease. In other systems specialized diseases are classified only, this model can classify general diseases which are common now.

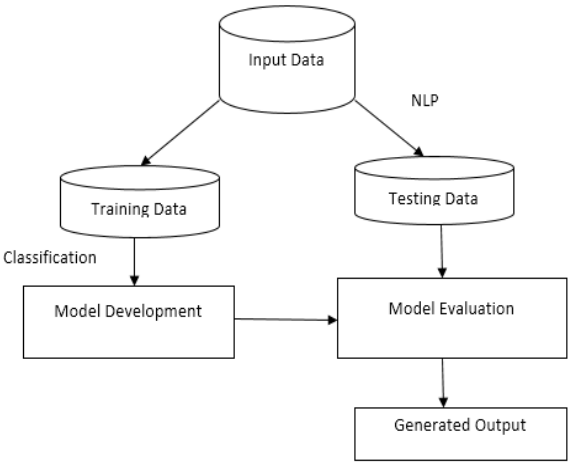


Fig 2: Flowchart of our system

C. Natural Language Processing (NLP)

NLP is used here for extracting the key words excluding the stop words and irrelevant parts of speech words through generating tokenization. When an user gives input as an enquiry about hospital information then the general database gives relevant information to the user. Such as if any user search that whether that hospital have card payment options? Then the system can reply that yes, we have card payment options. Thank you for your enquiry. Besides any user can search any specialist doctor and can have an appointment through online which is too time consuming and hustle in real life. Most importantly anyone can input his symptoms and can get the probable disease name.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Machine learning is an application of artificial intelligence that helps any system to learn. Its aim is to allow the machines learn automatically without human interaction. There are two types of machine learning algorithms [9].

1) *Supervised Machine Learning:* In supervised machine learning, system is trained through data to predict outcome for any given input. It has two phase 1) training b) testing. Through training it gathers information to learn and through testing it can provide answer itself based on training. Two types of supervised machine learning algorithms are available.

a) *Classification:* When selection has to be done among many classes then classification technique will be used. It is a statistical method to predict. There are many classification techniques such as Naïve Bayes, Support Vector Machine (SVM), Neural Network etc.

b) *Regression:* Regression technique predicts a single output value using training data.

2) *Unsupervised Machine Learning:* In unsupervised machine learning, system is not trained through training phase. It can generate output from its own supervision. Cluster algorithms, K-means, Hierarchical clustering etc. are common unsupervised techniques.

D. Naïve Bayes Classification

Naïve Bayes is a machine learning algorithm which is based on bayes theorem. It can generate output through a probabilistic method. It is very simple and fast classification technique. But in complex system with large dependency sometimes it does not give accurate result [10]. It can be done through calculating the probability for each class, assuming conditional independence of the attributes of class. The NB technique has been provided a remarkable classification in medical diagnosis and system performance measurement [11]. Laryngeal cancer based CDSS is developed on the basis of Bayesian Network having 1000 variables with about 1300 dependencies. A subsystem of 303 variables reached 100% correct predictions [12]. Naïve bayes is used for text classification [13]. Naïve Bayes gives probability of liver diseases from EMR text data using [14].

Bayes Rule:

In training phase, P(Evidence | Output) is given. Here outcome is known based on evidence. In testing phase,

P(Output| Evidence) prediction of output will be obtained based on evidence from training phase.

According to Bayes rule,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

In any cases we can have multiple X variables. When the variables or features (X) are independent of each other then the Bayes rule converts to Naïve Bayes.

$$P(Y=k|X1..Xn) = \frac{P(X1|Y=k)*P(X2|Y=k)...*P(Xn|Y=k)*P(Y=k)}{P(X1)*P(X2)...*P(Xn)}$$

Here k is a class of Y.

$$\text{Probability of Outcome| Evidence} = \frac{Probability\ of\ Likelihood\ of\ Evidence*Prior\ Probability}{Probability\ of\ Evidence}$$

(Probability of Outcome| Evidence) can be called posterior probability. Probability of evidence is same for all classes of y.

Let's consider a following sample dataset.

Disease name	shivering	headache	tiredness	sweating	vomit	Weight loss	fever	cough
Malaria	0	1	0	1	0	1	1	0
Tuberculosis	0	0	0	1	0	0	0	1
Stroke	0	1	0	0	1	0	0	0
Malaria	1	0	0	1	1	0	1	0
Stroke	0	0	0	0	1	0	0	0
Tuberculosis	0	0	0	0	1	1	1	1
Malaria	0	1	0	0	0	1	1	0

Table 1: Sample dataset

After pre-processing a patient has given the symptoms as Headache, Sweating, Weight Loss and Fever frequently. What could be the disease based on above data?

We can summarize the above dataset in the following table.

Disease Name	Headache	Sweating	Weight Loss	Fever	Total
Malaria	2	2	3	3	3
Tuberculosis	0	1	1	0	2
Stroke	1	0	0	0	2
Total	3	3	4	3	7

Table 2: Summarized table for given input

Step1: To compute the prior probabilities of each of the class of diseases
P (Y = Malaria) = 3/7 = 0.42
P (Y = Tuberculosis) = 2/7 = 0.29
P (Y = Stroke) = 2/7 = 0.29
Step 2: To compute the probability of evidences
P (x1 = Headache) = 3/7 = 0.42
P (x2 = Sweating) = 3/7 = 0.42
P (x3 = Weight Loss) = 4/7 = 0.57
P (x4 = Fever) = 3/7 = 0.42
Step 3: To compute the probability of likelihood of evidences
Probability likelihood for Malaria:
P (x1= Headache| Y= Malaria) = 2/3 = 0.67
P (x1= Sweating | Y= Malaria) = 2/3 = 0.67
P (x1= Weight Loss | Y= Malaria) = 3/3 = 1
P (x1= Fever | Y= Malaria) = 3/3 = 1
Probability likelihood for Tuberculosis:
P (x1= Headache| Y= Tuberculosis) = 0/2 = 0
P (x1= Sweating | Y= Tuberculosis) = 1/2 = 0.5
P (x1= Weight Loss | Y= Tuberculosis) = 1/2 = 0.5
P (x1= Fever | Y= Tuberculosis) = 1/2 = 0.5
Probability Likelihood for Stroke:
P (x1= Headache| Y= Stroke) = 1/2 = 0.5
P (x1= Sweating | Y= Stroke) = 0/2 = 0
P (x1= Weight Loss | Y= Stroke) = 0/2 = 0
P (x1= Fever | Y= Stroke) = 0/2 = 0
Step 4: Combining all the three steps into the Naïve Bayes formula to get the probability
Probability that the disease is Malaria:

$$P(Malaria|Headache, Sweating, Weight\ Loss, Fever)$$
$$= \frac{P(Headache|Malaria)*P(Sweating|Malaria)*P(Weight\ Loss|Malaria)*P(Fever|Malaria)*P(Malaria)}{P(Headache)*P(Sweating)*P(Weight\ Loss)*P(Fever)}$$
$$= \frac{0.67*0.67*1*1*0.42}{P(Evidence)}$$

P(Evidence) is same for class Malaria. It can be assumed as constant.
$P(Tuberculosis|Headache, Sweating, Weight\ Loss, Fever)$ = 0
$P(Stroke|Headache, Sweating, Weight\ Loss, Fever)$ = 0
So, the possible disease class is Malaria.
We can get a percentage value from probability of Malaria
Percentage = Probability * 100
Here sometimes we get the probability as zero. By Laplace correction we add very small value from 0 to 1 such as 0.1 to every count so that it can never be zero.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

III. TOOLS

There are many types of tools used to build this model. The resources are used as PHP, MySQL, JavaScript, jQuery Artificial Intelligence Markup Language and Bootstrap as framework.

IV. EXPERIMENTAL ANALYSIS

The result is analyzed in two hospital with doctor's result before patient diagnosis. We asked 100 sets of input questions of different users to an actual doctor and collect the actual probability of the disease and compare them with our system's answers. Considering this we calculate the performance of our system and it is almost 80% accurate in accordance with doctor's result based on only symptoms. It can be used by doctors, general people and any hospital management to provide their all available features through a website.

System's Accuracy = $\frac{Number\ of\ true\ outputs}{Number\ of\ true\ outputs + Number\ of\ false\ outputs}$

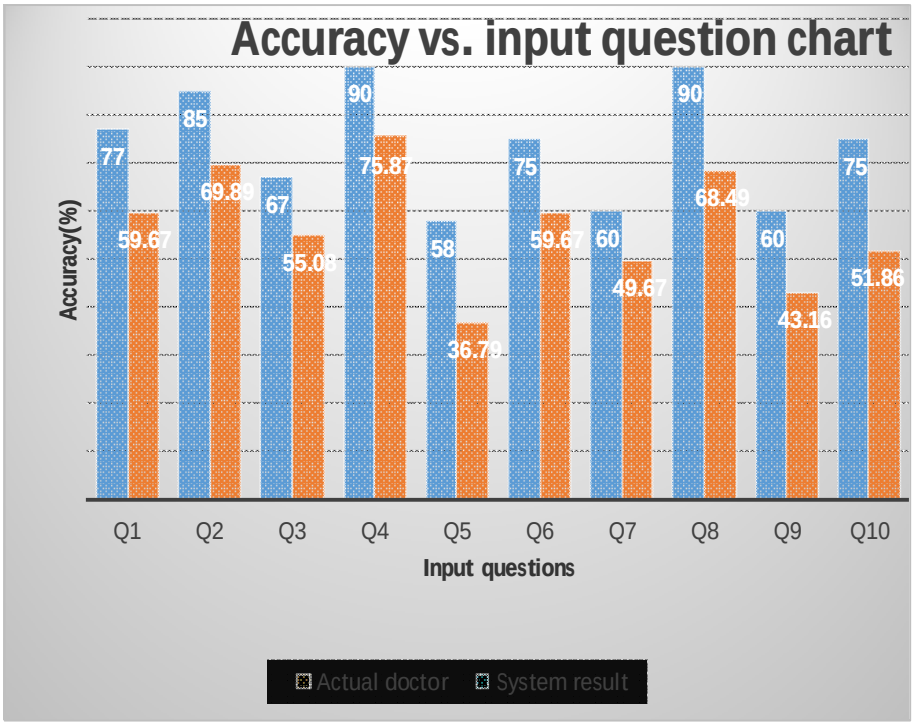


Fig 3: Accuracy chart

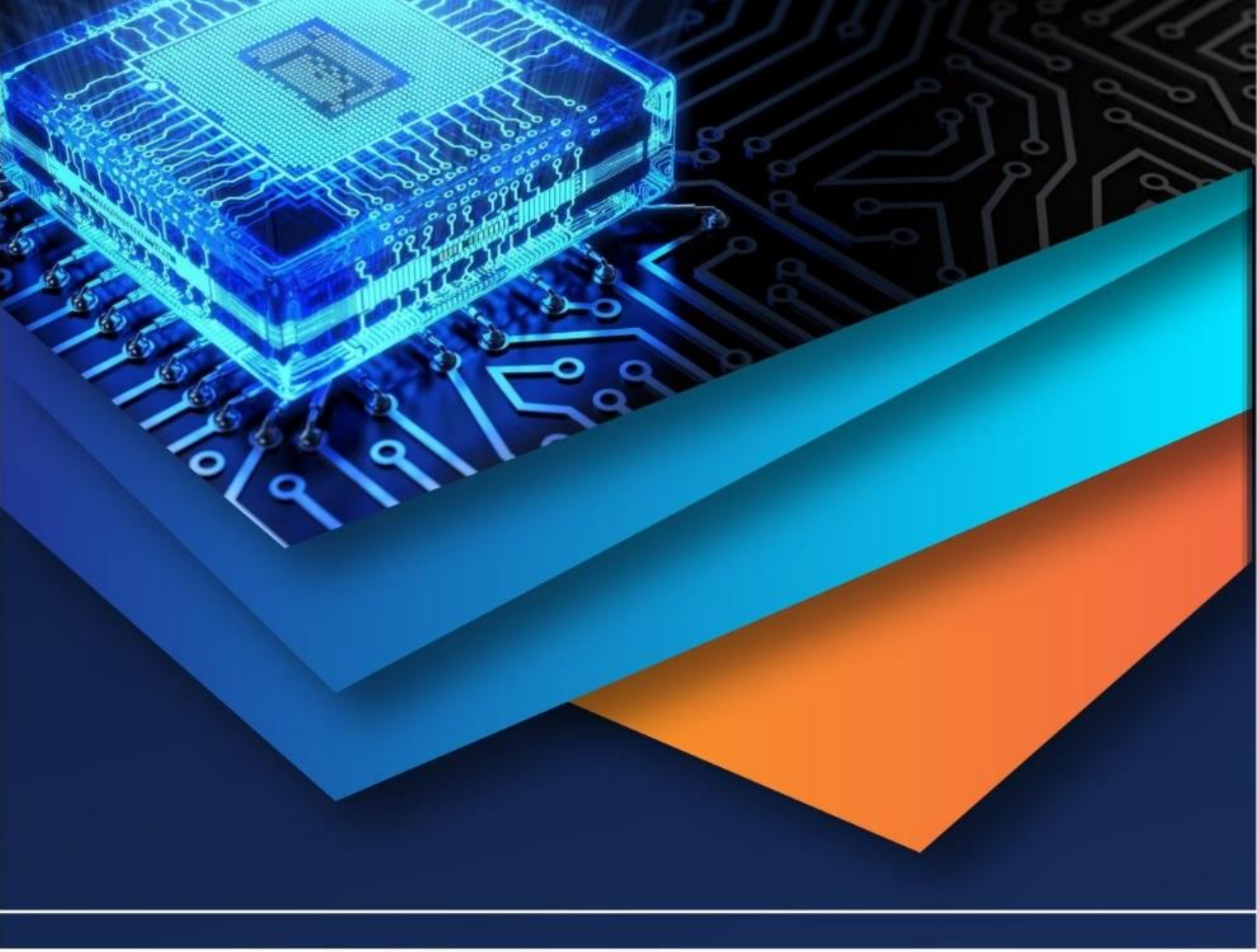
V. CONCLUSIONS

CDSS is a framework to use patient information with medical knowledge to derive models so that it can predict the possible disease and give patients instant decision based on predictions .The diseases are classified successfully. The attributes of diseases are selected with the help of a doctor. Any patient gets a prediction of a disease depending on his/her symptoms. The predicted value is calculated through Naïve bayes classifer which acts as a classifier and gives a probabilistic output. Naïve bayes classify the diseases easily whether a patient have a chance of the selected diseases or not. Besides it can assist health professionals while taking decisions. It is 80 % accurate in accordance with doctor's result. Besides it is 100% accurate in its trained dataset.

According to World Health Organization (WHO), tele-health program provides surveillance, health promotion programs and public health functions to raise awareness in people about their health issues. Tele-health indicates non-clinical services [16]. If people get to know about the probable disease and their basic solution at early stage then the impact of disease can be reduced greatly. It can be done by CDSS. The main idea of it to raise awareness among patients. In medical data mining each of minor features should be counted with common pattern for accurate prediction. The future work can be extended with large data to analyse and classify more diseases and the feedback system from users so that the machine can teach itself from the feedback analysis for any particular class. Moreover it can be modified for any hospital. History of each users can be recorded for supervision further. It can be modified as e-health monitoring system.

REFERENCES

[1] Masic I. and Novo A., "History of Medical Informatics in Bosnia and Herzegovina", Medical Faculty, University of Sarajevo, MedArh 2006.
[2] Heart Disease and Stroke Statistics-2019 At a Glance, American Heart Association, 2019.
[3] A K Mohiuddin, "Internationale Journal of Diabetes Research," Vol 2, No 1, 24 February 2019, pp 14-20.
[4] S.Mahajan and G.Shrivastava, " Effective Diagnosis of Diseases through Symptoms Using Artificial Intelligence and Neural Network", International Journal of Engineering Research and Applications (IJERA), 2013.
[5] M. Srivastava, M. Goyal, "Personalization of Automatic E-mail Response for the University System", The Next Generation Information Technology Summit (4th International Conference), pp. 485 – 489, Sept. 2013.
[6] W. JC and Liu JL.J, "Basic concepts in medical informatics", Epidemiol Community Health, 56(11), pp. 808-812, 2002.
[7] Little and J.D.C., "Models and Managers: The Concept of a Decision Calculus", M,anagement Science 16(B), B466-B485, 1970.
[8] A.Aamodt, O.E.Gundersen, J.H.Loge, E.Wasteson and T.Szczepanski, "Case-Based Reasoning for Assessment and Diagnosis of Depression in Palliative Care", IEEE 23RD International Symposium on computer based Medical system, page 480-485, doi: 10.1109/CBMS.2010.6042692, October 2010.
[9] https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/.
[10] S. Joshi and M. K. Nair, "Survey of Classification Based Prediction Techniques in Healthcaere", Indian Journal of Science and Technology, Vol 11(15), April 2018, DOI: 10.17485/ijst/2018/v11i15/121111.
[11] Y. Kumar and G. Sahoo, "Prediction of different types of liver diseases using rule based classification model", Technology and Health Care 21(2013), pp417-432, 2013, DOI: 10.3233/THC-10742.
[12] M. A. Crypko and M. Stoehr, "Digital patient models based on Bayesian networks for clinical treatment decision support", Minimally Invasive Therapy and Allied Technologies, 27 February,2019. DOI: 10.1080/13645706.2019.1584572.
[13] Y.Jiang, H.Lin, X.Wang, D.Lu, and Z. Gong, "A Technique for Improving the Performance of Naive Bayes Text Classification", WISM 2011, Part II, LNCS 6988, pp. 196–203, 2011., Springer-Verlag Berlin Heidelberg.
[14] Y.Shen,Y.Li,Hai.T.Z,B.Tang and M.Yang, "Enhancing ontology-driven diagnostic reasoning with a symptom-dependency-aware Naïve Bayes classifier", DOI: 10.1186/s 12859-0192924-0, 2019.
[15] L.Gastaldi and M.Corso , "Managing ICT to Solve the Exploration -Exploitation Paradox In Healthcare", 2015.
[16] www.chironhealth.com.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY