2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Disease Prediction using Machine Learning Algorithms

Sneha Grampurohit
Electronics and communication
K.L.E Institute of Technology
Hubli, India
snehagrampurohit5@gmail.com

Chetan Sagarnal
Electronics and communication
K.L.E Institute of Technology
Hubli, India

Abstract- **The development and exploitation of several prominent Data mining techniques in numerous real-world application areas (e.g. Industry, Healthcare and Bio science) has led to the utilization of such techniques in machine learning environments, in order to extract useful pieces of information of the specified data in healthcare communities, biomedical fields etc. The accurate analysis of medical database benefits in early disease prediction, patient care and community services. The techniques of machine learning have been successfully employed in assorted applications including Disease prediction. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage. A Sample data of 4920 patients' records diagnosed with 41 diseases was selected for analysis. A dependent variable was composed of 41 diseases. 95 of 132 independent variables(symptoms) closely related to diseases were selected and optimized. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier. The paper presents the comparative study of the results of the above algorithms used.**

Keywords: Machine Learning, Data mining, Decision Tree classifier, Random forest classifier, Naive Bayes classifier.

I. INTRODUCTION

The healthcare and medical sector are more in need of datamining today. When certain data mining methods are used in a right way, valuable information can be extracted from large database and that can help the medical practitioner to take early decision and improve health services. The spirit is to use the classification in order to assist the physician.

Diseases and health related problems like malaria, dengue, Impetigo, Diabetes, Migraine, Jaundice, Chickenpox etc., cause significant effect on one's health and sometimes might also lead to death if ignored. The healthcare industry can make an effective decision making by "mining" the huge database they possess i.e. by extracting the hidden patterns and relationships in the database. Data mining algorithms like Decision Tree, Random Forest and Naïve Bayes algorithms can give a remedy to this situation. Hence, we have developed an automated system that can discover and extract hidden knowledge associated with the diseases from a historical(diseases-symptoms) database according to the rule set of the respective algorithms.

II. OVERVIEW

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have.

The considered symptoms are:

TABLE I. SYMPTOMS

Symptoms		
Back pain	Bloody stool	scurrying
Constipation	depression	Passage of gases
Abdominal pain	Irritation in anus	Weakness in limbs
diarrhea	Neck pain	Fast heart rate
Mild fever	dizziness	Internal itching
Yellow urine	cramps	Toxic look
Yellowing of eyes	bruising	palpitations
Acute liver failure	obesity	Painful walking
Fluid overload	Swollen legs	Prominent veins on calf
Swelling of stomach	irritability	Fluid overload
Swelled lymph nodes	Swollen blood vessels	Excessive hunger
malaise	Muscle pain	Black heads
Blurred and distorted vision	Pain in anal region	Pain during bowel movements
phlegm	Brittle nails	Rusty sputum
Throat irritation	Belly pain	Mucoid sputum
Redness of eyes	Enlarged thyroid	Puffy face and eyes
Sinus pressure	Slurred speech	Hip joint pain
Runny nose	Knee pain	polyuria
congestion	Skin peeling	Family history
Chest pain	Extra marital contacts	Swollen extremities

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

Symptoms		
Yellow crust ooze	Swelling joints	Coma
Loss of smell	Stiff neck	Unsteadiness
Movement stiffness	Muscle weakness	Drying and tingling lips
Spinning movements	Red sore around nose	Weakness of one body side
Bladder discomfort	Foul smell of urine	Continuous feel of urine
Altered sensorium	Red spots over body	Abnormal menstruation
Dyschromic patches	Watering from eyes	Increases appetite
Lack of concentration	Visual disturbances	Receiving blood transfusion
Receiving unsterile injections	Distention of abdomen	History of alcohol consumption
Puss filled pimples	Blood in sputum	Stomach bleeding
Silver like dusting	Small dents in nails	Inflammatory nails
blister		

The diseases considered are:

TABLE II. DISEASES

Diseases		
Fungal Infection	Malaria	Varicose veins
Allergy	Chickenpox	Hypothyroidism
Gerd	Dengue	Vertigo
Chronic cholestasis	Peptic ulcer disease	acne
Drug reaction	Hepatitis A	Urinary tract infection
Piles	Hepatitis B	Psoriasis
AIDS	Hepatitis C	Impetigo
Diabetes	Hepatitis D	Hyperthyroidism
Gastroenteritis	Hepatitis E	Hypoglycemia
Bronchial Asthma	Alcoholic hepatitis	Cervical Spondylosis
Hypertension	Tuberculosis	Arthritis
Migraine	Common cold	Osteoarthritis
Paralysis	Pneumonia	Typhoid
Jaundice	Heart Attack	

The generalized prediction model can be given as:

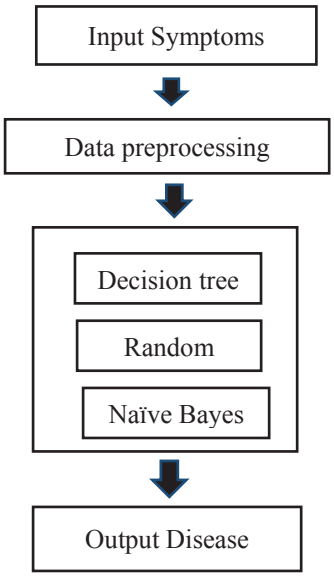

Fig. 1. PREDICTION MODEL

A. *Input (Symptoms)*

While designing the model we have assumed that the user has a clear idea about the symptoms he is experiencing. The Prediction developed considers 95 symptoms amidst which the user can give the symptoms his processing as the input.

B. *Data preprocessing*

The data mining technique that transforms the raw data or encodes the data to a form which can be easily interpreted by the algorithm is called data preprocessing. The preprocessing techniques used in the presented work are:

- *Data Cleaning*: Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data.
- *Data Reduction*: The analysis becomes hard when dealing with huge database. Hence, we eliminate those independent variables(symptoms) which might have less or no impact on the target variable(disease). In the present work, 95 of 132 symptoms closely related to the diseases are selected**.**

C. *Models selected*

The system is trained to predict the diseases using three algorithms

- Disease Tree Classifier
- Random forest Classifier
- Naïve Bayes Classifier

A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database.

D. *Output(diseases)*

Once the system is trained with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model, those symptoms are processed according the rule set developed,

thus making classifications and predicting the most likely disease.

III. METHODOLOGY

The disease prediction system is implemented using the three data mining algorithms i.e. Decision tree classifier, Random forest classifier and Naïve Bayes classifier. The description and working of the algorithms are given below.

A. Decision Tree Classifier

The classification models built by decision tree resemble the structure of tree. By learning the series of explicit if-then rules on feature values (symptoms in our case), it breaks down the dataset into smaller and smaller subsets that results in predicting a target value(disease). A decision tree consists of the decision nodes and leaf nodes.

- *Decision node*: Has two or more branches. In our work presented, all the symptoms are considered as decision nodes.
- *Leaf node*: Represents the classification that is, the Decision of any branch. Here the Diseases correspond to the leaf nodes.

A. ID3 Algorithm:

One of the core algorithms we have used in our work is the ID3 algorithm invented by J.R. Quinlan. ID3 uses a top down, greedy search through the given columns, where each column(attribute=symptoms) at every node is tested and selects the attribute(symptom) that is best for classification of a given set. To choose which symptom is best to build a decision Tree, ID3 uses Entropy and Information Gain.

1) Entropy: Amount of uncertainty or randomness. That is, it represents predictability of a certain event.
To build a decision tree, we need to calculate two types of entropy using frequency tables of each attribute as follows:

- Entropy E(C) using the frequency table of one attribute, where C is a current state (existing outcomes) and P(h) is a probability of an event h of that state C:

$$\boldsymbol{E(C)} = \sum_{h\in H} -P(h)\log_2 P(h) \qquad (1)$$

- Entropy E (C, A) using the frequency table of two attributes- C and A, where C is a current state with an attribute A and A is the considered attribute, P(h) is a probability of an event H of an attribute A.

$$\boldsymbol{E(C,A)} = \sum_{h\in H} [P(h) * E(C)] \qquad (2)$$

E(C) is the Entropy of the whole set, while the second term ***E (C, A)*** corresponds to an attribute A.

2) Information Gain: Information gain (also called as Kullback-Leibler divergence) represented by IG (C, A) for a state C is an effective change in entropy after finalizing an attribute A. It measures the relative change (decrease) in entropy with respect to the symptoms, as below:

$$\boldsymbol{IG(C,A)} = E(C) - E(C,A) \qquad (3)$$

B. Example:

Let us consider a medical record of 12 patients who were prone to Dengue and Malaria. The symptoms shown by them are as follows: High fever, Vomiting, Shivering, Muscle paining

TABLE III. SAMPLE MEDICAL RECORD

HIGH FEVER	VOMITING	SHIVERING	MUSCLE PAINING	DISEASE
1	0	1	0	Dengue
1	1	0	0	Malaria
0	1	0	1	Malaria
0	0	1	1	Dengue
1	0	1	1	Dengue
1	1	0	1	Dengue
1	1	1	1	Malaria
0	1	1	1	Dengue
1	1	1	0	Malaria
0	1	1	0	Dengue
1	0	0	1	Dengue
0	1	0	0	Dengue

In the above example, 1 represents presence of that symptom and 0 represents the absence of the symptom.

Step 1:
Calculation of ***E(C):***

Dengue	Malaria	Total
8	4	12

Using formula (1)

$$E(C) = \sum_{h\in H} -P(h)\log_2 P(h)$$

$$= -\frac{8}{12}\log_2\left(\frac{8}{12}\right) - \frac{4}{12}\log_2\left(\frac{4}{12}\right)$$

$$= 0.91822$$

The distribution is *Fairly random.*

Step 2:
In order to build a decision tree, Root Node should be decided first. The symptom with the highest *Information gain* is considered as the Root node. Starting with "*High Fever*" symptom, calculate ***E (C, High fever)*** and ***IG (C, High fever):***

Using formula (2) and (3)

$$IG(C, High\ fever)$$

$$= E(C)$$

$$- E(C, High\ fever)$$

$$IG(C, High\ fever) = E(C) - \sum_{h \in H} [P(h) * E(C)]$$

Where 'h' in *P(h)* are the possible values for a symptom. Here, symptom "High Fever" takes two possible values in considered dataset i.e. *Present* (1) and *Absent* (0). Hence

x= (Present, Absent).

$$IG(C, High\ fever) = [E(C) - P(C_{present}) * E(C_{present}) - P(C_{absent}) * * E(C_{absent})]$$

Among 12 examples, we have 7 cases, where "High *fever*" symptom is present and 5 cases where the symptom is absent.

Present	Absent	Total
7	5	12

$$P(C_{present}) = \frac{Number\ of\ present\ events}{Total\ events} = \frac{7}{12}$$

$$P(C_{absent}) = \frac{Number\ of\ present\ events}{Total\ events} = \frac{5}{12}$$

Now out of 7 present cases, 4 of them lead to Dengue and 3 of them lead to Malaria So Entropy for "*Present*" Values of *High Fever* attribute:

$$E(C_{present}) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.98518$$

Similarly, out of 5 absent cases, 4 of them lead to Dengue and 1 of them lead to Malaria So Entropy for "*Absent*" Values of *High Fever* attribute:

$$E(C_{absent}) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0.721992$$

Hence,

$$IG(C, High\ fever) = E(C) - P(C_{present}) * E(C_{present}) - P(C_{absent}) * E(C_{absent})$$

$$= 0.91822 - \frac{7}{12} * 0.98518 - \frac{5}{12} * 0.721992$$

$$= 0.3483$$

Step 3:

Calculate the information gain for all the symptoms in the similar manner, thus we have:

$$IG(C, High\ fever) = 0.3483$$
$$IG(C, Vomiting) = 0.25155$$
$$IG(C, Shivering) = 0.0102$$
$$IG(C, Muscle\ wasting) = 0.0102$$

Hence, $\boldsymbol{IG(C, High\ fever)}$ has the highest information gain i.e. **0.3483**. Therefore, we choose *High fever* symptom as the root node. At this stage, decision tree looks like:

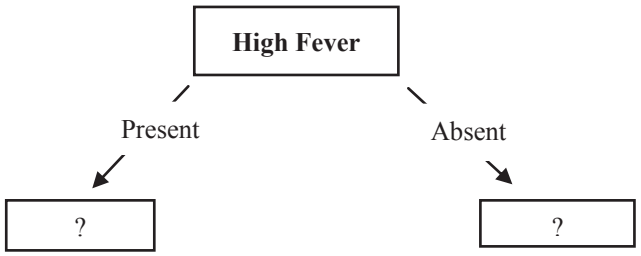

Fig. 2. Decision tree flowchart

Since we have used High fever, there are three other symptoms remaining: Vomiting, Shivering, and muscle wasting. And there are two possible values of High fever*: Present* and *Absent*, these form the sub-trees. Starting the *Present* sub-tree:

Among all the 7 examples the attribute value of High fever is Present, 4 of them lead to Dengue and 3 of them lead to Malaria

$$E(Highfever_{present}) = \sum_{h \in H} -P(h)\log_2 P(h) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.98514$$

Similarly, we calculate the Information gain values w.r.t other symptoms as follows:

$$IG(Highfever_{present}, Vomiting) = 0.6518$$
$$IG(Highfever_{present}, Shivering) = 0.07718$$
$$IG(Highfever_{present}, Muscle\ wasting) = 0.0772$$

From the above calculations, we observe that highest information gain is given by the symptom *Vomiting*. Now the tree can be modified as:

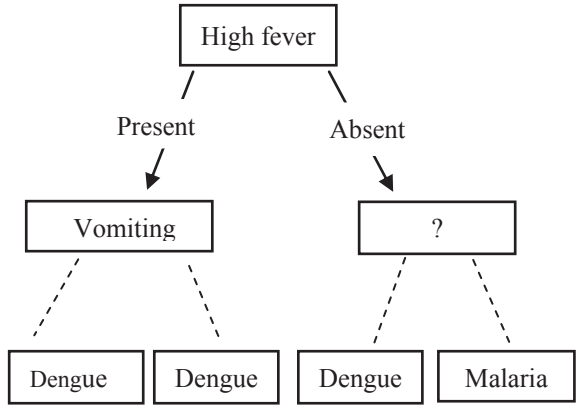

Fig. 3. Decision Tree flowchart

Therefore, proceeding in the same way the entire Decision tree can be Built leading to Dengue and Malaria at the end.ID3 follows the rule: A branch with entropy zero is a leaf node. A branch with entropy greater than zero needs splitting. In case it is not possible to achieve zero entropy, the decision is made by the method of a simple majority. As in the above case the final decision will be **Dengue.**

C. Limitation:

When all the 132 symptoms were considered from the original dataset instead of 95 symptoms, it led to **Overfitting**. i.e. the tree seems to memorize the dataset given and hence fails to classify the new data. Hence only 95 symptoms were considered choosing the optimized ones during data- cleaning step.

D. Random forest Classifier:

Random forest is a flexible, easy to use machine learning algorithm that provides exceptional results most of the time even without hyper-tuning. As mentioned in the Decision tree, the major limitation of decision tree algorithm is **overfitting.** It appears as if the tree has memorized the data.

Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to using multiple algorithms or same algorithm multiple times. Random forest is a team of Decision trees. And greater the number of these decision trees in Random forest, the better the generalization.

More precisely, Random forest works as follows:

1. Selects k symptoms from dataset (medical record) with a total of m symptoms randomly (where k<<m). Then, it builds a decision tree from those k symptoms.
2. Repeats **n** times so that we have **n** decision trees built from different random combinations of k symptoms (or a different random sample of the data, called *bootstrap sample*)
3. Takes each of the **n**-built decision trees and passes a random variable to predict the Disease. Stores the predicted Disease, so that we have a total of **n** Diseases predicted from **n** Decision trees.
4. Calculates the votes for each predicted Disease and takes the mode (most frequent Disease predicted) as the final prediction from the random forest algorithm.

E. Naïve Bayes Classifier:

The fundamental Naïve Bayes assumption

n is that each feature makes an:

- Independent
- Equal

Contribution to the outcome. Its advantage is that it works fast even on a large dataset as it requires less computational power.

1) Bayes theorem

Naïve Byes algorithm is based on Bayes theorem given by:

$$P(s/h) = \frac{P(h/s)P(s)}{P(h)} \qquad (4)$$

Where

$P(s/h)$= Posterior probability
$P(h/s)$=Likelihood
$P(s)$ = Class prior probability
$P(h)$= Predictor Prior probability

In the formula above 's' denotes class and 'h' denotes features. In *P(h),* the denominator consists the only term that is a function of data(features)- it is not a function of the class we are currently dealing with. Thus, it will be same for all the classes. Traditionally in naïve Bayes Classification, we ignore this denominator as it does not affect the result of the classifier in order to make the prediction:

$$P(s/h) \propto P(h/s)P(s) \qquad (5)$$

Key Terms:

- *Prior probability* is the proportion of Disease in the considered data set.
- *Likelihood* is the probability of classification a disease in presence of some other symptoms.
- *Marginal Likelihood* is the proportion of symptoms in the considered dataset.

2) Example:

Considering the same medical record, which was considered for decision tree, we estimate the Naïve Bayes results for set of symptoms i.e.

- High fever= Present (denoted by vale'1')
- Vomiting=Absent (denoted by vale '0')
- Shivering=Present (denoted by value '1')
- Muscle wasting=Present (denoted by value '1')

From the Table (1):

- features considered = 4 symptoms=High fever, Vomiting, Shivering, Muscle wasting.
- Classes = Dengue, Malaria
-

As per the dataset we have considered:

1. Likelihood = *P (Feature=symptoms / Class=Dengue, Malaria)*
2. Marginal Likelihood= *P(Features=symptoms)*
3. Prior Likelihood= *P(Class)*

The prediction is thus made by comparing the posterior probabilities for each class (i.e. For each disease) after observing the input symptoms. To do this, we will use expression (2). Therefore, in order to deflate our formula, let us consider the following notations:

- 'F1' for '*High fever'*, 'F2' for '*Vomiting'*, 'F3' for '*Shivering'*, 'F4' for '*Muscle Wasting'* and 'D' for '*Diseases*(class)'.

Firstly, the probability for *Dengue* is estimated *(*i.e. the class=Dengue with input symptoms as follows: "*High fever=Present*"; *"Vomiting=Absent*"; "*Shivering*=Present"; "*Muscle Wasting=Present*")

The formula thus modifies to:

P (S=Dengue | F1=Present, F2=Absent, F3=Present, F4=Present) =
= P (F1=Present, F2=Absent. F3=Present, F4=Present

| S=Dengue) * P(S=Dengue)

= P (F1=Present | S=Dengue) * P (F2=Absent | S=Dengue) * P (F3=Present | S=Dengue) * P (F4=Present | S=Dengue) * P(S=Dengue)

$$= \frac{4}{12} * \frac{4}{12} * \frac{5}{12} * \frac{5}{12} * \frac{8}{12}$$

=0.01286

Secondly, the probability for *Malaria* is estimated (i.e. the class=Malaria with the same input symptoms as mentioned in above step)

P (S=Malaria | F1=Present, F2=Absent, F3=Present, F4=Present) =

P (F1=Present, F2=Absent. F3=Present, F4=Present | S=Malaria) * P(S=Malaria) =

P (F1=Present | S=Malaria) * P (F2=Absent | S=Malaria) * P (F3=Present | S=Malaria) * P (F4=Present | S=Malaria) * P(S=Malaria)

$$= \frac{3}{12} * 0 * \frac{2}{12} * \frac{2}{12} * \frac{4}{12}$$

=0.002348

As per the calculations above:

0.0128 > 0.0023 --- > P(S=Dengue) > P(S=Malaria)

Thus, we can predict that the considered data point belongs to the class "**Dengue**", i.e. the patient with the symptoms *High fever*, *Shivering*, and *Muscle wasting* is more likely to have Dengue than Malaria.

IV. IMPLIMENTATION AND RESULTS

A. Performance of Algorithms on Training data:

The system was trained on medical record of 4920 patients prone to 41 diseases which was due to the combination of various symptoms. We have considered 95 symptoms out of 132 symptoms to avoid overfitting.

We used the K fold cross validation technique (K=5) to check the performance of all three algorithms on the dataset.

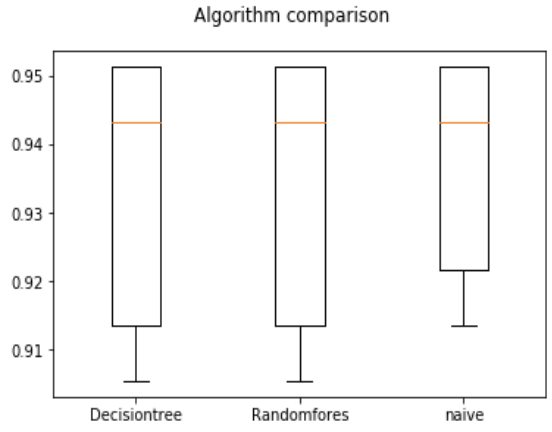

Fig. 4. Box and whisker plot of comparison of algorithms performance on training set

The above figure is a box and whisker plot showing the spread of the accuracy scores across each cross-validation fold (K=5) for each algorithm.

From these results, we can infer that all the three algorithms work exceptionally well on the dataset. However, Naïve Bayes is perhaps working a little better when compared to the other two algorithms.

The accuracy score of each algorithm after training were:

TABLE IV. ACCURACY TABLE

Algorithm used	Accuracy score
Decision Tree	0.932927
Random Forest	0.932927
Naïve Bayes	0.936179

A. Performance of Algorithms on test data

After training, the system was tested on 41 new patients records considering 95 symptoms. The accuracy score and the confusion matrix is given as by:

TABLE V. ACCURACY AND CONFUSION MATRIX

Algorithm used	Accuracy score	Confusion matrix	
		Correctly classified	Incorrectly classified
Decision Tree	0.951219	39	2
Random forest	0.951219	39	2
Naïve Bayes	0.951219	39	2

From the above table, we can infer that all the algorithms have equal accuracy score. The accuracy in terms of percentage: 95.12 percentage.

B. GUI results:

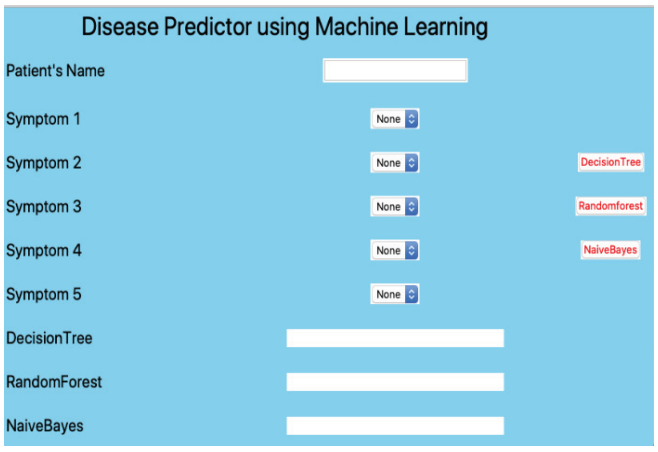

Fig. 5. An empty Disease prediction GUI

The GUI created takes in 5 symptoms from user. The user can choose the symptoms from the list of symptoms which appears when clicked on "None" option. The user can give a maximum of 5 symptoms he is facing.

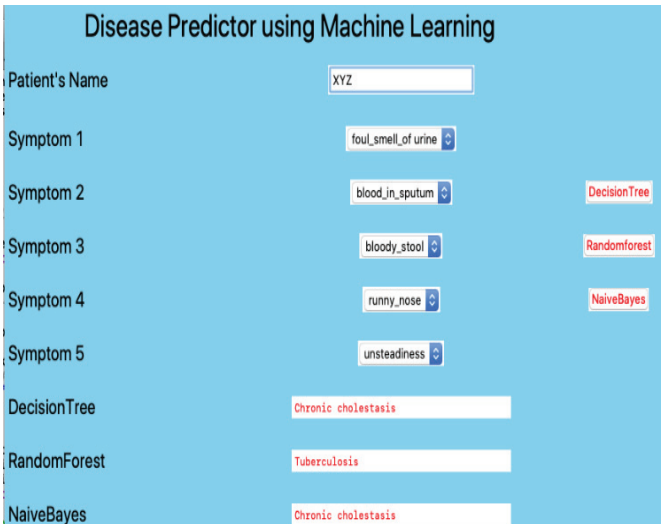

Fig. 6. Resulting GUI when user XYZ has given 5 symptoms he is facing

Once the symptoms are given, the algorithms are to be selected. As the algorithms are selected, the symptoms are processed, and the disease is searched based on the rule set that has been defined in the Methodology section.

The symptoms given by the patient XYZ were: *"fuel smell of urine*", "*blood in sputum*", "*bloody stool*", "*runny nose*", "*unsteadiness*".

Predictions by algorithms were:

Decision Tree: Chronic cholestasis

Random forest: Tuberculosis

Naïve Bayes: Chronic cholestasis

Hence the physician can go by the majority of the results i.e. the patient is more likely to have **Chronic cholestasis**

V. CONCLUSION

From the historical development of machine learning and its applications in medical sector, it can be shown that systems and methodologies have been emerged that has enabled sophisticated data analysis by simple and straightforward use of machine learning algorithms. This paper presents a comprehensive comparative study of three algorithms performance on a medical record each yielding an accuracy up to 95 percent. The performance is analyzed through confusion matrix and accuracy score. Artificial Intelligence will play even more important role in data analysis in the future due to the availability of huge data produced and stored by the modern technology.

REFERENCE

[1] Qulan, J.R. 1986. "Induction of Decision Trees". Mach.Learn. 1,1 (Mar. 1986),81-10

[2] Sayantan Saha, Argha Roy Chowdhuri et,al "Web Based Disease Detection System",*IJERT*, ISSN:22780181,Vol.2 Issue 4, April-2013

[3] Shadab Adam et.al "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[Accepted- 12/06/2012].

[4] Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", *IEEE*[Access 2017]

[5] Mr Chintan Shah,Dr. Anjali Jivani, "Comparison Of Data Mining Classification Algorithms for Breast Cancer Prediction", *IEEE*-31661

[6] Palli Suryachandra, Prof.Venkata Subba Reddy,"Comparison of Machine Learning algorithms For Breast Cancer", *IEEE.*

[7] Andrew Alikberov, Stephan Broadly et.al "*The Learning Machine",* Accessed on: March 26,2020. [Online]. Available: https://www.thelearningmachine.ai.