ISSN No:-2456 –2165

Use of Data Mining to Predict Human Diseases

Saumya Shandilya
Computer Science Department, Symbiosis Institute of Technology
saumyanda@gmail.com

Abstract- In this project, we intend to make an intelligent agent that asks the user about their medical symptoms and tries to predict the most probable diseases/medical conditions that they might be suffering from. Based on the results, it can also direct the user/patient to go to pharmacy or consult a doctor or to go for medical emergency services. It is truly said that "Prevention Is Better Than Cure". Sometimes diseases like cancers have very minor symptoms in the early stages but if detected this could save a patient's life. There is no harm in taking preventive medical advice than regretting later. Artificial Neural Networks (ANN) is currently a 'hot' research area in medicine and it is believed that they will receive extensive application to biomedical systems in the next few years. An application called the "Instant Physician" trained an auto associative memory neural network to store a large number of medical records. After training, the net can be presented with input consisting of a set of symptoms; it will then find the full stored pattern that represents the "best" diagnosis and treatment. This product can be useful for various users such as:

1. ***General Population/Patients:***

a. This can act as a preliminary advice mechanism for patients before they consult a doctor.
b. They can get suggestions as to whether they need to consult a doctor, or a visit to the local pharmacy would be fine for them.

2. ***Medical Professionals:***

a. To speed up the process of diagnosis and to reduce human errors involved in finding the possible ailments.

3. ***Medical Undergraduate/Students:***

a. To understand the common diseases and the symptoms related to them.
b. To understand all possible medical conditions which could be present in the patient who is exhibiting a said symptom.

4. ***Hospitals:***

a. Based on the diagnosis, hospital websites can display their specialist doctors that the patients can visit.

Index Terms- Artificial Neural Networks, Associative Memory Neural Network, Data Mining

I. INTRODUCTION

Sometimes people ignore some medical symptoms or conditions that they might be suffering from and do not feel like going to the doctor for every small medical problem that they are facing. Hence, we felt that there is a need for a medical health advisor that would guide people about the diseases or medical conditions that they might be suffering from. This medical health advisor is an intelligent learning and heuristics based system that predicts the diseases based on the symptoms that they enter. Based on this prediction the application would also suggest if they need to take medical advice from a doctor for their condition and if yes what kind of medical specialist do they need to visit. This application would also be useful for medical professionals and new medical students if they need to know about all the possible diseases that might be related to one particular symptom. Thus, particularly in the Indian context where medical advice is not readily available especially in rural areas, tie-ups could be done with local health centers and the state government in extending this application's reach. Medical ignorance could be life-threatening thus it is important to stay informed to stay safe.

II. LITERATURE SURVEY

Research phase is very crucial for the success of any project. The capabilities and strengths of a project depend on how strong the research is. We devoted 40% of our time towards research on various Natural Language Processing Algorithms, Sentiment Analysis Tools and various APIs.

II.1 Method of diagnosing cerebral infarction (US Patent No. 5590665 A) developed by Kazuyuki Kanai. Publication Date: Jan 7, 1997

ABSTRACT: A novel method of diagnosing cerebral infarction using a neural network, wherein plural sets of data previously obtained from healthy and sick persons, each including an age, measured values of coagulo-fibrinolytic molecular markers (e.g., D-dimer, TAT and PAP) , an index indicative of the state of cerebral infarction (e.g., 0 for healthy persons and 1 for sick persons) and the like, are repeatedly input

into a neural network to let it learn the correlation of these characteristics and, thereafter, a set of data of a person to be diagnosed, including his age, measured values of the coagulo-fibrinolytic molecular markers and the like, are input in the neural network to obtain an index indicative of his state of cerebral infarction as a degree of dangerousness of cerebral infarction. This method is significantly higher inaccuracy as compared with the prior art methods using the same data.

II.2 Artificial Neural Networks in Medical Diagnosis.

ABSTRACT: An extensive amount of information is currently available to clinical specialists, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. Each type of data provides information that must be evaluated and assigned to a particular pathology during the diagnostic process. To streamline the diagnostic process in daily routine and avoid misdiagnosis, artificial intelligence methods (especially computer aided diagnosis and artificial neural networks) can be employed. These adaptive learning algorithms can handle diverse types of medical data and integrate them into categorized outputs. In this paper, we briefly review and discuss the philosophy, capabilities, and limitations of artificial neural networks in medical diagnosis.

II.3 A Data Mining Approach for prediction of heart disease using neural networks

ABSTRACT: Heart disease diagnosis is a complex task which requires much experience and knowledge.Traditional way of predicting heart disease is doctor's examination or number of medical tests such as ECG, Stress Test, and Heart MRI etc. Nowadays, health care industry contains huge amount of heath care data, which contains hidden information. This hidden information is useful for making effective decisions. Computer based information along with advanced Data mining techniques are used for appropriate results. Neural network is widely used tool for predicting heart disease and other diseases in human beings. In this research paper, a Heart Disease Prediction system(HDPS) is developed using Neural network. The HDPS system predicts the likelihood of patient getting a Heart disease. For prediction, the system uses sex, blood pressure, cholesterol like 13 medical parameters. Here two more parameters are added i.e. obesity and smoking for better accuracy. From the results, it has been seen that neural network predict heart disease accurately.

III. RESEARCH ELABORATION

We have a unique approach to the classification algorithm for this project, i.e. we have developed our own classification algorithm for the dataset. This is because no standard algorithm such as Random Forests or Bayesian networks could be employed in this use case. Also, we intended to question the user dynamically, hence to find the order of questions was difficult using the standard algorithms.

To classify the diseases based on the symptoms, we thought of implementing a rule-based algorithm, which is the basis of AI. The algorithm which we initially thought of implementing was Apriori Algorithm, which talks about generating the most frequent itemset from a set of transactions and gives the support count of the items occurring in a said order. In essence, Apriori algorithm talks about rule based mining. Upon implementing the same on the dataset, we couldn't get accuracy more than 70%. Hence, we discarded the approach.

Next, we thought of Longest Common Subsequence (LCS) approach to understand the patterns of the dataset and generate the dynamic questions according the most frequent longest subsequence. This approach was significantly better than Apriori Algorithm as it was giving an accuracy of 85%. Upon testing with unknown data we found that this approach couldn't yield the required results.

We then thought of performing a frequency analysis of the entire data to understand the sparsity of the data and subsequently to generate the dynamic nature of questions based on the clusters and outliers of the data. The frequency analysis was done using a MultiValueMap, a class in the org.apache.commons.collections library. The MultiValueMap map stores the data set in the format such that one key can have multiple values mapped to it. In this map the key is the frequency of the symptom and value array stores all the symptoms which have the frequency same as the key. Hence, we can say that the MultiValueMap does the clustering of the dataset upon feeding the entire dataset into it. The keyset of the MultiValueMap was sorted and used as the input of the Binary Search Tree (BST) which was made to understand the nature of the frequency distribution.

Every node of the BST has the structure as follows:

1. Frequency of the node : integer value
2. Symptom list associated with said frequency: ArrayList <String> data type

A mirroring operation is performed on the BST data structure to exchange the left and right subtrees of each node. This is done to ensure that that the most frequent symptoms fall in the left subtree of the root node, hence making the traversal of the BST simple. We are implementing an inorder traversal for the entire BST to get the symptoms in decreasing order of frequency with every traversal. At every traversal, we get the symptoms associated with the node which is then used by the dynamic questioning interface to intelligently ask questions to the user. Hence, our classification algorithm builds a decision tree from the dataset and intelligently asks relevant questions based the user interactions with the system. The output of the algorithm is all the possible set of diseases associated with the set of symptoms selected by the user on runtime.

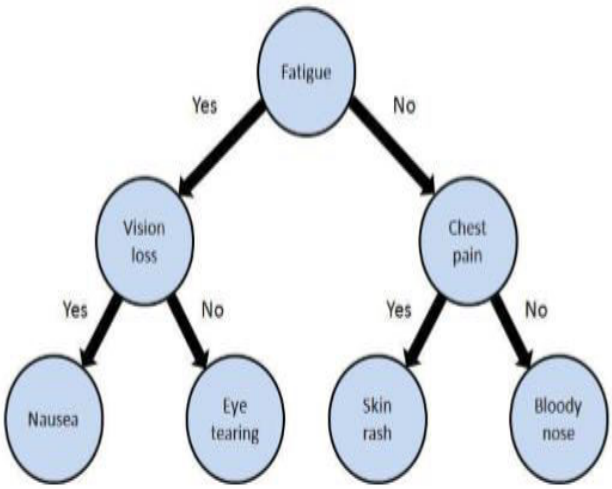

Fig.1: Example of a Decision Tree

Common symptom 2 (nausea) present?	
Yes	No
3	2

FIG 8.16: Construction of decision tree 2

- Find entropy based on one single attribute yes/no.

Entropy(Common Symptom1) = Entropy(5,9)

=Entropy(0.36,0.64)

= -(0.36 log2 0.36) - (0.64 log2 0.64)

=0.94

Fig.2: Entropy Calculation for Decision Tree

IV. RESULTS AND FINDINGS

A. *Sample Code*

```
def body_systems_description(url)
i=0
system_parts={}
doc = Nokogiri::HTML(open(url)) do |config|
 config.noblanks
end


arr_extensions=[]
doc.xpath('//div[@class="tp_rdbox_bborder_c"]/ul/li/a/text()').each do |x|
 i += 1
 count =0
 x.xpath('//div[@class="tp_rdbox_bborder_c"]/ul/li/a/@href').each do |y|
  count +=1
  if count.eql?i
   system_parts[x.to_s]=y.content.to_s[1..-1]
   arr_extensions<<y.content.to_s[1..-1]
  end
 end
end


i=0
keys = system_parts.keys


arr_extensions.each do |url_ext|
 temp_url = url+'#'+url_ext
 puts temp_url
 individual_doc = Nokogiri::HTML(open(temp_url)) do |config|
  config.noblanks
 end
 disease_arr=[]
 puts url_ext[0..-3]
 individual_doc.xpath("//div[@id=\"#{url_ext[0..-3]}\"]/div[@class=\"tp_rdbox_bborder\"]/div
                [@class=\"tp_rdbox_bborder_c\"]/div/ul/li/a/text()").each do |x|
  disease_arr<<x.content.to_s
 end
      system_parts[keys[i]]=disease_arr
      i +=1
     end


     puts system_parts.inspect
     system_parts
     end
```

This sample code uses the gem “Nokogiri” for the purpose of fetching the structure of a said webpage, which is passed as a parameter to the function body_systems_descriptions(url). The url is then parsed using the gem and the required element of the HTML page is selected using the xpath. Tree structure of the HTML node required is passed to the xpath and the processing of data is done to populate the dataset.

B. Screenshots and Outputs

```
431,[vomiting]
359,[nausea]
312,[fever]
260,[fatigue]
255,[headache]
193,[diarrhea]
183,[confusion]
```

Fig.3: Symptoms and Their Frequencies

A
Aarskog syndrome;Belly button that sticks out";"Bulge in the groin or scrotum";"Delayed sexual maturity";"Delaye
Aase syndrome;"Absent or small knuckles";"Cleft palate";"Decreased skin creases at finger joints";"Deformed ears
Abdominal aortic aneurysm;"Pain in the abdomen or back. The pain may be severe
Abdominal pain - children under age 12;"Constipation";"Gas";"Food allergy or intolerance";"Heartburn or acid refl
ABO incompatibility;"Back pain";"Blood in urine";"Chills";"Feeling of impending doom";"Fever";"Yellow skin and

Fig.4: Raw Data from Data Crawler

	A	B	C
1	Aarskog syndrome	258	595
2	Aase syndrome	98	258
3	Abdominal aortic aneurysm	1	2
4	Abdominal pain - children under age 12	27	38
5	ABO incompatibility	3	14

Fig.5: Diseases and Their Associated Symptoms

```
ut - TinyDocs (run) #2
Task completed:Creation of Multi-Map
Total time:6278 ms

Task completed:Writing mappings
Total time:203 ms

Task completed:DataSet Creator Execution
Total time:8805 ms

end of DataSet_Creator

Welcome to Tiny Docs!
--------------------------------
Do you have vomiting
yes
--------------------------------
Do you have nausea
yes
--------------------------------
Do you have fever
no
--------------------------------
Do you have fatigue
yes
--------------------------------
Do you have headache
yes
--------------------------------
Do you have diarrhea
yes

Getting the diseases:
List:1,2,4,5,6
Line numbers:[537, 928, 1057, 1460]
Diseasees:[Drug-induced hepatitis, Kidney stones, Methemoglobinemia, Shellac poisoning]
```

Fig.6: CLI Application

Fig.7: GUI Application

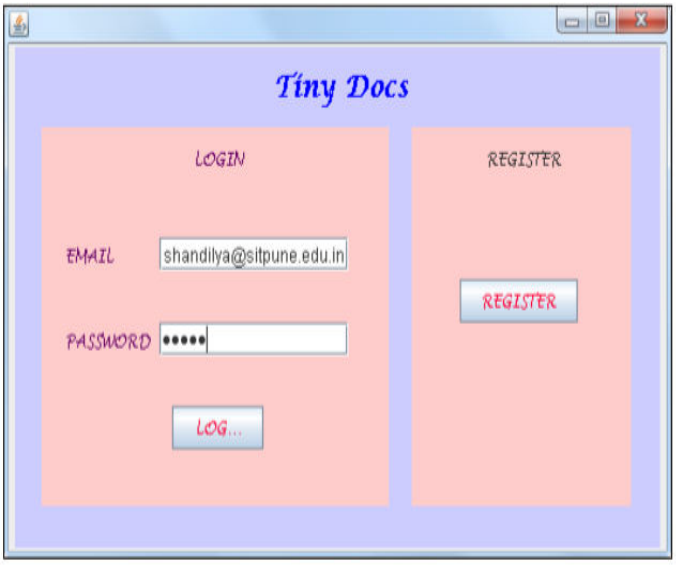

Fig.8: login page

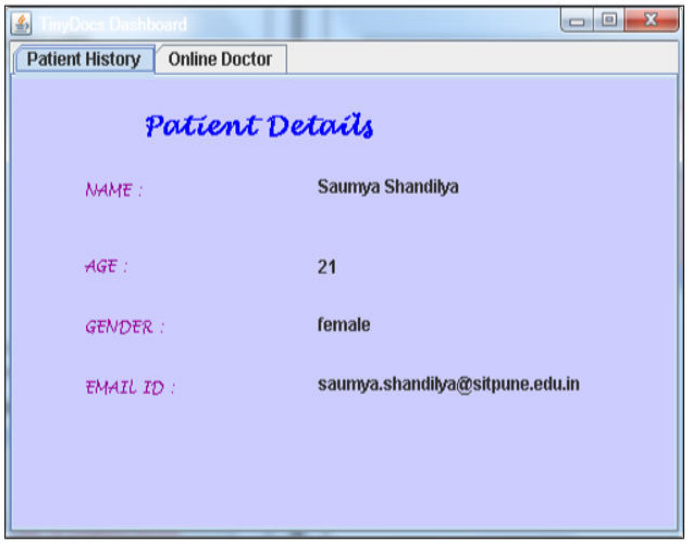

Fig.9: Patient Details Page

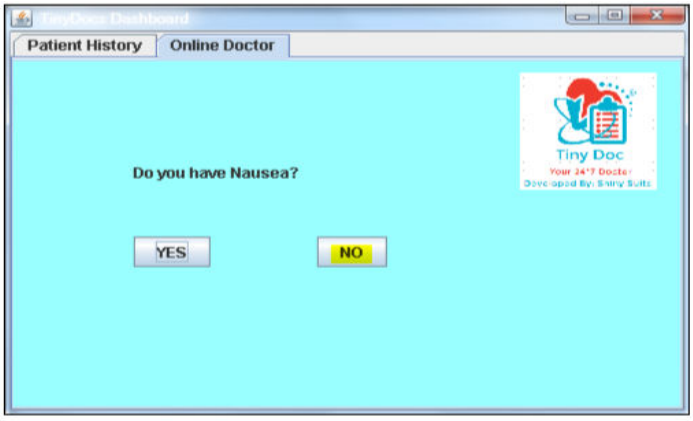

Fig.10: Sample Questions Page

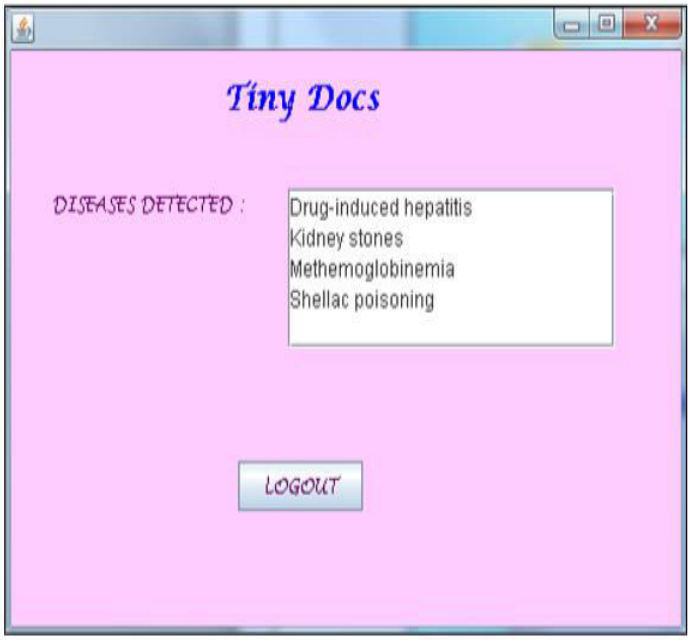

Fig.11: Sample Output Page Showing Diseases Detected

III. CONCLUSION

After extensive study about diseases and their symptoms, we have developed a preliminary health assessing tool for a common man to use. We aimed to tell the user about the possible diseases that the user may be suffering from depending on the symptoms. This application could be very useful for people who are uncertain about the diseases that they might have but do not have prompt access to medical services. At the same time, we do not intend to take the place of a general physician or OPD clinics; we just aim to guide the patient to the right type of medical assistance. While working on this project, we realized that the true Indian doesn't really have the knowledge of what he/she may be having and are ignorant about the diseases that they may be suffering from. Hence, we feel that this project will be a big contribution in this area where people hesitate are ignorant about their health or those who don't have access to medical services.

REFERENCES

[1]. Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl3 and Josef Havel1; Artificial neural networks in medical diagnosis .

[2]. Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte ; A data mining approach for prediction of heart disease using neural networks . International Journal of Computer Science and Technology, vol. 2, Issue 2, June (2011).

[3]. Kazuyuki Kanai ; Method of diagnosing cerebral imfarction . US Patent no : US005590665A, Jan. 7,1997.