Disease Prediction with Recommendation system using Deep learning

Deepa S M[1], Sivakami K[2], Revathi N[3], Dr. T. Jayaprakash[4]

Assistant Professor/ECE, Nehru Institute of Engineering and Technology
Assistant Professor/ECE, Nehru Institute of Engineering and Technology
Assistant Professor/ECE, Nehru Institute of Engineering and Technology
Professor, Department of Science and Humanities, Nehru Institute of Technology,

nietdeepa@nehrucolleges.com, ksivakaamii@gmail.com, nietrevathi@nehrucolleges.com, nitjayaprakash@nehrucolleges.com

Abstract

Health and medicine has gained a lot of importance in today's digital world, where evolving technology is being used to fight against almost all the known diseases thoroughly. But, according to reports, more than two lakh people in China and one lakh in United States of America are dead every year due to the mistakes made while prescribing errors. This paper aims to propose a system which takes as input the symptoms from the patient to predict the disease, which is followed by recommending the correct medicine. This system consists of database system module where disease dataset contains attributes like "DiseaseID", "Disease", "Symptoms" and "count of occurrences" and drug dataset has additional attributes such as "drug" and "drug_rating". In the data preparation module, a new dataframe is created where each disease with its associated symptom is given "1" and "0" if it is not the symptom related. After all these steps of preprocessing are performed on the raw dataset, the dataset is now ready to be acted upon by a machine learning algorithm. Models like Decision Tree, Random Forest and Naive Bayes were applied on the training-dataset. Entire knowledge database was fitted in models and it was seen if not much dataset is used Random Forest and Naïve Bayes classifiers showed 84% accuracy whereas Decision Tree gives 92%.But Random Forest also gives accuracy similar to Decision Tree for large dataset .Hence for disease prediction we use Decision Tree. The recommendation is done on the basis of the rating of the drug. After this, only the drug with the highest rating is included in a newly formed dataset and the rest of the drugs are ignored. The system is implemented for user access by making use of Tkinter a python GUI Library for constructing the desktop end user interface.

This paper deals with the design and implementation of a system which performs the function of prediction of diseases and the recommendation of medicines, which adds to the capabilities of the present systems in health infrastructure.

Keywords: *A Decision tree map, classification algorithms and data mining*

I. Introduction

1.1 Proposed System

Statistics show that more than 70% of Indians are prone to general body maladies every 3 months. Quite often any people do not realize that the usual body ailment could lead to side effects to something really dreadful, 26% of the people surrenders to death because of

ignoring the early general body symptoms [1]. This could be a dangerous event for the population and can be alarming. Therefore recognizing or predicting the ailment at the earliest is important to maintain strategic distance from any unrequired losses. The present available systems are the systems that are either devoted to a certain sickness or are in the research phase for algorithms when it comes to a particular disease.

This situation is not only limited to India, but similar observations have been seen across the world, even in countries with great healthcare systems. According to reports more than two lakh people in China and one lakh in United States of America are dead every year due to the mistakes made while prescribing errors. In addition to this, various studies have been showing that almost lakhs of people die due to the medication errors [10]. These errors can be attributed to doctors, who write medicines for patients based on their career experience.

Technologies such as data mining and recommender systems provide opportunity for investigating show ways to inspect potential knowledge from historical records pertaining to diagnosis and help medical specialists to perform analysis on the clinical mistakes and endorse prescription accurately to reduce error in medication significantly.

In this paper, we have proposed a system which takes the symptoms as necessary inputs from the patient to predict the disease, which is followed by searching and provide correct medicine. The main motivation behind the paper is to correctly diagnose general and frequently occurring ailments that when left unseen could turn into fatal disease and recommend medicines and reduce manual errors. This paper deals with the design as well as implementation of a system which performs the function of prediction of diseases and the recommendation of medicines.

1.2 Existing System

The traditional way consists of doctors performing a patient's diagnosis and recommending medication by doctor's experience in his/her career, which might sometimes lead to the doctors prescribing wrong medicines or an overdose to patients, which causes severe side effects for the patients.

The patient, who is suffering with a particular disease, is predicted. No external support like medication or tracking the progression of the disease is included in these systems.

1.3 Proposed Design of System

In our paper, we have created a system which takes symptoms present in the patient as input to predict the disease, which is followed by giving the answers a correct medicine [8].Instead of going through the task of answering many questions which usually are a form of consultation, the user would have to just enter the symptoms present in the patient. In the dataset the symptoms related to medical data are stored. The dataset includes categories such as disease's name and list of associated symptoms. Here there is a need for the system to give response to the query entered by the user by deploying a machine learning model on dataset.

The symptoms of the patients are taken by the doctors for the continuous evaluation of vitals like heart rate, blood pressure, sugar level etc. for the analysis [9]. A doctor can search in the system or can fire questionnaires to the system. The system will respond according to the corresponding dataset. This system is mainly designed to help doctors integrate prediction modules and recommendation modules so that it can recommend medicines based on the respective disease.

Naive Bayes model, Random Forest algorithm and decision tree are used to achieve the result of predicting disease and recommending a medicine. Since we require efficient accuracy and high potency is important for such a system which predicts disease based on symptoms and recommends medicine, thus to get high accuracy and efficiency we need to asses some data preprocessing approaches.

Advantages

- Creates a medication system for supporting doctors in the diagnosis of diseases.
- Reduces the scope of errors made by doctors manually and increases the probability of avoiding errors.
- Takes historical medical data into account.

II. Literature Survey

Machine learning which has been the core area of AI has enabled computers to learn-by-self without being externally programmed which can practice to learn, develop, grow, and change. The ML learns data from the algorithms and makes simple and easy to find perceptive information without any programs. Many people have done research to build models which can predict diseases.

Darcy Davis, Nitesh V. Chawla, Nicholas Blumm, and Albert Laszlo published a paper titled "Predicting individual disease risk based on medical history,"[2] where there has been developed a system where the primary concern is to recognize the disease and take necessary action. This system uses Collaborative filtering and recommendation engine where patient's own medical history and similar patient's medical history are used to predict the risk that a patient might have from a particular disease .Use of ICD-9-CM codes are present to predict future disease risk. This step is necessary to note the progression of disease which is important for any healthcare initiative.

Md Aliul Islam Abir at [3] proposed work shows how efficiently k-means algorithm has been employed to construct a master system for identifying human disease by finding out the disease-symptoms data for improving the decision making systems that highly depend on huge data. Doctors and medical professionals use this for rectifying diagnosis. Also how the process of "essential feature selection is performed is also shown where an attribute which is below "significant level" are eliminated and rest of them are considered as significant features.

This paper [4] mention the use of recommender systems for making personalized recommendation with user-item interaction ratings and implicit feedback. A matrix factorization model has been constructed for predicting a personalized ranking over a set of items for an individual user with similarities among users and items. With the matrix as input, a deep structure learning architecture is presented in recognizing a common low dimensional space for the representations of users and items.

In this paper at [5] a Naive Bayesian approach is used for predicting causes with respect to the heart disease. The attributes or the causes acts as input for the Navies Bayesian machine learning model to predict the disease. The dataset collected is divided into two separate sets, 80% of the dataset is used for training and other is used for testing.

Here in this paper at [6] the use of tkinter for building end user applications by making use of event driven-scripts where event handlers and widgets can be constructed and the flow of execution maintained by the use of Python language which offers great ease to any developer. Also few necessary code lines have been mentioned which makes it useful for our paper.

Another paper "A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," by X. Huang and Q. Hu, described the term "medical retrieval" as the dominant way for knowledge exchanging and sharing [7]. Huang etal. Proposed a re-ranking model for promoting diversity in medical.

Few limitations exist in the previous papers only tracking the progression of the disease does not solve the problem or estimating the risk of the disease a recommender system module should be integrated along with prediction module. Also most of the papers related to "disease prediction" contain only one classifier model to calculate accuracy we propose to implement three models which are Decision Tree, Naive Bayesian and Random Forest.

In the above papers only the procedure for data preprocessing and data prediction module is present which describes how the entire process works but to put it into use for the end-user no interactive end-user application is mentioned therefore we propose to use, Tkinter a GUI python library to create the required end-user application window for a doctor.

III. Methodology

Database System Module

The diseases and the symptoms related to them are contained in first dataset. The motive is to predict the ailment the person might be suffering based on the symptoms that are shown by them. The attributes are 'Disease Id', 'Disease', 'Symptoms', and 'Count of Disease Occurrence'. 'Disease Id' helps in identifying each disease uniquely. The 'Count' attribute indicates us which diseases quite often.

Once the ailment has been predicted, the patient has been diagnosed correctly; the next step is to aid the recovery process by prescribing the right kind of medication. The second dataset contains attributes such a 'Disease Id', 'Drug' and 'Drug Rating'. The attribute 'Disease Id' acts as the link between the two separate processes of disease prediction and drug recommendation. The 'Rating' attribute is based on the user-feedback and provides information about the potential of the drug.

Data Preparation Module

The main purpose of this module is to clean the data as in reality data is raw and has many dissimilarities .Hence we need to perform this task of cleaning. Also it involves finding consistent and authenticated data. The symptoms are contained as significant features in the second dataset with each row representing the disease. Also the empty columns are filled with same values of the disease name.

After this, the proper disease name needs to be extracted from the elongated disease description containing a unique identifier for each disease. For instance, 'UMLS: C0010528_depression mental' is rewritten as 'depression mental'. . For example, 'UMLS: C0004031_depressive disorder' is modified to be 'depressive disorder' instead.
Many diseases can be associated with a single symptom. Here, the special character '^' is used to indicate the presence of more than one disease. In such cases, we split the given entry

around the character '^' and then make two separate entries of the different diseases showing the same symptom. Whereas some entries of disease attribute is present like this 'UMLS: C0010528_depression mental^ UMLS: C0004031_depressive disorder'. Hence a new dataset has to be created which has only single disease and single symptom in every row.
Since in machine learning algorithms dataset should be processed fast we convert the disease names and symptoms into numerical values. Another dataframe object might be created which possess exact same values of symptoms. This dataframe has all the possible symptoms as its column values or attributes. The rows values indicate the corresponding disease. If that disease is associated with a particular symptom, the respective entry in the data frame is given the value of '1' and '0' otherwise. Now the dataset is ready.

Disease Prediction Module

Disease Prediction Module deals with choosing the suitable algorithm to be deployed on the data. The data which has been cleaned is fragmented into two parts. The first part of the dataset is called the training data and it is used for developing the model. The second part of the dataset is called the test data, and is used as a reference to test the model. Models like Random Forest, Naive Bayes and decision tree have been applied on the training dataset. This was followed by testing of the model by applying it on the testing dataset. Later the accuracy was given. But, the accuracy turned out to be zero. The reason why this particular model behaved that way was due to the reason that it does not have any first-hand knowledge of disease prior to the testing of it. Three different classification algorithms are imported from the specific libraries and deployed on the dataset. After this, the model is evaluated on the testing dataset. The algorithm with highest accuracy is used to predict the diseases.

Recommendation Module

Recommendation Module is used to recommend the frequently used medicine for a particular ailment. This module particularly helps in assessing the quality of medicine which is most suitable and recommends you to use the required medicine after the prediction of the disease.
The drug dataset which consists of attributes Disease Id, Drug, and Rating is used. Multiple drugs can be used to for a single disease. The objective is to find the most suitable drug and suggest it. This is done on the basis of rating through collaborative filtering approach. After the ratings of all are ordered in a sequence for a particular disease and the drug with highest rating is used.

Model Evaluation

Model Evaluation Module tests the performance of the different prediction and recommendation models that are used on our dataset. Since it is a precarious task to analyze a model's accuracy such as how it works on the data that was not used for model building. Multiple diseases have same symptoms so for easily detecting every disease and its related symptom is stored in knowledge base. After using Naive Bayes and Random Forest the accuracy obtained is 84 percent which is quite low. The paper has used decision tree to increase the accuracy which is 92 percent.

Algorithms:

1. Database System Module

Steps:

1. Importing the dataset which consists of diseases and symptoms.
2. Import the dataset which consists of diseases and the corresponding drugs for it.

2. Data Preparation Module

Steps:

1) Preprocess the dataset which consists of diseases and symptoms.
   1.1) Fill in the values which are missing
2) Retrieve the names of diseases and the symptoms associated with it.
3) Process the names of diseases and the symptoms associated with it.
   3.1 Exclude the disease id and include the name of disease only.
   3.2 Extract the corresponding symptoms
   3.3 Get the Names of the disease.
   3.4 Get the Symptoms with respect to the Diseases
4) Obtain dummy values of corresponding symptoms of a given disease
5) For each of the disease name convert every one of them to a natural number .

3. Disease Prediction Module

Steps:

1) Training the disease dataset

1.1 Naive Bayes classifier

The original Naïve Bayesian strategy is based on the conditional probability and the occurrence of the maximum probability. The Bayesian Naive algorithm is as follows

Begin

- Initialization pc → Number of classes nca → Number of attributes Nu → Number of samples
- for each class Cj do Calculate prior probability $P(Ci) = \sum Cj / \sum Nu, j \in \{1,pc\}$
- for each class Cj do For each attribute Ai do Calculate the conditional probability $P(Ai|Cj) = \sum Cj$ with $Ai / \sum ji , j \in \{1,..,pc\}$ and $i \in \{1,..,nca\}$
- for each class Cj do Calculate the conditional probability of the tuple K i.e $P(K|Cj) = P(A1|Cj) * P(A2|Cj) * \ldots * P(Anca|Cj)$
- for each class Cj do Calculate the posterior probability of the tuple K i.e $P(Cj) * P(K|Cj)$
- Prediction If$((P(C p) * P(K|Cp)) > (P(Cq) * P(K|Cq)))$ Prediction → C p Else Prediction → Cq Where $p,q \in \{1,..,pc\}$ and $p \neq q$
- End

1.2 Decision Tree

Decision tree is a simpler, easy to implement classifier which needs no domain knowledge. The key benefit of this is that the decision tree approach can be extended to enormous data to be interpreted. Decision tree is a tree-like structure which consists of arcs, nodes, and branches.

A general algorithm for a decision tree can be described as follows:

1. Choose the best attribute / function. The best attribute is one that best separates or divides the data.
2. Ask the question which is important.
3. Follow the direction of answers.
4. Go to step 1 before the response arrives.

1.3 Random Forest Algorithm

Random forest Algorithm is a supervised machine learning technique that is based on ensemble learning. It takes average based on many of the decision trees applied on the different subset of the data and predict accordingly.

1.3.1 Select random K data points from the training set.

1.3.2 Build the decision trees associated with the selected data points (Subsets).

1.3.3Choose the number N for decision trees that you want to build.

1.3.4Repeat the steps mentioned in 1.3.1 and 1.3.2

2) Test all the models

3) Calculate accuracy

4. Recommendation module

1) Importing the dataset which consists of diseases and medicines.

2) Describe the dataset

3) For a particular disease with rating retrieve the one with highest rating.

4) Use the Disease Prediction Module to calculate the predicted disease and then map it to the dataset to get the corresponding medicine.

5. Model Evaluation

1) Compare the accuracy of all the three mentioned models

Accuracy = Number of Correct Predictions / Total Number of Predictions Made

2) Select the model with highest accuracy.

3) Display the results to the user.

6. User Interface

1) Create a Tkinter window object.

2) Add widgets such as buttons, text-boxes and option menu.

3) Create the required event handlers for all widget to invoke necessary functions.

4) Display the output to the end-user.

System Design

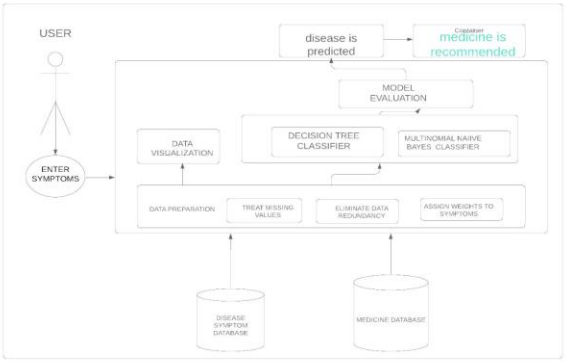


Figure 1: Architecture of System

The above Figure 1 describes the proposed architecture of our paper “disease prediction and medicine recommendation system”. It indicates how the architecture of system has been designed and shows the flow across different modules throughout the system .Since the advent of artificial intelligence there has been significant strides in prediction system and recommendation system which have found to be useful. Like many current recommender systems where they are used only foe e-commerce applications. Our system tries to provide an animated doctor for beginner doctors and patients in mistreatment right medication. Since high accuracy and reliability is important for such a symptom based disease prediction and medication recommender system.

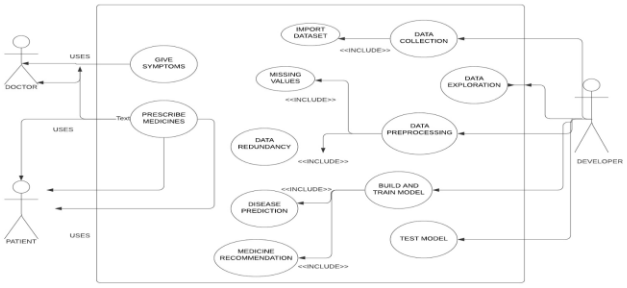


Figure 1.2. Use Case Diagram

The above Figure 1.2 shows the different tasks performed by different stakeholders of the system. There are 3 different types of users. They are 1) Patient 2) Developer and 3) Doctor. The tasks are performed in the system and are represented by ovals which are actors. . Dashed lines represent actions performed by user.

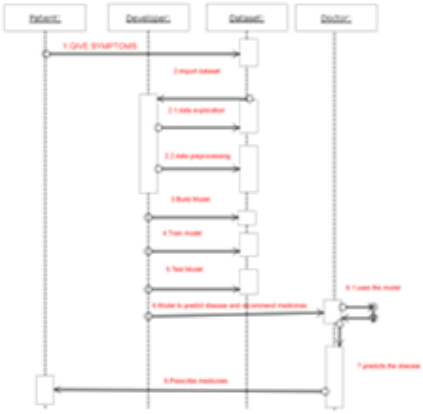


Figure 1.3. Sequence Diagram

The Figure 1.3 describes the direction of messages in the system among different stakeholders of the system. The objects which are invoke by doctor are denoted by rectangles at top.The lines dropping from boxes which are dashed represent the stakeholder of the entire system at particular point of time. Some boxes present on the dotted line imply that they are events and the transfers of messages across those events are represented by these arrows.

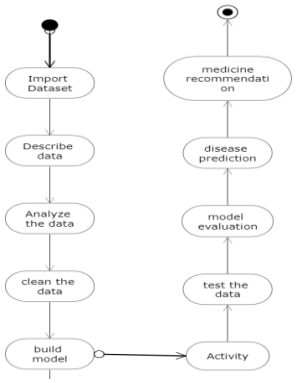
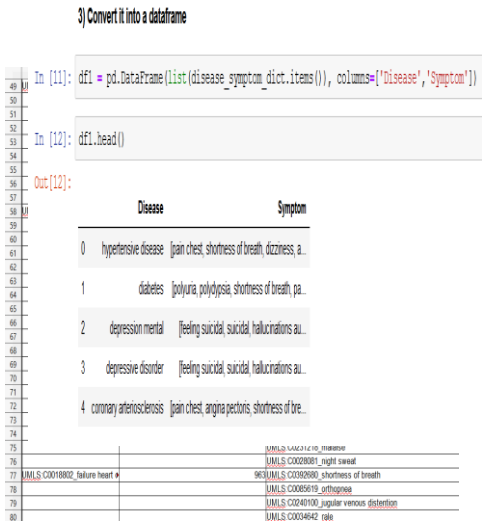


Figure 1.4. Activity Diagram

The above Figure 1.4 describes the activity diagram of the system represents the message direction of activities in our paper. Dot at the starting indicates the ending of the start and point with circle, and an action is represented as a curvy-edged rectangle. These are the flow of behaviors from starting till ending.



3) Convert it into a dataframe

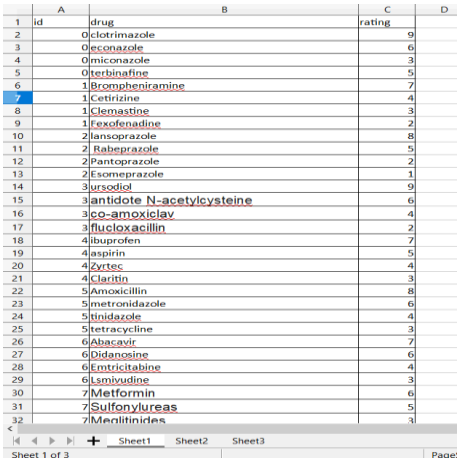
```
In [11]: df1 = pd.DataFrame(list(disease_symptom_dict.items()), columns=['Disease','Symptom'])

In [12]: df1.head()

Out[12]:
```

	Disease	Symptom
0	hypertensive disease	[pain chest, shortness of breath, dizziness, a...
1	diabetes	[polyuria, polydipsia, shortness of breath, pa...
2	depression mental	[feeling suicidal, suicidal, hallucinations au...
3	depressive disorder	[feeling suicidal, suicidal, hallucinations au...
4	coronary arteriosclerosis	[pain chest, angina pectoris, shortness of bre...

UMLS:C0028081_night sweat
963 UMLS:C0392680_shortness of breath
UMLS:C0085619_orthopnea
UMLS:C0240100_jugular venous distention
UMLS:C0034642_rale
UMLS:C0016802_failure heart

Figure 2. Disease Dataset

The above Figure 2 represents the dataset which contains diseases and symptoms. It has significant attributes like Disease, count of disease occurrence and symptoms. The count of the disease occurrence \are also taken for predicting the ailment based on the symptoms that have been provided. This data set is mainly used for predicting the chances of a particular ailment for a given set of symptoms provided by patient.



	A	B	C	D
1	id	drug	rating	
2	0	clotrimazole	9	
3	0	econazole	6	
4	0	miconazole	3	
5	0	terbinafine	5	
6	1	Brompheniramine	7	
7	1	Cetirizine	4	
8	1	Clemastine	3	
9	1	Fexofenadine	2	
10	2	lansoprazole	8	
11	2	Rabeprazole	5	
12	2	Pantoprazole	2	
13	2	Esomeprazole	1	
14	3	ursodiol	9	
15	3	antidote N-acetylcysteine	6	
16	3	co-amoxiclav	4	
17	3	flucloxacillin	2	
18	4	ibuprofen	7	
19	4	aspirin	5	
20	4	Zyrtec	4	
21	4	Claritin	3	
22	5	Amoxicillin	8	
23	5	metronidazole	6	
24	5	tinidazole	4	
25	5	tetracycline	3	
26	6	Abacavir	7	
27	6	Didanosine	6	
28	6	Emtricitabine	4	
29	6	Lsmivudine	3	
30	7	Metformin	6	
31	7	Sulfonylureas	5	
32	7	Meglitinides	3	

Sheet1 Sheet2 Sheet3
Sheet 1 of 3

Figure 2.2. Medicine Dataset

The above Figure 2.2 shows the MEDICINE dataset which contains significant attributes like id of the ailment, drug, rating of the drug. The unique id attribute refers to the id of the disease in the symptom dataset. The drug attribute includes the drugs used for the given disease.The rating attribute tells us about the feedback of its effectiveness of the drug.

3) Convert it into a dataframe

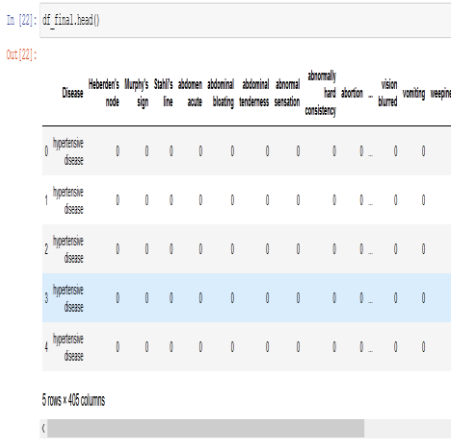
```
In [11]: df1 = pd.DataFrame(list(disease_symptom_dict.items()), columns=['Disease','Symptom'])

In [12]: df1.head()
```

Out[12]:

	Disease	Symptom
0	hypertensive disease	[pain chest, shortness of breath, dizziness, a...
1	diabetes	[polyuria, polydipsia, shortness of breath, pa...
2	depression mental	[feeling suicidal, suicidal, hallucinations au...
3	depressive disorder	[feeling suicidal, suicidal, hallucinations au...
4	coronary arteriosclerosis	[pain chest, angina pectoris, shortness of bre...

Figure 2.3. Data Preprocessing

The above Figure 2.3 is showing the preprocessing of the disease dataset to build the model by performing data exploration by extracting disease names, symptoms names, including disease ids and then obtaining the duplicate values of corresponding symptoms of a particular disease and finally considering the disease ids and their symptoms for processing it faster.



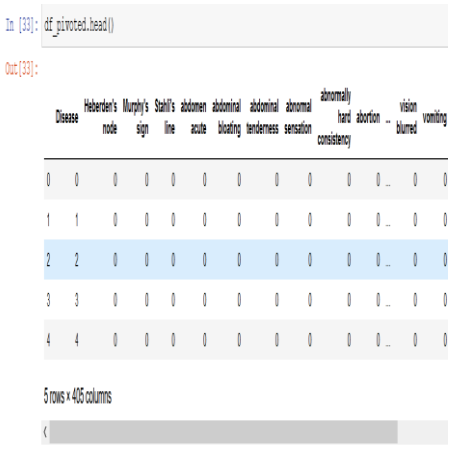
```
In [22]: df_final.head()
```

Out[22]:

	Disease	Heberden's node	Murphy's sign	Stahli's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnormally hard consistency	abortion	...	vision blurred	vomiting	weepin
0	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	
1	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	
2	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	
3	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	
4	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 405 columns

Figure 2.4.1. Processed dataset of disease with associated symptoms



In [33]: df_pivoted.head()

Out[33]:

	Disease	Heberden's node	Murphy's sign	Stahl's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnormally hard consistency	abortion	...	vision blurred	vomiting
0	0	0	0	0	0	0	0	0	0	0	...	0	0
1	1	0	0	0	0	0	0	0	0	0	...	0	0
2	2	0	0	0	0	0	0	0	0	0	...	0	0
3	3	0	0	0	0	0	0	0	0	0	...	0	0
4	4	0	0	0	0	0	0	0	0	0	...	0	0

5 rows × 405 columns

Figure 2.4.2. Processed dataset of disease with associated symptoms after converting disease names to IDs

The below Figure 2.4.3 is showing converting the disease names into disease ID's for processing fastly by the machine learning model.

In [27]:
```
df_pivoted = df_final.groupby("Disease").sum()
df_pivoted = df_pivoted.reset_index()
df_pivoted.head()
```

Out[27]:

	Disease	Heberden's node	Murphy's sign	Stahl's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnormally hard consistency	abortion	...	vision blurred	vomiting	weepi
0	Alzheimer's disease	0	0	0	0	0	0	0	0	0	...	0	0	
1	HIV	0	0	0	0	0	0	0	0	0	...	0	0	
2	Pneumocystis carinii pneumonia	0	0	0	0	0	0	0	0	0	...	0	0	
3	accident cerebrovascular	0	0	0	0	0	0	0	0	0	...	0	0	
4	acquired immuno-deficiency syndrome	0	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 405 columns

Figure 2.4.3. Distinct dataset of disease with unique ID.

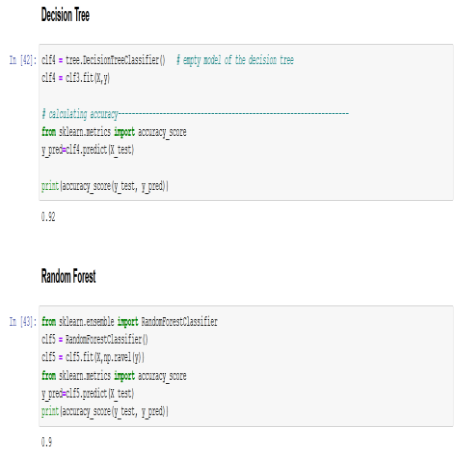
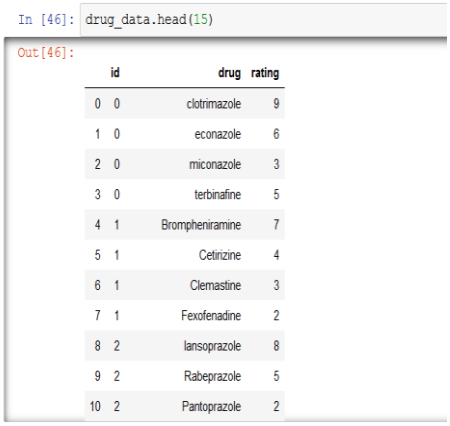


Figure 3. Accuracy of Decision Tree and Random Forest

The above Figure 3 mentioned figure is showing the accuracy of the decision tree model for prediction of the disease. The accuracy of all the models is compared and the model with highest accuracy is selected which is decision tree with 92%.



	id	drug	rating
0	0	clotrimazole	9
1	0	econazole	6
2	0	miconazole	3
3	0	terbinafine	5
4	1	Brompheniramine	7
5	1	Cetirizine	4
6	1	Clemastine	3
7	1	Fexofenadine	2
8	2	lansoprazole	8
9	2	Rabeprazole	5
10	2	Pantoprazole	2

Figure 4. Rating of Drugs

The above Figure-4 shows the highest rated drug for a particular disease. The disease is represented in the form of id.The medicine dataset contains the disease id, medicines along with the ratings, the only medicine is recommended to the user which has the highest rating i.e., most accurate and effective medicine.

```
[49]: highest_rated_drug
t[49]:
```

	index	id	drug	rating
0	0	0	clotrimazole	9
1	4	1	Brompheniramine	7
2	8	2	lansoprazole	8
3	12	3	ursodiol	9
4	16	4	ibuprofen	7
5	20	5	Amoxicillin	8
6	24	6	Abacavir	7
7	28	7	Metformin	6
8	32	8	Ampicillin	9
9	36	9	Albuterol	8
10	40	10	hydrochlorothiazide	8
11	44	11	Grenil	8
12	48	12	Fexmid	9
13	52	13	clonazepam	7
14	56	14	antidote N-acetylcysteine	9
15	60	15	Chloroquine phosphate	9
16	64	16	T.cetrizen	7
17	68	17	Tylenol	7
18	72	18	ciprofloxacin	8

Figure 5. Highest Rated Drugs

The above Figure-5 shows only medicine which has the highest rating is displayed which is most effective and has good accuracy.

IV. Result

The proposed work is implemented by the use of python and it's modules such as pandas, NumPy and Tkinter, a GUI library. The Table below shows the observations from the proposed system for different test cases.

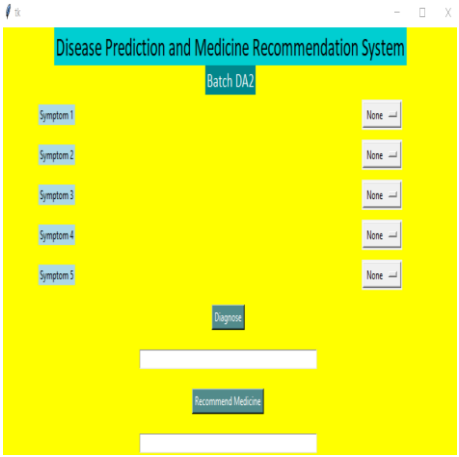


Figure 6. User Interface

The above Figure 6 is the user interface of our work "Disease Prediction and Medicine Recommendation system". It has five drop down options for symptoms and two buttons "Diagnose" and "Recommend Medicine".

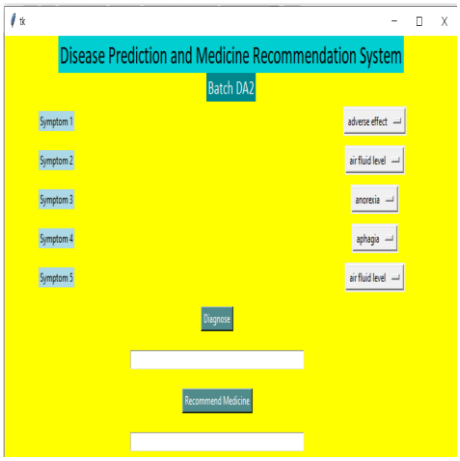


Figure 6.2. Symptoms as Input

In the above Figure 6.2 all symptoms are selected which are present in the patient.

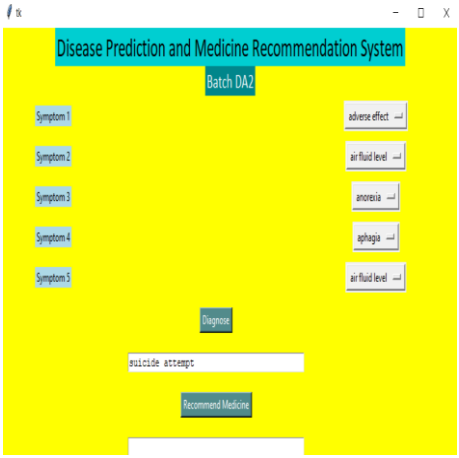


Figure 6.3. Prediction of Disease

In the above figure 6.3 it is shown that after clicking the “Diagnose” button the disease predicted is displayed.

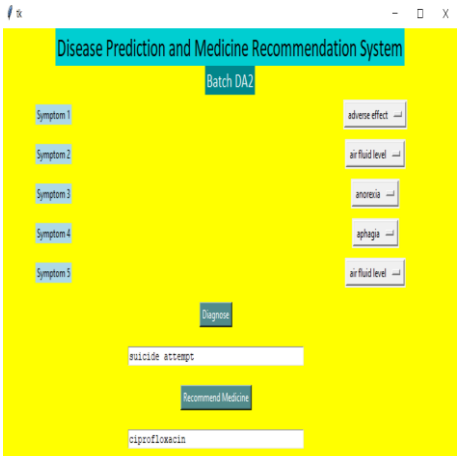


Figure 6.4. Medicine Recommendation

In the above figure 6.4 it is shown after clicking the "Recommend Medicine" button .The required medicine which needs to be taken for the disease is displayed.

Test Cases

All the below test cases have been executed with no defects.

Table 1.Test cases for data exploration and database module

S.NO	TEST CASE ID	TEST STEPS	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
1.	ID-1	Inserting the dataset	Dataset is inserted without any errors	Dataset is inserted	Pass
2.	ID-2	Inserting the dataset	Dataset is inserted without any errors	Dataset inserted with wrong path name	Fail
3.	ID-3	Describe the dataset	Reveals the details of dataset	Dataset is empty	fail
4.	ID-4	Describe the dataset	Reveals the details of dataset	All the attributes and details of dataset are shown	pass
5.	ID-5	Check the missing values for condition	Values not present	Values not present	pass

6.	ID-6	Check the missing values for condition	Values not present	Values not present for condition	fail
7.	ID-7	Check the redundant data	No redundant data is available	Redundancy present	fail
8.	ID-8	Checking if the disease name is valid or not	Valid disease name	Disease name is invalid	fail

Table 2. Test cases for disease prediction module

S.NO	TEST CASE ID	TEST STEPS	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
1.	ID-1	Bedridden, egophony, asthenia	neuropathy	Neuropathy	pass
2.	ID-2	Shortness of breath, orthopnea, cough, wheezing	Heart failure	Pneumonia	fail
3.	ID-3	Wheezing,cough,shortness of breath, chest tightness	asthma	asthma	pass
4.	ID-4	Awakening early,agitation,adverse reaction	hepatitis	hepatitis	pass

Table 3. Test cases for medicine recommendation module

S.NO	TEST CASE ID	TEST STEPS	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
1.	ID-1	Bedridden, egophony, asthenia	Dab witch hazel	Dab witch hazel	pass
2.	ID-2	Shortness of breath, orthopnea, cough, wheezing	Aspirin	Levaquin	fail
3.	ID-3	Wheezing, cough, shortness of breath, chest tightness	Theophylline	Theophylline	pass
4.	ID-4	Awakening early, agitation, adverse reaction	dramamine	Dramamine	pass

Table 4. Model Accuracy for respective algorithms

Model Name	Accuracy
Doctors Manual Prediction	65 %
Naïve Bayes	90%
Decision Tree	92%
Random Forest	90 %

The Doctors who have experience at many hospitals treating various kinds of diseases and having a great deal of knowledge about the type of medicines correctly predicted 65% of diseases only where they asked the patient limited number of of questions about his or her illness. [13]. Hence we go with decision tree algorithm as it has accuracy of 92% and in terms of large dataset random forest has lesser probability of succeeding than decision tree.Also for shorter dataset Naïve Bayes gives accuracy of 84%.

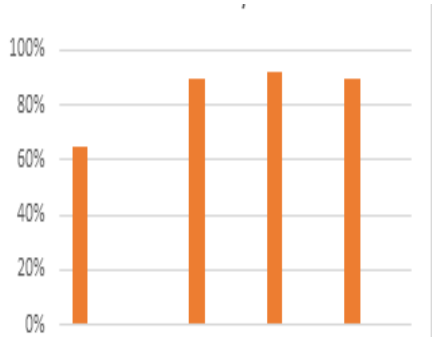


Figure 7. Accuracy comparison of various models and Doctor's Manual Prediction

V. Conclusion and Future Enhancements

Finally we conclude that we have created a system where diseases prediction takes place by inputting symptoms and recommend appropriate medicine to tackle manual errors caused by doctor while giving medication keeping in mind their experience in this profession and thus reducing human casualties in society .This is built with distinct models which are data preprocessing, inputting symptoms ,analyzing them and predicting disease and recommending medicine. Since the required model provides accuracy of up to 92 % it is reliable in the health infrastructure.

We would like to add some more interesting features in future that would increase the productivity and the reliability of the proposed system more. We will add in future the recommendation of treatments for diseases which are not only dependent on medicines.

References

1. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K.Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map",IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.
2. Darcy Davis, Nitesh V. Chawla, Nicholas Blumm, and Albert Laszlo,January 2018, "Predicting individual disease risk based on medical history",in 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2018
3. Md Aliul Islam Abir , 06th July 2019 , "DISEASE PREDICTION THROUGH SYNDROMES USING K-MEANS ALGORITHM" Department of Computer Science and Engineering Daffodil International University ,DHAKA ,BANGLADESH
4. Hong-Jian Xue, Xin-Yu Dai, Jianbing Zhang, Shujian Huang, Jiajun Chene,2017 Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) ,Nanjing University, Nanjing 210023, China
5. Anjan Nikhil Repaka,Sai Deepak Ravikanti,Ramya G Franklin ,10 October 2019, 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI),Tirunelveli, India

6. Shridhar B Dandin , Mireille Ducasse,DECEMBER 2020 , "Tkinter - A Boon for Designing GUI in Applications like ComVisMD" ,Department of Computer Science Engineering Sarala Birla University, Ranchi, India
7. Xiangji Huang,Qinmin Hu,July 2009 ,A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval in 32nd International ACM SIGIR conference on Research and Development in information retrieval
8. Rahul Isola, Rebeck Carvalho and Amiya Kumar Tripathy.Knowledge discovery in Medical system by using Differential Diagnosis, AMSTAR and K-NN, IEEE Transaction on Information Technology in Biomedicine, Vol.16, No.6, November 2012.
9. S.Fox and M. Duggan. Health online 2013.Pew Internet and American LifePaper.http://pewinternet.org/Reports/2013/Health-online.aspx,2013.
10. Ashish Chhabbi, LakhanAhuja, SahilAhir, and Y. K. Sharma,19 March 2016,"Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Advent Technology,E-ISSN:2321-9637,Special Issue National Conference "NCPC-2016", pp. 104-106.
11. Varun Kumar –Nisha Rathee 2011, March .Knowledge discovery from database Using an integration of clustering and classification. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011).
12. Durairaj M Assistant Professor School of Comp. Sci., Engg. &Applications, BharathidasanUniversity,Trichy,TN,Indiadurairaj, 2015, February PREDICTION OF ACUTE MYELOID LEUKEMIA CANCER USING DATAMINING-A SURVEY. (International Journal of Emerging Technology and Innovative Engineering Volume I, Issue 2, February 2015 ISSN: 2394 - 6598).
13. Takanor iUehara, Masatomi Ikusaka,[1] Yoshiyuki Ohira,[1] Mitsuyasu Ohta, Kazutaka,Noda,[1] Tomoko,Tsukamoto,[1] Toshihiko Takada,[1] and Masahito Miyahara[1] ,2013 Dec 6,Accuracy of diagnoses predicted from a simple patient questionnaire stratified by the duration of general ambulatory training: an observational study in International Journal of General Medcine,US National Library of Medicine.