Disease Prediction

Mentor

Dr.Neha Nandal (Associate Professor)

Mohammad Khaja Faizan	: 19241A05W5
K.Rajinikanth	: 19241A05W2
D.Rohit Rajan	: 19241A05U9
Vasu Sena Gunda	: 19241A05V6

Abstract

Now-a-days, people face various diseases due to environmental conditions and their living habits. So the prediction of a disease at an earlier stage becomes an important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctors. The correct prediction of disease is the most challenging task.Disease prediction is a way to recognize patient health by applying data mining and machine learning techniques on patient treatment history. With the help of a disease prediction system, it was possible to diagnose people based on symptoms. Disease prediction systems provide only possible outcomes it does not guarantee that it will predict the disease correctly. But it has significantly higher accuracy for predicting possible diseases.

Abstract

We design a disease prediction system using multiple ML algorithms.Based on the symptoms of an individual,the disease prediction system gives the output as the disease that the individual might be suffering from.Our disease prediction system can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved.

Software Requirements :-

Technology/Language	:- Python
Operation system	:- Windows 7 or above
IDE	:- Visual Studio Code.
Source	:-Data set from kaggle
Libraries	:- pandas , numpy,Tkinter and sklearn.

Hardware Requirements:-

Processor	:- Intel core i3 or above , AMD Ryzen 3 or above
hard disk	:-100 GB or above.
Ram	:- 4 GB or above

SCOPE

The scope of this project is primarily on the performance analysis of disease prediction approaches using different variants of supervised machine learning algorithms. Disease prediction and in a broader context, medical informatics, have recently gained significant attention from the data science research community in recent years. This is primarily due to the wide adaptation of computer-based technology into the health sector in different forms (e.g., electronic health records and administrative data) and subsequent availability of large health databases for researchers. These electronic data are being utilised in a wide range of healthcare research areas such as the analysis of healthcare utilisation , measuring performance of a hospital care network , exploring patterns and cost of care, developing disease risk prediction model , chronic disease surveillance, and comparing disease prevalence and drug outcomes.

Existing System

In the existing system the patient has to undergo several stages to get know the disease he is suffering from.

Initially the patient has to get doctor’s appointment to meet the doctor

The he has to undergo several basic treatments in the hospital which is both the time consuming and cost efficient

After undergoing the doctor identifies the disease the patient is suffering from

So in Existing system both patient and doctor’s time will be consuming

Disadvantages of Existing System

1. To undergo all types of test like CT Scanning , Endoscopy etc... requires all a lot of time .So Existing system is time consuming

2. Undergoing all those test may not gives accurate accuracy because in the process human errors may occurs since human is involving

3. To undergo all these tests it is not economic efficient

Proposed System

Most of the chronic diseases are predicted by our system. It accepts the structured type of data as input to the machine learning model. This system is used by end-users i.e. patients/any user. In this system, the user will enter all the symptoms from which he or she is suffering. These symptoms then will be given to the machine learning model to predict the disease. Algorithms are then applied to which gives the best accuracy. Then System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. Naïve Bayes algorithm is used for predicting the disease by using symptoms, for classification KNN algorithm is used, Logistic regression is used for extracting features which are having most impact value, the Decision tree is used to divide the big dataset

Advantages of Proposed System

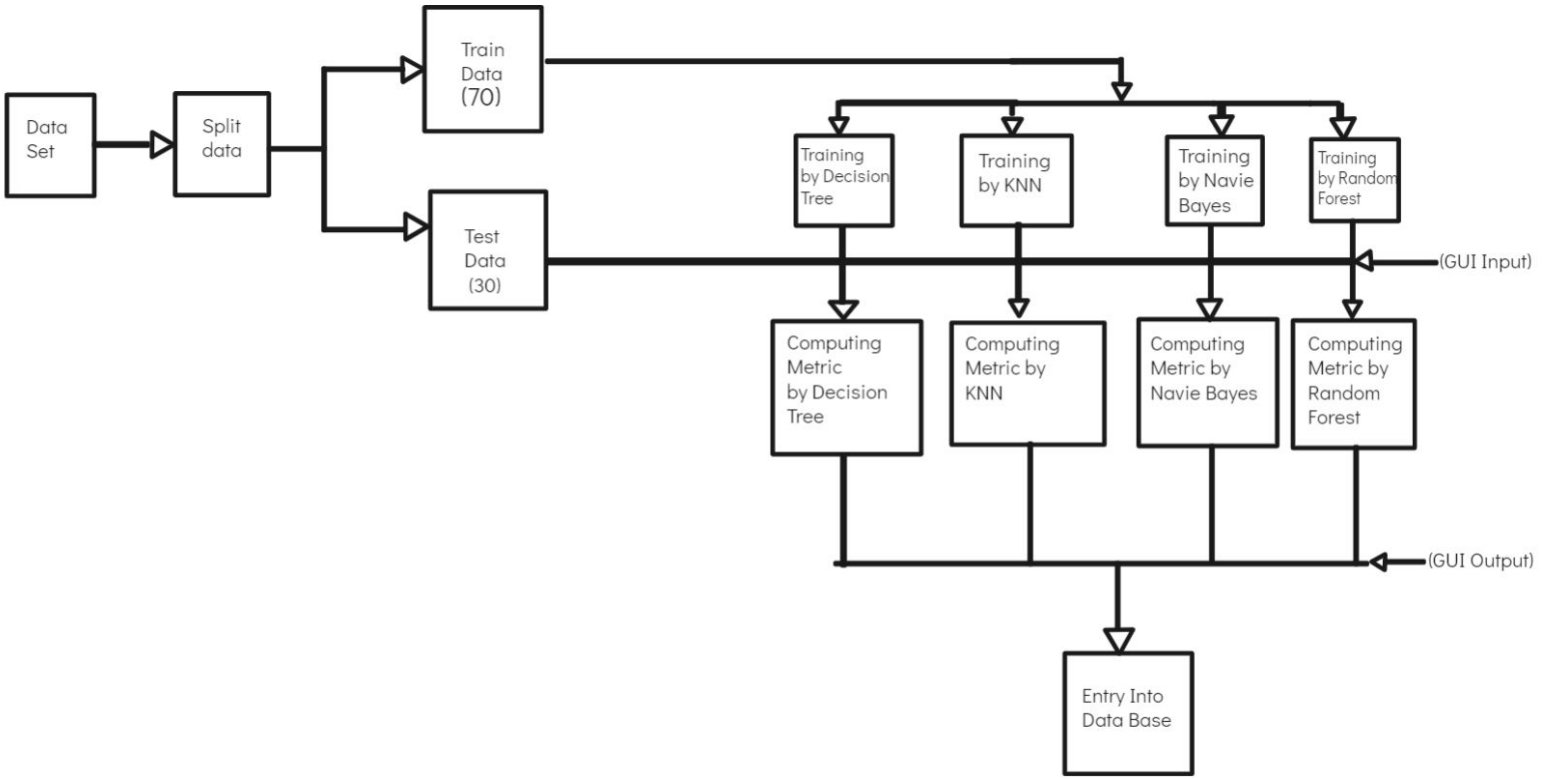
1. It can identify patients at risk of disease or health conditions

2. Clinicians can then take appropriate measures to avoid or minimise the risk and in turn, improve quality of care and avoid potential hospital admission

3. So it is economic efficient and time consuming

4. Disease prediction has the potential to benefit stakeholders such as the government and health insurance companies.

Methodology



Algorithms Used

- Decision Tree
- KNN
- Naive Bayes
- Random Forest

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

KNN

K Nearest Neighbour is a supervised learning algorithm. It is a basic yet essential algorithm.

It finds extensive use in pattern finding and data mining.

The next step is to calculate the distance between the data points who class is to predicted and all the training data points.

It works by finding a pattern in data which links data to results and it improves upon the pattern recognition with every iteration.

We have used K Nearest Neighbour to classify our dataset

Naive Bayes

Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems.

We can use probability to make predictions in machine learning. Perhaps the most widely used example is called the Naive Bayes algorithm. Not only is it straightforward to understand, but it also achieves surprisingly good results on a wide range of problems

Bayes' Theorem provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge.

P(class|data) = (P(data|class) * P(class)) / P(data)

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification.

Steps involved in random forest algorithm:

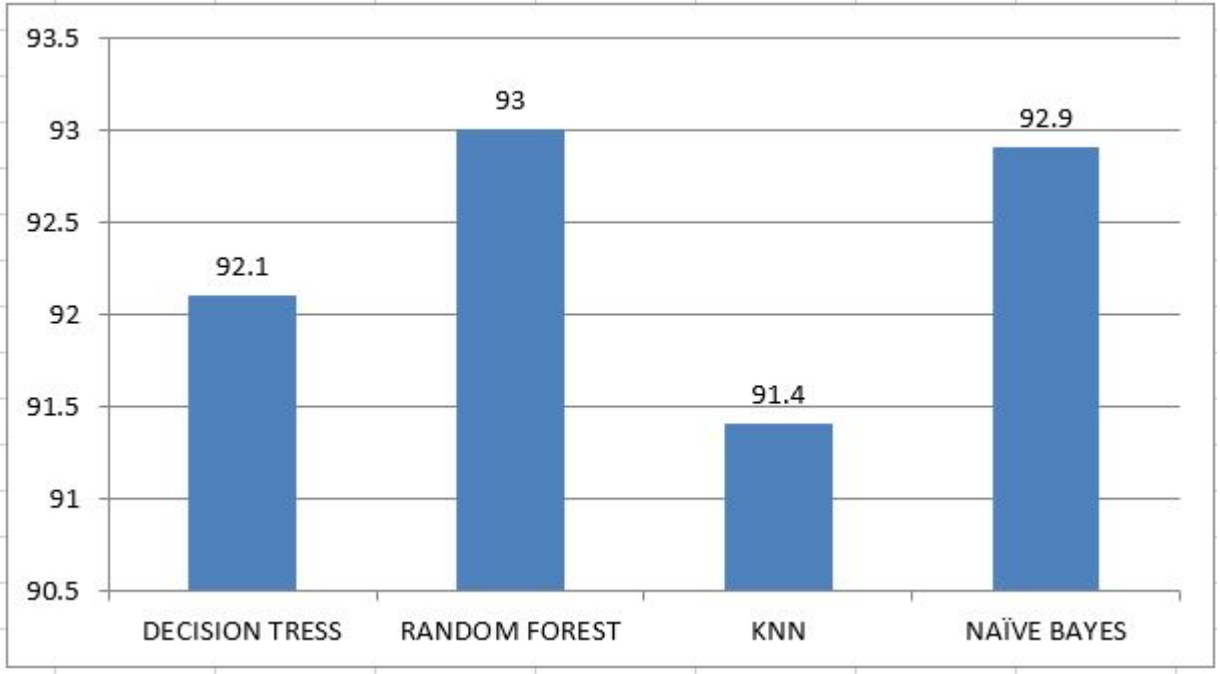
Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Accuracy of Algorithms



Symptoms

- Back_pain
- Mild_fever
- Loss_of_smell
- Muscle_weakness
- Internal_itching
- Skin_peeling
- Fast_heart_rate
- Neck_pain
- painful_walking
- Blackheads
- Knee_pain
- Chest_pain
- Congestion
- Muscle_pain
- Watering_from_eyes
- Visual_disturbances
- Obesity
- Swollen_legs
- .....

Total of 106 Symptoms were taken in the data set

Diseases

- Diabetes
- Bronchial Asthma
- Hypertension
- Dengue
- Typhoid
- Heartattack
- Allergy
- Common Cold
- Pneumonia
- Acne
- Malaria
- Jaundice
- Bronchial Asthma
- ....

Total of 42 Diseases were taken in the data set

Training Data

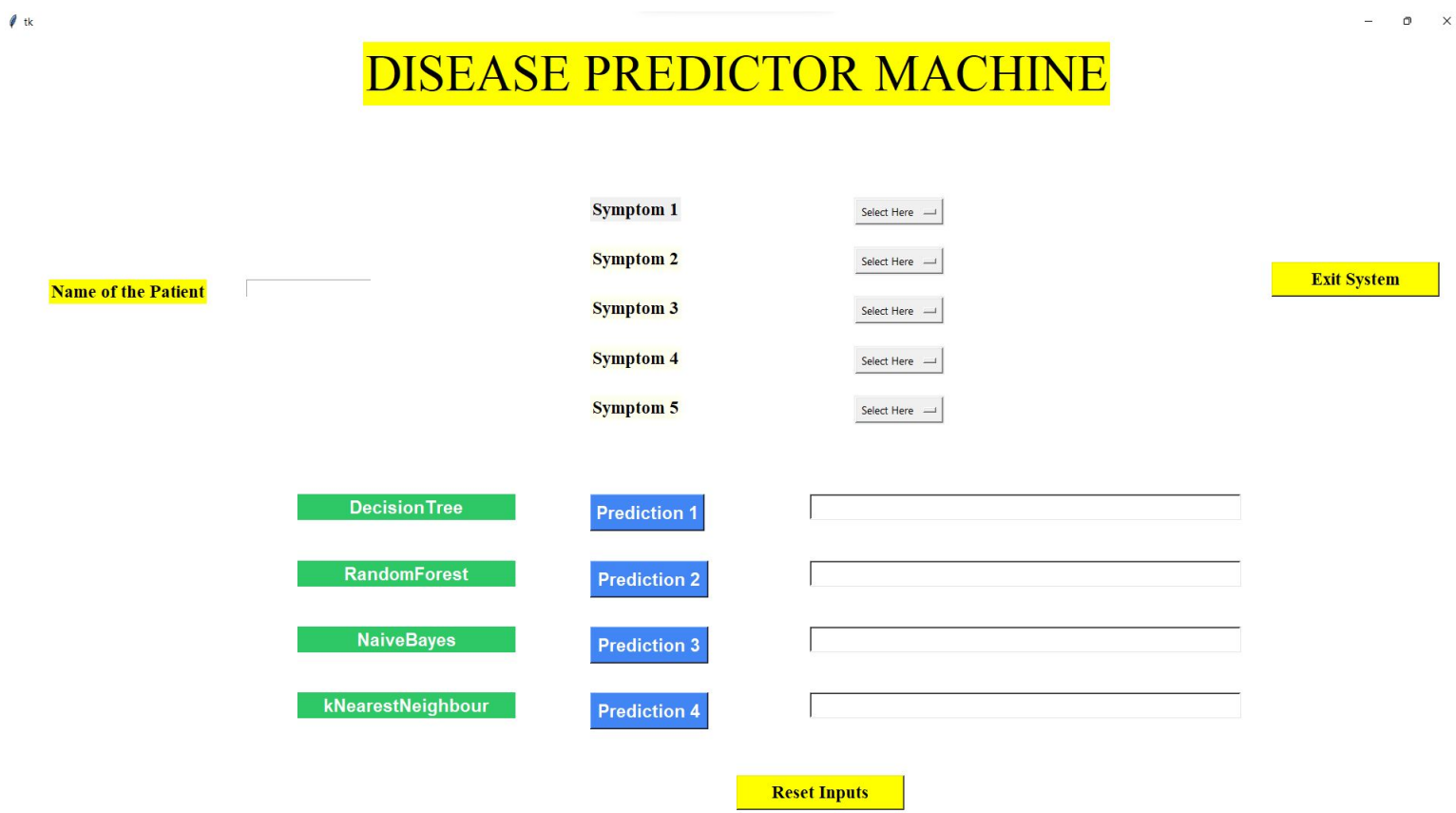
```
#Reading the training .csv file
df=pd.read_csv("training.csv")

#Replace the values in the imported file by pandas by the inbuilt function replace in pandas.

df.replace({'prognosis':{'Fungal infection':0,'Allergy':1,'GERD':2,'Chronic cholestasis':3,'Drug Reaction':4,
    'Peptic ulcer diseae':5,'AIDS':6,'Diabetes ':7,'Gastroenteritis':8,'Bronchial Asthma':9,'Hypertension ':10,
    'Migraine':11,'Cervical spondylosis':12,
    'Paralysis (brain hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken pox':16,'Dengue':17,'Typhoid':18,'hepatitis A':19,
    'Hepatitis B':20,'Hepatitis C':21,'Hepatitis D':22,'Hepatitis E':23,'Alcoholic hepatitis':24,'Tuberculosis':25,
    'Common Cold':26,'Pneumonia':27,'Dimorphic hemmorhoids(piles)':28,'Heart attack':29,'Varicose veins':30,'Hypothyroidism':31,
    'Hyperthyroidism':32,'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,
    '(vertigo) Paroymsal  Positional Vertigo':36,'Acne':37,'Urinary tract infection':38,'Psoriasis':39,
    'Impetigo':40}},inplace=True)
```

Testing Data

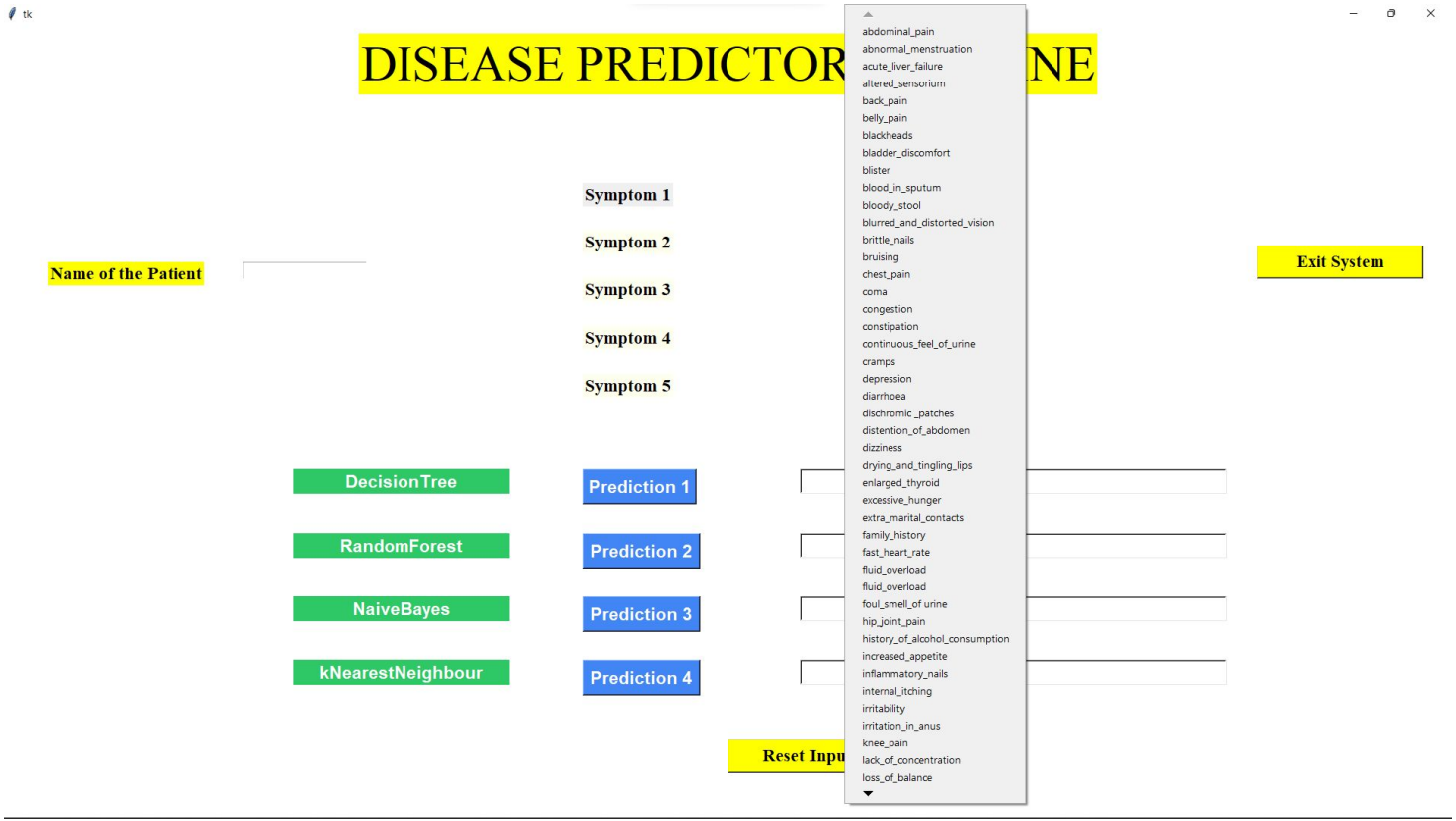
```
#Reading the  testing.csv file
tr=pd.read_csv("testing.csv")

#Using inbuilt function replace in pandas for replacing the values

tr.replace({'prognosis':{'Fungal infection':0,'Allergy':1,'GERD':2,'Chronic cholestasis':3,'Drug Reaction':4,
    'Peptic ulcer diseae':5,'AIDS':6,'Diabetes ':7,'Gastroenteritis':8,'Bronchial Asthma':9,'Hypertension ':10,
    'Migraine':11,'Cervical spondylosis':12,
    'Paralysis (brain hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken pox':16,'Dengue':17,'Typhoid':18,'hepatitis A':19,
    'Hepatitis B':20,'Hepatitis C':21,'Hepatitis D':22,'Hepatitis E':23,'Alcoholic hepatitis':24,'Tuberculosis':25,
    'Common Cold':26,'Pneumonia':27,'Dimorphic hemmorhoids(piles)':28,'Heart attack':29,'Varicose veins':30,'Hypothyroidism':31,
    'Hyperthyroidism':32,'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,
    '(vertigo) Paroymsal  Positional Vertigo':36,'Acne':37,'Urinary tract infection':38,'Psoriasis':39,
    'Impetigo':40}},inplace=True)
```

GUI

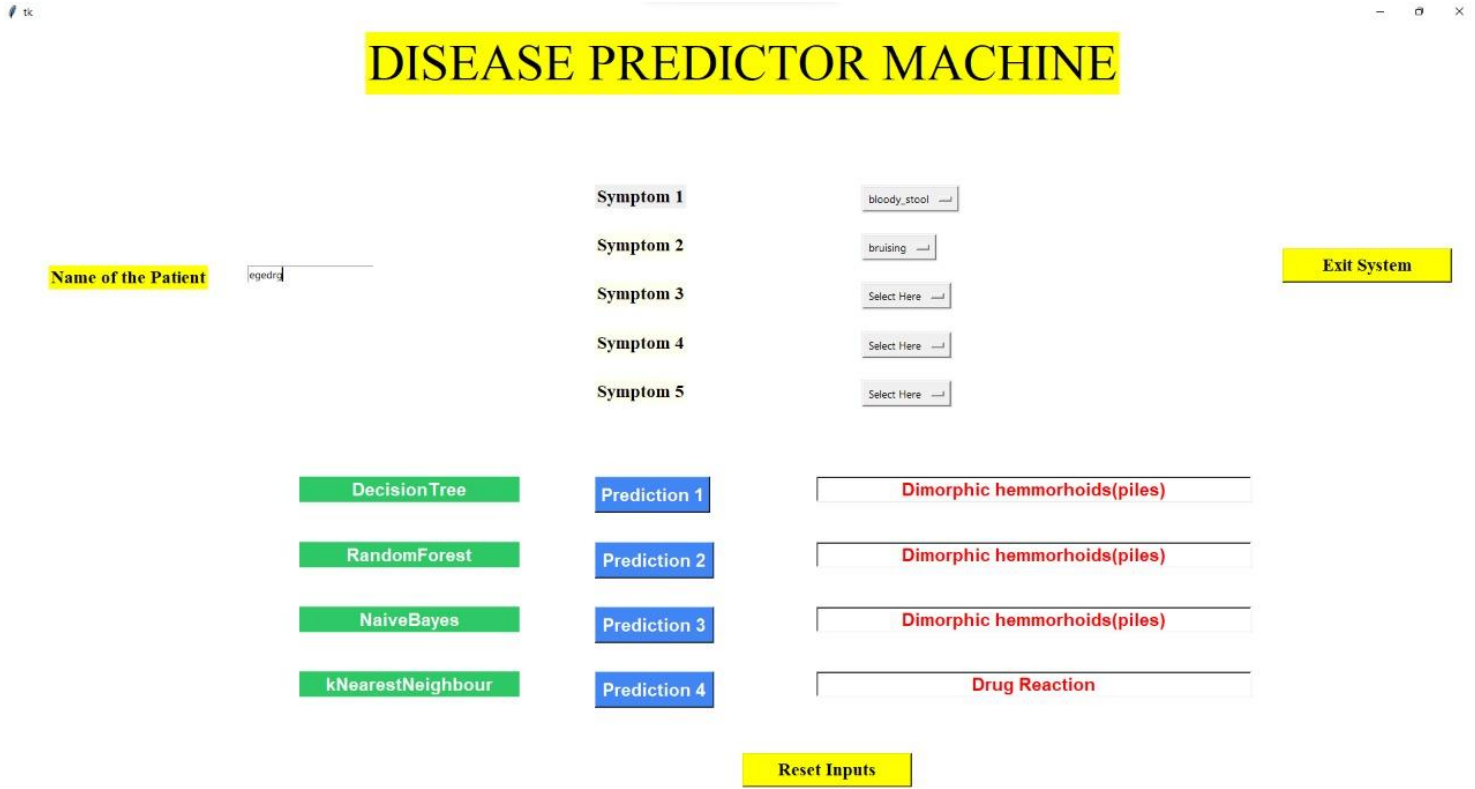
Before giving Symptoms as inputs



Symptoms Selection



Final Output



Storing In DataBase

	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	egedrg	bloody_stool	bruising	Select Here	Select Here	Select Here	Dimorphic hemmorhoids(piles)
2	sdvsd	blurred_and_distorted_vision	chest_pain	Select Here	Select Here	Select Here	GERD
3	sdfs	bruising	chest_pain	Select Here	Select Here	Select Here	GERD
4	sdv	bruising	chest_pain	Select Here	Select Here	Select Here	Varicoseveins
5	sdfs	chest_pain	chest_pain	Select Here	Select Here	Select Here	GERD
6	svs	bruising	congestion	Select Here	Select Here	Select Here	GERD

Conclusion

We set out to create a system which can predict disease on the basis of symptoms given to it.

Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff.

We were successful in creating such a system and use 4 different algorithm to do so

On an average we achieved accuracy of ~92%. Such a system can be largely reliable to do the job.

Creating this system we also added a way to store the data entered by the user in the database which can be used in future to help in creating better version of such system.

References

- Data Set from Kaggle

Thank You