NITTE
EDUCATION TRUST

N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
: 08258 - 281039 – 281263, Fax: 08258 – 281265

# Disease Prediction System

**Shreema S Suvarna[1], Shreya [2], Smitha [3],Sonali Suresh Shetty[4].**

---------------------------------------------------------------------***---------------------------------------------------------------

**Abstract -** *The world is moving with a fast speed and in order to keep up with the whole world we tend to ignore the symptoms of disease which can affect our health to a large extent. Many working professional's get heart attacks, bad cholesterol, eye disease and they are unable to treat it at the right time as they are busy coping up with progressive world. God has granted each and every individual a beautiful gift called life, so it is our responsibility to live our life to fullest and try to stay safe from the dangers of the world. So we have developed a logistic regression model with the help of machine learning algorithms like decision tree, random forest,k-nearest neighbour and naïve Bayes which take into account the symptoms felt by a person and according to those symptoms it predicts the disease which the person can be suffering from. It saves time as well as makes it easy to get a warning about your health before it's too late.*

***Key Words*: k-nearest neighbor, Decision tree, Random Forest, Naive Bayes Algorithm**

## INTRODUCTION

Our project is based on disease prediction according to the symptoms shown by the patient. This model which we have built comes under the umbrella of data analysis. Prediction of disease by looking at the symptoms is an integral part of treatment. In our project we have tried to accurately predict a disease by looking at the symptoms of the patient.We have used 4 different algorithms for this purpose and gained an accuracy of 92-95%.Such a system can have a very large potential in medical treatment of the future. We have also designed an interactive interface to facilitate interaction with the system. We have also attempted to show and visualize the result of our study and this project.

For this we are using python as a platform to run our machine learning algorithms. The first step to any analysis is to decide the problem we want to solve. Then getting the dataset to work on .Then we visualize our data with the help of scatter plot or any different plot and see it on an excel file. By doing this we can reduce redundancy in our data i.e. outliers, missing values etc. Then we treat our data by replacing the missing values, as python is a case sensitive programming language we transform all the letters into capital. Creating dummy variables to sort our data into mutually exclusive categories also means the no of dummy variables should be less than the no of categories of a qualitative variable. Also many people make the mistake of replacing the missing values with the mean of that variable but by doing so you can miss very important variations in the data.

## 2. LITERATURE SURVEY

### 2.1 Comparative Analysis

In the paper "Disease Prediction System using data mining techniques"[1] the author has discussed data mining techniques like association rule mining, classification, clustering to analyse the different kinds of heart based problems. The database used contains a collection of records, each with a single class label; a classifier performs a brief and clear definition for each class that can be used to classify successive records. The data classification depends on MAFIA algorithms that cause accuracy, the info is calculable exploitation entropy primarily based cross validations and partition techniques and also the

results are compared. C4.5 algorithmic rule is employed because of the coaching algorithmic rule to indicate rank of attack with the choice tree. The heart unwellness information is clustered mistreatment the K-means clump algorithmic rule, which will remove the data applicable to heart attack from the database. Some limitations are square measure faced by the system like, time complexity is more due to DFS traversal, C4.5- Time complexity increases while searching for insignificant branches and lastly no precautions are defined.

In the paper "A study on data mining prediction techniques in the healthcare sector" [2] the fields that are mentioned are, information Discovery method (KDD) is the method of adjusting the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial removal of implicit, antecedently unknown and doubtless helpful data from information in databases. The repetitious method consists of the subsequent steps: information cleansing, information integration, information choice, information transformation, data processing, Pattern analysis, Knowledge. Healthcare data processing prediction supported data processing techniques are as follows: Neural network, Bayesian Classifiers, call tree, Support Vector Machine. The paper states the comparative study of various aid predictions, Study of information mining techniques and tools for prediction of cardiovascular disease, numerous cancers, and diabetes, disease and medicine conditions. Few limitations are that if attributes are not related then Decision trees prediction is less accurate and ANN is computationally intensive to train also it does not lead to specific conclusion.

The paper "Predicting Disease by Using Data Mining Based on Healthcare Information System" [4] applies the information mining process to predict high blood pressure from patient medical records with eight alternative diseases. The data was extracted from a true world health care system info containing medical records. Under- sampling technique has been applied to come up with coaching knowledge sets, and data processing tool wood hen has been wont to generate the Naive Bayesian and J

48 classifiers were created to improve the prediction performance, and rough set tools were wont to scale back; the ensemble supported the concept of second- order approximation. Experimental results showed a bit improvement of the ensemble approach over pure Naive Bayesian and J-48 in accuracy, sensitivity and F-measure. Initially they'd a classification and so ensemble the classifiers and so the reduction of Ensemble Classifiers was employed. But the choice trees generated by J-48 are typically lacking within the leveling therefore the overall improvement of the victimization ensemble approach is a smaller amount.

The paper "An approach to devise an Interactive software solution for smart health prediction using data mining" [5] aims in developing a computerized system to check and maintain your health by knowing the symptoms. It has a symptom checker module which actually defines our body structure and gives us liability to select the affected area and checkout the symptoms. Technologies implemented in this paper are: The front end is designed with help of HTML, Javascript and CSS. The back end is designed using MySQL which is used to design the databases. This paper also contains the information of testing like Alpha testing which is done at server side or we can say at the developer's end, this is an actual testing done with potential users or as an independent testing process at server end. Beta testing is done after performing alpha testing, versions of a system or software known as beta versions are given to a specific audience outside the programming team. Only the limitation of this paper is it suggests only the award winning doctors and not the nearby doctors to the patient.

## 3. PROPOSED WORK

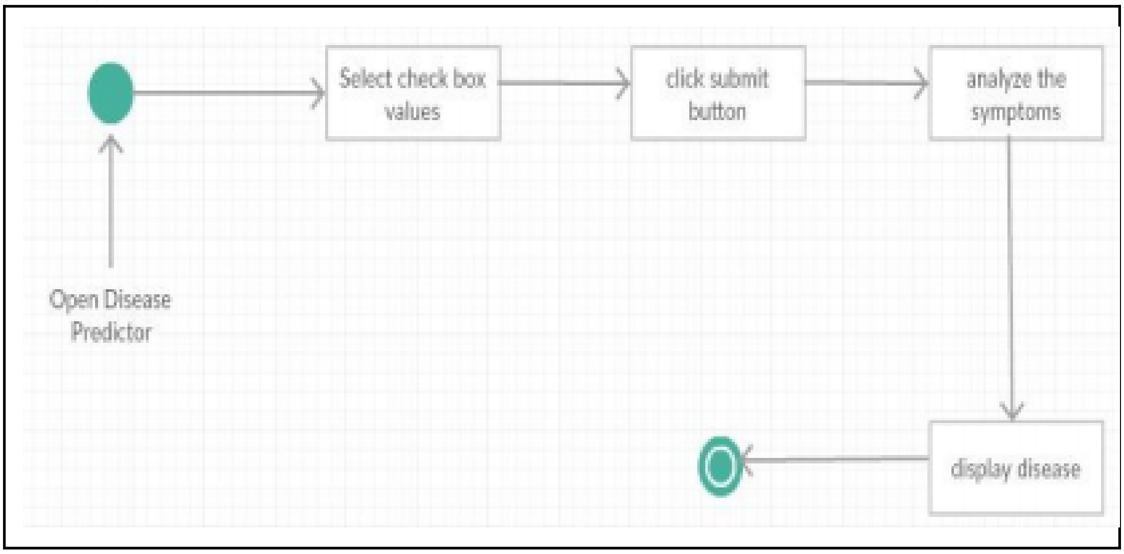


**Fig – 1:** State Diagram

We are predicting a disease which a person is suffering from depending upon the symptoms he or she is suffering. Here we take five symptoms from the patient and evaluate them by using algorithms such as Random Forest , k-nearest neighbour,Decision Tree, Naïve Bayes.

Steps of model building:

### i. Objective

We want to predict the disease suffered by a patient depending upon the symptoms.

### ii. Collecting data

Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning. The better the variety, density and volume of relevant data, the better the learning prospects for the machine becomes.

### iii. Preparing the data

Any analytical process thrives on the quality of the data used. One needs to spend time determining the quality of data and then taking steps for fixing issues such as missing data and treatment of outliers. Exploratory analysis is perhaps one method to study the nuances of the data in detail thereby burgeoning the nutritional content.

### iv. Training a model

This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.

### v. Evaluating the model

To test the accuracy, the second part of the data (holdout / test data) is used. This step determines the precision in the choice of the algorithm based on the outcome. A better test to check accuracy of a model is to see its performance on data which was not used at all during model build.

### vi. Improving the performance

This step might involve choosing a different model altogether or introducing more variables to augment the efficiency. That's why a significant amount of time needs to be spent in data collection and preparation.

## 4.METHODOLOGY

## Dataset

Dataset for this project was collected from a study of the University of Columbia performed at New York Presbyterian Hospital during 2004. Link of dataset is given below.

http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

## Library Used

In this project standard libraries for database analysis and model creation are used. The following are the libraries used in this project.

1. tkinter: It's a standard GUI library of python. Python when combined with tkinter provides a fast and easy way to create GUI. It provides a powerful object-oriented tool for creating GUI.It was used in this project to create our GUI namely messagebox, button, label,Option Menu, text and title. Using tkinter we were able to create an interactive GUI for our model.

2. Numpy: Numpy is the core library of scientific computing in python. It provides powerful tools to deal with various multi-dimensional arrays in python. It is a general purpose array processing package.

3. pandas : it is the most popular python library used for data analysis. It provides highly optimized performance with back-end source code purely written in C or python.

Data in python can be analysed with 2 ways

Series

Dataframes

Series is a one dimensional array defined in pandas used to store any data type.

Dataframes are two-dimensional data structures used in python to store data consisting of rows and columns.

Pandas dataframe is used extensively in this project to use datasets required for training and testing the algorithms. Dataframes make it easier to work with attributes and results. Several of its inbuilt functions such as replace were used in our project for data manipulation and preprocessing.

4. sklearn: Sklearn is an open source python library which implements a huge range of machine-learning, pre-processing, cross-validation and visualization algorithms. It features various simple and efficient tools for data mining and data processing. It features various classification,regression and clustering algorithm such as support vector machine, random forest classifier,decision tree, gaussian naïve-Bayes, KNN to name a few.In this project we have used sklearn to get advantage of inbuilt classification algorithms like decision tree, random forest classifier, k-nearest neighbour and naïve Bayes. We have also used inbuilt cross validation and visualization features such as classification report, confusion matrix and accuracy score.

In this project we are using four algorithms to predict disease based on symptoms.They are

1.Decision tree

2.Random forest tree

3.Gaussian Naive Bayes

4.kNN

**1. Decision tree**

Decision tree is a supervised learning technique program used for classification problems.It is also capable of engaging problems of higher dimensionality. It mainly consists of three parts: root, nodes and leaf.

This prediction method gives accuracy of ~95%.

**2. Random Forest**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble . Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree), **nij=wjCj - wleftjCleftj - wrightjCrightj**

where,

ni sub(j)= the importance of node

W sub(j)=weighted number of samples reaching node j

C sub(j)=the impurity value of node j

left(j) = child node from left split on node j

right(j)= child node from right split on node j

This prediction method is used with 100 random samples and gives accuracy of ~95%.

**3. k Nearest Neighbor**

kNN is a simple , easy to implement supervised learning algorithm used for classification and regression problems. It works by finding a pattern in data which links data to results and it improves upon the pattern recognition with every iteration.

Assume we are given a dataset where X is a matrix of features from an observation and Y is a class label. We will use this notation throughout this article. k-nearest neighbors then, is a method of classification that estimates the conditional distribution of Y given X and classifies an observation to the class with the highest probability. Given a positive integer k, k-nearest neighbors looks at the k observations closest to a test observation x0 and estimates the conditional probability that it belongs to class j using the formula,

$$\mathbf{Pr(Y=j|X=x0)=(1/k)\sum_{i\in N0} I(yi=j)}$$

where,
N0=set of k nearest observations
I(yi=j)=indicator variable that evaluates to 1 if given observation(xi,yi) in N0 is member of j,and 0 if otherwise

This prediction method has accuracy of ~92%.

### 4. Naive Bayes Algorithm

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification problems.

Bayes theorem is a mathematical formula used for calculating conditional probability. Conditional probability is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred. Formula is :

$$\mathbf{P(A/B) = (P(B/A)*P(A)) / P(B)}$$

This prediction method has an accuracy of 95%.

**Working of this model :**

There will be a text field where the name of the user needs to be entered.After entering name,user has to enter minimum two symptoms to predict a disease.There are four buttons called prediction1,prediction 2, prediction 3 and prediction 4 ,which predicts disease using decision tree ,random forest,naive bayes and kNN respectively.When user clicks on particular button,result will be displayed in the text field.There are two more buttons ,reset inputs and exit systems ,where reset inputs is used to clear the inputs last entered and exit system is used to exit.

## 5. RESULTS

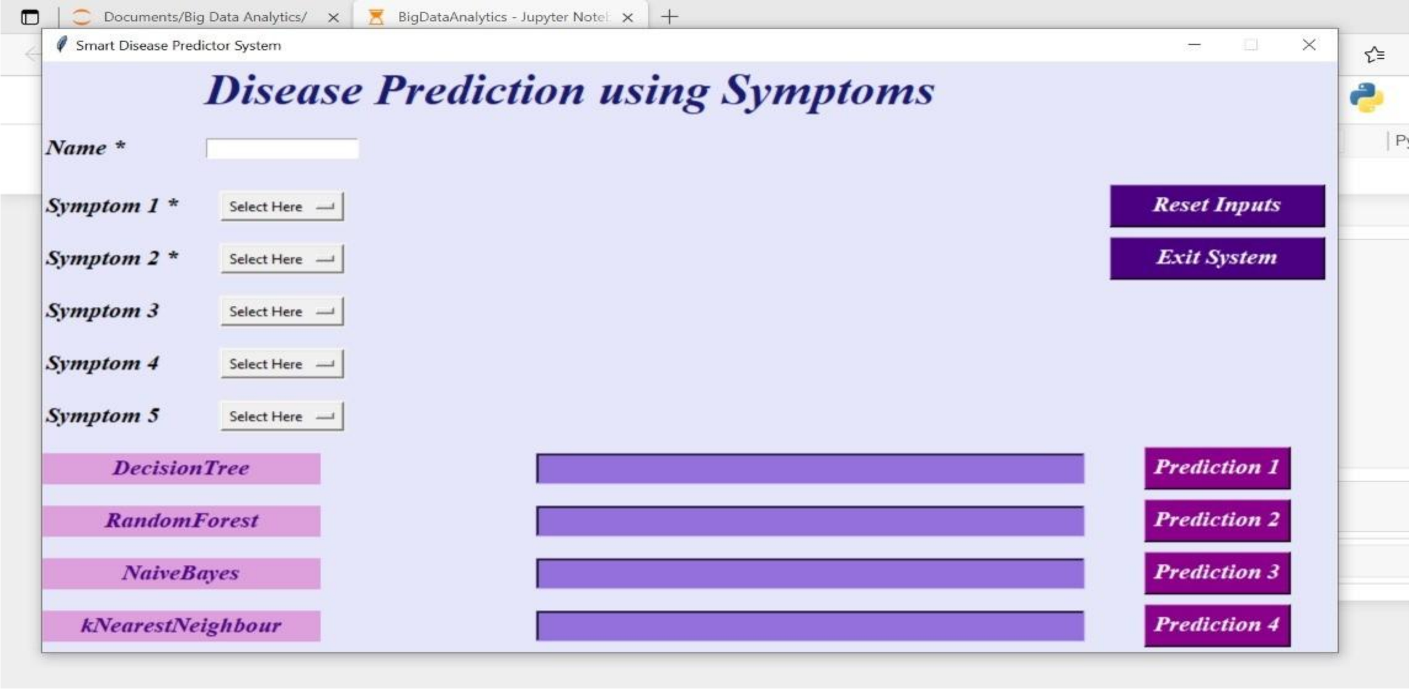
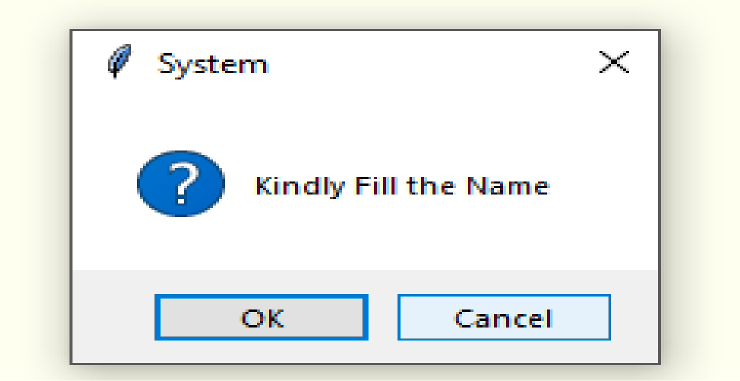
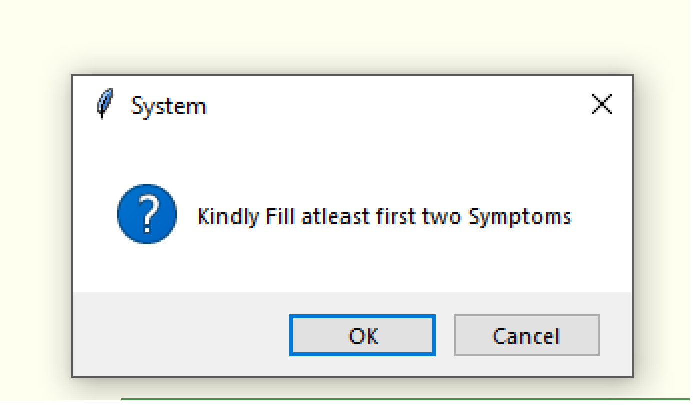


**Fig -2:** Welcome Page



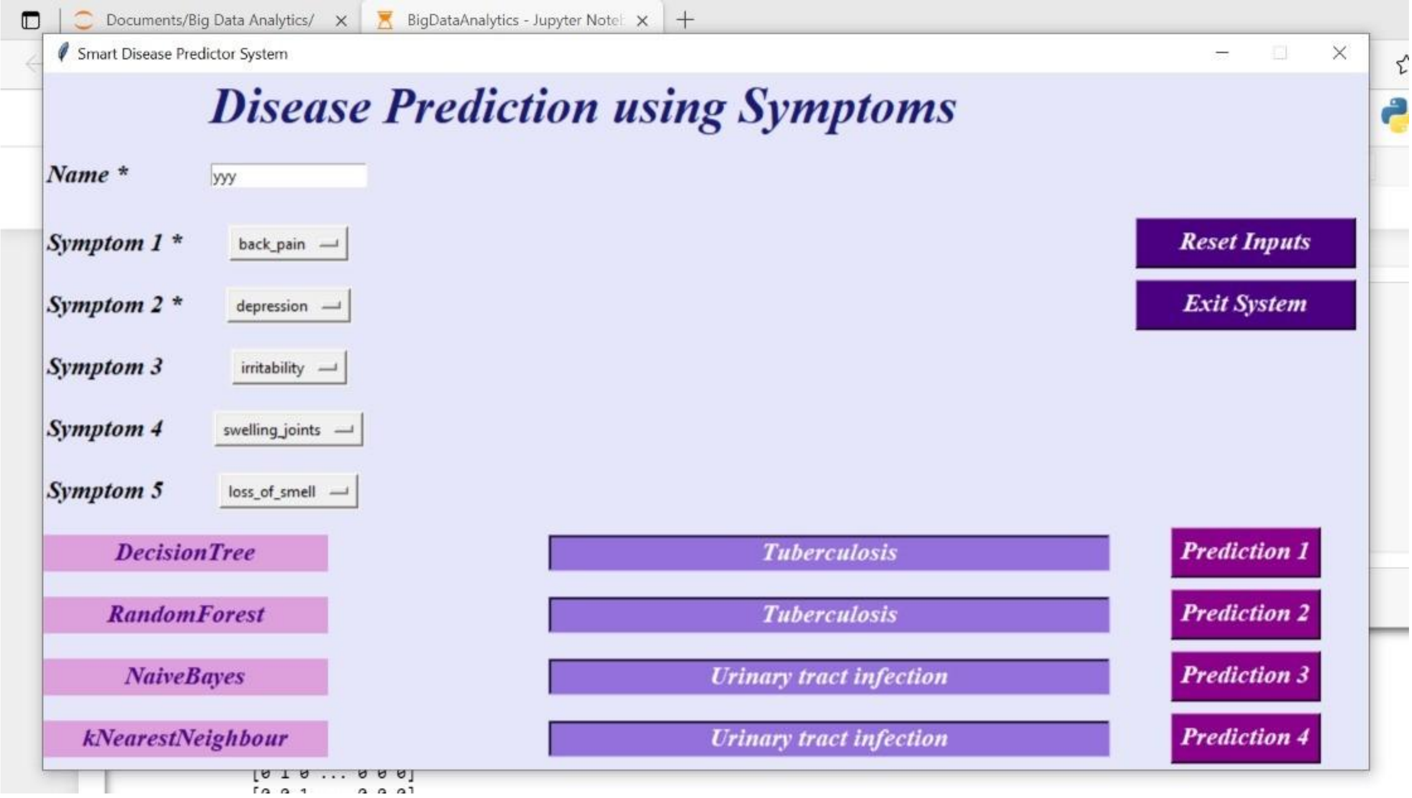
Kindly Fill the Name

**Fig-3**: Message box displayed when name field is not filled



Kindly Fill atleast first two Symptoms

**Fig-4**: Message box displayed when at least 2 symptoms are not filled



Smart Disease Predictor System

Disease Prediction using Symptoms

Name * yyy

Symptom 1 * back_pain

Symptom 2 * depression

Symptom 3 irritability

Symptom 4 swelling_joints

Symptom 5 loss_of_smell

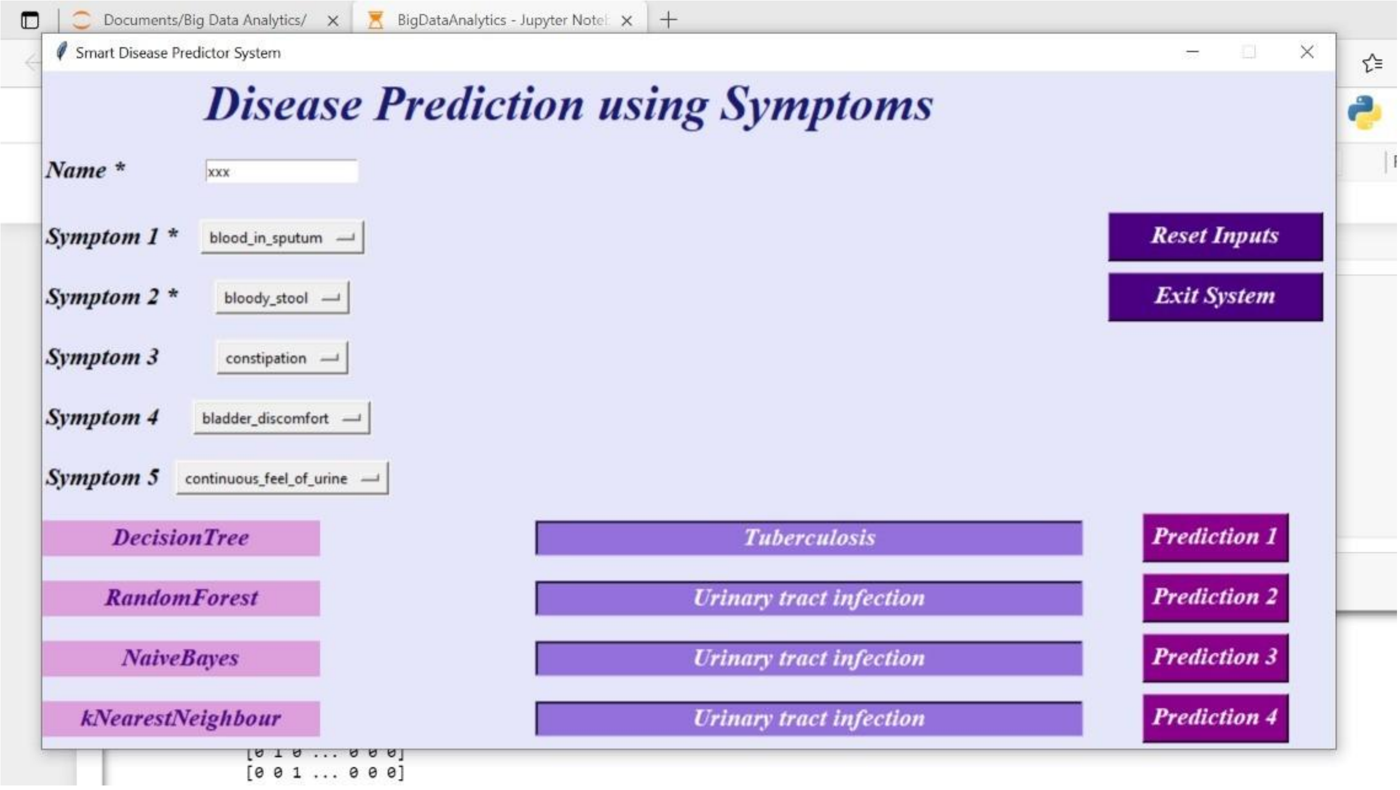
DecisionTree Tuberculosis

RandomForest Tuberculosis

NaiveBayes Urinary tract infection

kNearestNeighbour Urinary tract infection

**Fig-5:** Result Page(Input 1)



Smart Disease Predictor System

Disease Prediction using Symptoms

Name * xxx

Symptom 1 * blood_in_sputum

Symptom 2 * bloody_stool

Symptom 3 constipation

Symptom 4 bladder_discomfort

Symptom 5 continuous_feel_of_urine

DecisionTree Tuberculosis

RandomForest Urinary tract infection

NaiveBayes Urinary tract infection

kNearestNeighbour Urinary tract infection

```
[0 0 1 ... 0 0 0]
```

**Fig-6**: Result page(Input 2)

## 6. CONCLUSION

In this paper, our ultimate goal was to make an easy user interface for predicting the disease from various symptoms. In developing nations, predictive analytics are the next big idea in medicine and play a very important role . By this project, it is easy for doctors to predict which disease he/she is facing. Based on the disease predicted from the symptoms given by the doctors in the gui, doctors can easily treat patients for the predicted disease. From this project, it is going to consume less time for predicting which disease he/she is having and easily helps in giving better treatment for patients. Hospitals, pharmaceutical corporations and insurance suppliers can see changes furthermore. These changes which will virtually revolutionize the manner drugs are practiced for higher health and unwellness reduction. Relationship formation is one of the reasons doctors say they went into medicine, and when these diminish, so does their satisfaction with their profession

## FUTURE WORK

Every one of us would like to have a good medical care system and doctors are expected to be medical experts and make good decisions all the time. But it's highly unlikely to memorize all the knowledge, patient history, records needed for every situation. Although they have a massive amount of data and information, it's difficult to compare and analyse the symptoms of all the diseases and predict the outcome. So, integrating information into a patient's personalized profile and performing in-depth research is beyond the scope of a doctor. Predictive analytics is the process to make predictions about the future by analyzing historical data. For health care, it would be convenient to make the best decisions for every individual. Predictive modeling uses artificial intelligence to create a prediction from past records, trends, individuals, diseases and the model is deployed so that a new individual can get a prediction instantly. Health and Medicare units can use these predictive models to accurately assess when a patient can safely be released.

## REFERENCES

[1] Aditya Tomar, "Disease Prediction System using data mining techniques", in International Journal of Advanced Research in Computer and Communication Engineering, ISO 3297, July 2016

[2] Dr. B.Srinivasan, K.Pavya, "A study on data mining prediction techniques in the healthcare sector", in International Research Journal of Engineering and Technology (IRJET), March-2016.

[3] Megha Rathi, Vikas Pareek, "An integrated hybrid data mining approach for healthcare" , in IRACST -International Journal of Computer Science and Information Technology Security (IJCSITS), ISSN: 2249-9555 , Vol.6, No.6,Nov-Dec 2016.

[4] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System" , in IEEE 2012.

[5] M.A. Nishara Banu,B Gomathy, "An approach to devise an Interactive software solution for smart health prediction using data mining, in International Journal of Technical Research and Applications , eISSN, Nov-Dec 2013.

[6] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. Information Technology Journal.

[7] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical History