**Disease Prediction using machine learning**

A Project Report

submitted in partial fulfilment of the requirements for the degree of

**Bachelor of Technology**

in

**Computer Science and Engineering**

by

# Mayank Tiwari 1901610100108

# Saksham Mishra 1901610100173

# Lakshy Kandpal 1901610100100

# Shashi Kumar 1901610100187

Under the Esteemed Guidance of :

## Dr. Pratibha Singh



# Krishna Engineering College, Ghaziabad- 201007 Affiliated to



# Dr. A.P.J. Abdul Kalam Technical University, Lucknow May, 2022

## UNDERTAKING

We hereby declare that the work presented in this dissertation entitled **“Diseases Prediction using machine learning”** of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, submitted to Dr. A.P.J. Abdul Kalam Technical University, Lucknow, is our work carried out during the period from *20/07/2021* to *30/05/2022* under the guidance of **Vishakha Chaudhary** Krishna engineering college, Ghaziabad.

The work reported in their dissertation has not been submitted by us for award of any other degree or diploma.

Date :

Place: Ghaziabad

Mayank Tiwari

1901610100108

Lakshy kandpal

1901610100100

Shashi Kumar

1901610100187

Saksham Mishra

1901610100173

## CERTIFICATE

This is to certify that the Project report entitled **“Disease Prediction using machine learning”** done by Mayank Tiwari 1901610100108, Lakshy kandpal 1901610100100, Shashi kumar 1901610100187, Saksham Mishra 1901610100173 carried out by them at Krishna Engineering College, Ghaziabad under my guidance. The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Date: \_

|  |  |
| --- | --- |
| Prof. (Dr.) Sachin Malhotra | Vishakha Chaudhary |
| Head of department | (Associate Professor) |

(Department of Computer Science and Engineering)

## ACKNOWLEDGEMENT

We would like to take the opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during the Project work.

Firstly, we would like to thank our supervisor, Vishakha Chaudhary for being a great mentor and the best adviser we could ever have. Her advice, encouragement and critics are source of innovative ideas, inspiration and causes behind the successful completion of this dissertation. The confidence shown on us by her was the biggest source of inspiration for me. It has been a privilege working with her from last one year. She always helped us during our project and many other aspects related to academics. Her talks and lessons not only help in Project work and other activities of college but also make us good and responsible professional.

## ABSTRACT

Disease Prediction exploitation Machine Learning is that the system that's want to predict the diseases from the symptoms that are given by the patients or any user. The system processes the symptoms provided by the user as input and offers the output because the chance of the sickness. Naïve Bayes classifier is employed within the prediction of the sickness that may be a supervised machine learning algorithmic program. The chance of the sickness is calculated by the Naïve Bayes algorithmic program. With a rise in medicine and aid knowledge, correct analysis of medical knowledge edges early sickness detection and patient care. By exploitation regression toward the mean and decision tree we tend to are predicting diseases like polygenic disease,

Malaria, Jaundice, Dengue, and Tuberculosis

## TABLE OF CONTENTS

Candidate’s Declaration.................................................................................................. (ii)

Certificate ................................................................................................................. (iii)

Acknowledgement .......................................................................................................... (iv)

Abstract ................................................................................................................. (v)

List of Figures ……………………………………………………………………….. (vii)

List of Table ………………………………………………………………………... (viii)

**Chapter 1. Introduction** ........................................................................................... 8

**Chapter 2. Literature Survey**……………………………………………………... 10

2.1 Prediction of disease employing a classification algorithm.. …………….. 10

2.2 Iot based mostly Health observation System……………………………... 11

2.3 AI‐ based smart prediction of clinical disease using random forest classifier 11

2.4 Existing Systems ………………………………………………………….. 11

2.5 Proposed System …………………………………………………………… 11

**Chapter 3. Research and Methodology** ……………………………………………… 13

3.1 Decision tree……………………………………………………………….. 13

3.2 KNN……………………………………………………………………….. 13

3.3 Naïve Bayes………………………………………………………………… 13

3.4 Random Forest……………………………………………………………… 13 **Chapter 4. Implementation** …………………………………………………………….

4.1 Library Used- Tkinter ,Numpy,Pandas, sklearn ……………………………. 16

4.2 Models …………………………………………………………………….... 16

1. 2.1Decision tree……………………………………………………………… 17

4.2.2 Random Forest…………………………………………………………... 18

4.2.3 KNN…………………………………………………………………….. 19

4.2.4 Naïve Bayes……………… …………………………………………….. 20

4.3 Database collection ………………………………………………………… 21

**Chapter 5. Conclusion** ……………………………………………………………… **39**

**References** ………………………………………………………………. **40**

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Fig no.** | **Figure Name** | **Page No.** |
| **Figure 1** | System architecture of our model | 14 |
| **Figure 2** | Working model of decision tree algorithm | 17 |
| **Figure 3** | Working model of random forest algorithm | 18 |
| **Figure 4** | Working model of KNN algorithm | 19 |
| **Figure 5** | List of various symptoms | 21 |
| **Figure 6** | List of diseases | 22 |
| **Figure 7** | User interface diagram | 23 |
| **Figure 8** | Root.title() | 23 |
| **Figure 9** | Label of heading and contribution section | 23 |
| **Figure 10** | Label with four sections | 24 |
| **Figure 11** | Option menu to create top-down menu | 24 |
| **Figure 12** | Buttons used to give predict the outcome | 24 |
| **Figure 13** | Message box are used one- to restrain the to enter name | 25 |
| **Figure 14** | two- to ask for at least two symptoms | 25 |
| **Figure 15** | Three from exit the system | 26 |
| **Figure 16** | Output of five initial rows data frame df | 26 |
| **Figure 17** | Block per column distribution | 27 |
| **Figure 18** | Output for distribution graph of training.cse file | 28 |
| **Figure 19** | Output for the print(x) in which different symptoms has values 0 or 1 | 29 |
| **Figure 20** | Decision tree algorithm | 30 |
| **Figure 21** | Function like scatterplot and scatterimp are used to pair the training data | 30 |
| **Figure 22** | Random forest algorithm classifier | 31 |
| **Figure 23** | KNN algorithm | 32 |
| **Figure 24** | Naïve byes algorithm | 33 |
| **Figure 25** | Code of GUI box | 34 |
| **Figure 26** | Code of message box | 34 |
| **Figure 27** | Code of option menu | 35 |

|  |  |  |
| --- | --- | --- |
| **Figure 28** | Code of buttons | 35 |
| **Figure 29** | Output for printy() | 36 |
| **Figure 30** | Scatter plot for various symptoms | 36 |
| **Figure 31** | Scatter plot for diseases on the basis of symptoms given by user | 37 |
| **Figure 32** | Final output shown to the user | 37 |
| **Figure 33** | Database created using SQLITE3 | 38 |

# CHAPTER 1 INTRODUCTION

Due to environmental changes new diseases have gotten active that bit by bit affects several soul. care is that the one amongst the key issue that everybody takes care of and it needs continuous watching of health factors like heart rate, heat rate, gas level of blood, pressure level, aldohexose level etc. since the 2019 irruption, coronavirus sickness has unfold worldwide inflicting several death in each country. This sickness started spreading quick that each country was full of this and bit by bit the economy was additionally affected. This sickness was deadly and, in such time, early prediction of sickness with symptoms may lead to avoid wasting several lives. Prediction of sickness at time may break several chains of the disease.

With the increase in variety of patient and sickness each year medical system is overladen and with time became expensive in several countries. Most of the sickness involves a consultation with doctors to urge treated. With spare information prediction of sickness by associate algorithmic rule is terribly straightforward and low cost. Prediction of sickness by watching the symptoms is associate integral a part of treatment. In our project we've got tried accurately predicting a sickness by watching the symptoms of the patient. we've got used four totally different algorithms for this purpose associated gained an accuracy of 92-95%. Such a system will have a awfully massive potential in medical treatment of the longer term. we've got additionally designed associate interactive interface to facilitate interaction with the system. we've got additionally tried to point out and pictured the results of our study and this project

## OBJECTIVE

The main objective of our model is to reduce the efforts that are taken by patients, only to diagnose the disease. Early detection of diseases can reduce the risk factor. Many patients are losing their life only because of the late diagnosis of their disease, so through this, we are helping the patients in early diagnosis of their disease. patients can be aware of their health within minutes, Due to modern lifestyle and unhealthy eating habits chances of heart diseases and other health issues will be more so it is become important to check body's condition time to time and get updated as soon as possible if feeling any doubt about the illness.

# CHAPTER 2

**LITERATURE SURVEY**

## 2.1 Prediction of disease employing a classification algorithm

The author has planned Article known as “prediction of disease employing a classification algorithm”. The classification algorithmic rule like supplying regression, SVM and KNN wont to predict disease. each algorithmic rule is compared supported their accuracy through a confusion matrix. Here the dataset of disease is downloaded from UCI with associate degree instance of 567, and also the information is collected from the ILPD (Indian Liver Patient dataset).

After running many algorithmic rule the collected result and when analyzing the result, i used to be discovered that the 2 algorithms, i.e. KNN and supplying regression was having the most effective accuracy result compared with alternative remaining algorithms and out of those 2 algorithms, supplying regression was the most effective and extremely responsive in term of true positive rate or redivision. From the confusion matrix KNN accuracy comes intent on be 73.97% and also the SVM accuracy comes intent on be 71.97%. So, the most effective model for disease prediction is supplying regression because it offers the most effective accuracy among all alternative algorithms. [1]

## 2.2 Iot based mostly Health observation System

The author has planned a piece of writing known as “Iot based mostly Health observation System”. they need planned a system which incorporates the utilization of detector i.e. Temperature detector, pulse detector, temperature detector of outer surroundings and detector of outer humidness. The system uses Arduino controller. Here detectors area unit embedded on the patient body and alternative sensor is placed reception to ascertain the humidness and also the temperature of the space. The detector price is additionally showed on the digital display.

The system collects the info, calculates the worth and so through associate degree IoT cloud it's transmitted to the bottom station and from base station values area unit accessed by the doctor. If the framework finds any sharp modification in rate or the temperature then it might straight off inform the patient regarding it exploitation alarm and conjointly the system shows live information of the from detector and might be seen from the net. the centre rate detector relies on principle of photograph plethysmography. This whole system helps in reducing the price of physical visit to hospital, testing etc. [2]

## 2.3 AI‐ based smart prediction of clinical disease using random forest classifier and Naive Bayes

Their main aim of work is to predict the diseases among the trained dataset using classification algorithms. They trained the naive bayes and random forest classifier model with three different disease dataset namely diabetes, coronary heart disease and cancer. ANN is being used to classify the labelled images based on the determination of true positive and false positive detection rates.

The ANN is segmented into two approaches, initially they applied the classifier to the image data with region of interest (ROI) and second includes the ANN learn the features from pre-processed image signals.

The training data is 70% original data and testing data is 30%, for efficient data analysis. They used the Diabetes dataset originally from the NIDDK. All patients’ data given here are females at least 21 years old of Pima Indian heritage. Coronary heart disease dataset is used by Framingham heart study which includes several demographic risk factors Breast Cancer dataset is breast cancer Wisconsin dataset.

Accuracy of their model in Bayesian Classification network is 74.46, 82.35 and 63.74% for diabetes, coronary heart disease and cancer data Similarly, classification with Random Forest model shows the accuracy of 74.03, 83.85 and 92.40%.

## 2.4 Existing Systems

The existing system predicts the chronic diseases that area unit for a selected region and for the actual community. solely explicit diseases area unit foretold by this method. during this System, massive information & CNN algorithmic rule is employed for illness risk prediction. For S sort information, the system is mistreatment Machine Learning algorithmic rule i.e. K-nearest Neighbours, decision Tree, Naïve theorem. The accuracy of the prevailing System is up to 94.8%

In the existing paper, they contour machine learning algorithms for the effective prediction of chronic illness eruption in disease-frequent communities. They experiment with the changed prediction models over real- life hospital information collected from central China. They propose a convolutional neural network-based multimodal illness risk prediction (CNN-MDRP) algorithmic rule mistreatment structured and unstructured information from the hospital

## 2.5 Proposed Systems

Most of the chronic diseases area unit expected by our system. It accepts the structured variety of information as input to the machine learning model. this method is employed by end-users

i.e. patients/any user. during this system, the user can enter all the symptoms from that he or she is suffering. These symptoms then are going to be given to the machine learning model to predict the illness. Algorithms area unit then applied to which supplies the simplest accuracy. Then System can predict illness on the premise of symptoms. this method uses Machine Learning Technology. Naïve Bayes rule is employed for predicting the illness by victimisation symptoms, for classification KNN rule is employed, supplying regression is employed for extracting options that area unit having most impact price, the choice tree is employed to

divide the massive dataset into smaller elements. the ultimate output of this method are going to be the illness expected by the model.

# CHAPTER 3

## RESEARCH METHODOLOGY

In this project we've got used for algorithms like decision tree, KNN, Random Forest and Naïve Byes to predict sickness by getting into symptoms.

A decision tree may be a structure which will be accustomed divide up an oversized assortment of records into with success smaller sets of records by applying a sequence of straightforward decision tree.

K Nearest Neighbour (KNN) can be really simple, easy to understand, versatile and one among the top machine learning algorithms.

Naive Bayes is a straightforward but astonishingly powerful rule for predictive modelling. The independence assumption that enables rotten joint probability into a product of marginal likelihoods is termed as 'naive'. This simplified Bayesian classifier is termed as naive Bayes.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression issues. It builds decision trees on totally different samples and takes their majority vote for classification and average just in case of regression.

## SYSTEM ARCHITECTURE

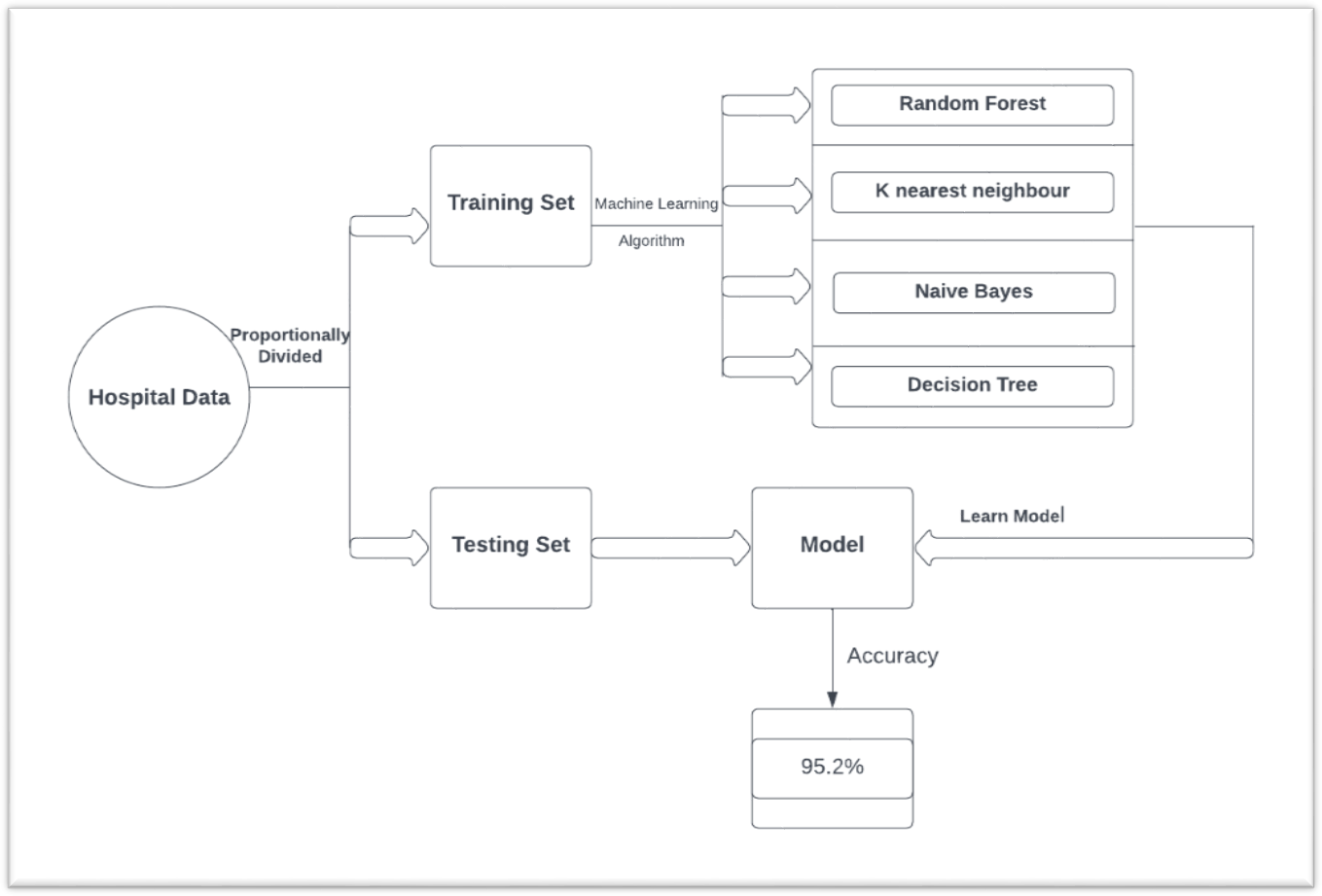


Figure 1 System architecture

# CHAPTER 4

## IMPLEMENTATION

### 4.1 LIBRARY USED

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

* tkinter: It’s a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It provides powerful object-oriented tool for creating GUI.

It provides various widgets to create GUI some of the prominent ones being:

* Button
* Canvas
* Label
* Entry
* Check Button
* List box
* Message
* Text
* Messagebox

Some of these were 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.

* Numpy: Numpy is 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.

Numpy’s main purpose is to deal with multidimensional homogeneous array. It has tools ranging from array creation to its handling. It makes it easier to create a n dimensional array just by using np.zeros() or handle its contents using various other methods such as replace, arrange, random, save, load it also helps I array processing using methods like sum, mean, std, max, min, all, etc

Array created with numpy also behave differently then arrays created normally when they are operated upon using operators such as +,-,\*,/.

All the above qualities and services offered by numpy array makes it highly suitable for our purpose of handling data. Data manipulation occurring in arrays while performing various operations need to give the desired results while predicting outputs require such high operational capabilities.

* 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 one dimensional array defined in pandas used to store any data type.

Dataframes are two-dimensional data structure 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 makes 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.

* sklearn: Sklearn is an open source python library with 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, KNN and naïve Bayes. We have also used inbuilt cross validation and visualization features such as classification report, confusion matrix and accuracy score.

## 4.2 Models

There are four different kinds of models present in our project to predict the disease these are

* Decision tree
* Random forest tree
* Gaussian Naïve Bayes
* KNN

**4.2.1 Decision tree** is classified as a very effective and versatile classification technique.

It is used in pattern recognition and classification for image. It is used for classification

in very complex problems due to its high adaptability. It is also capable of engaging problems of higher dimensionality. It mainly consists of three parts root, nodes and leaf.

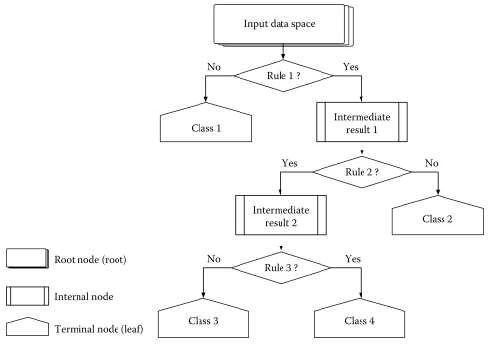


Figure 2 working model of decision tree algorithm

Roots consists of attribute which has most effect on the outcome, leaf tests for value Decision tree is the first prediction method we have used in our project. It gives us an accuracyof ~95%.

Decision trees are easy to comprehend, and information is transformed into IF-THEN rule which is also common kind of knowledge representation, choosing the next node is based on entropy of the dataset The more the decrease in entropy, the more information is gained.

Information Gain = (Entropy of parent node)-(Entropy of child node)

Consider a data set having a total number of N classes, than the entropy (E) can be determined with the formula below.

**4.2.2 Random Forest Algorithm** is a supervised learning algorithm used for both classification and regression. This algorithm works on 4 basic steps –

* It chooses random data samples from dataset.
* It constructs decision trees for every sample dataset chosen.
* At this step every predicted result will be compiled and voted on.
* At last most voted prediction will be selected and be presented as result of classification.

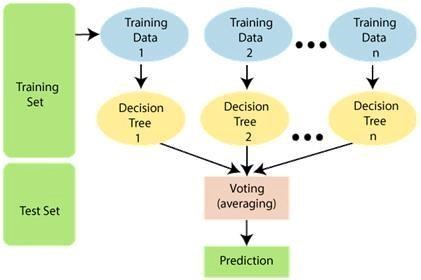


Figure 3 Random forest algorithms

In this project we have used random forest classifier with 100 random samples and the result given is ~95% accuracy.

**How Random Forest algorithm works?**

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1:** Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build

**Step-4**: Repeat Step 1 & 2

**4.2.3 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.

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

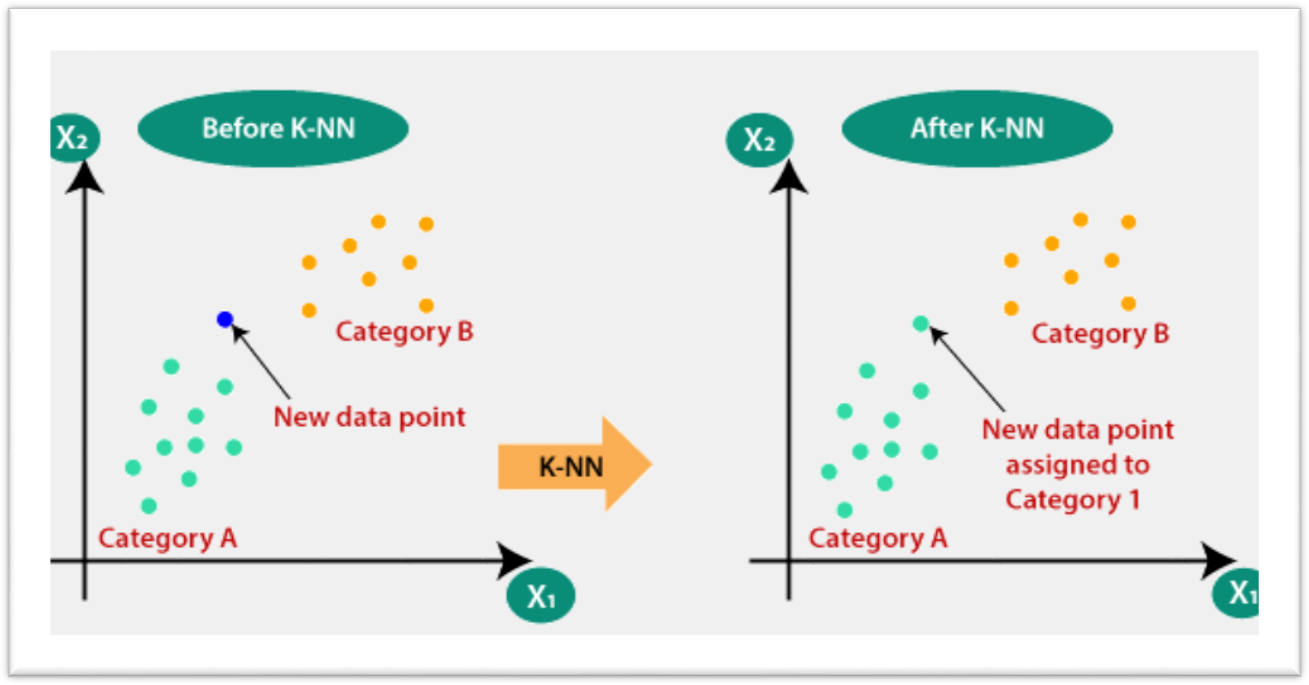


Figure 4 Working model of knn algorithms

**How does K-NN work?**

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbours

Step-2: Calculate the Euclidean distance of K number of neighbours

Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.

Step-4: Among these k neighbours, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.

Step-6: Our model is ready

We have used K Nearest Neighbour to classify our dataset and achieved ~92% accuracy

**4.2.4 Naïve Bayes** algorithm is a family of algorithms based on naïve bayes theorem. They share a common principle that is every pair of prediction is independent of each other. It also makes an assumption that features make an independent and equal contribution to the prediction.

Bayes' Theorem:

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:



Naïve Bayes Classifier Algorithm

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

In our project we have used naïve bayes algorithm to gain a ~95% accurate prediction.



Figure 5 L1 is the list made for various Symptoms which are generally showed up in people for various Diseases.

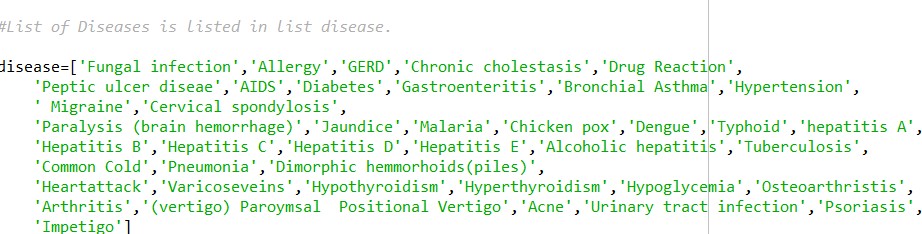


Figure 6 First L2 is the vacant list made. At that point, equivalent to a number of diseases in list L1, L2 is appended in a number of zeroes.

## GUI

GUI made for this project is a simple tkinter GUI consisting of labels, messagebox, button, text, title and option menu

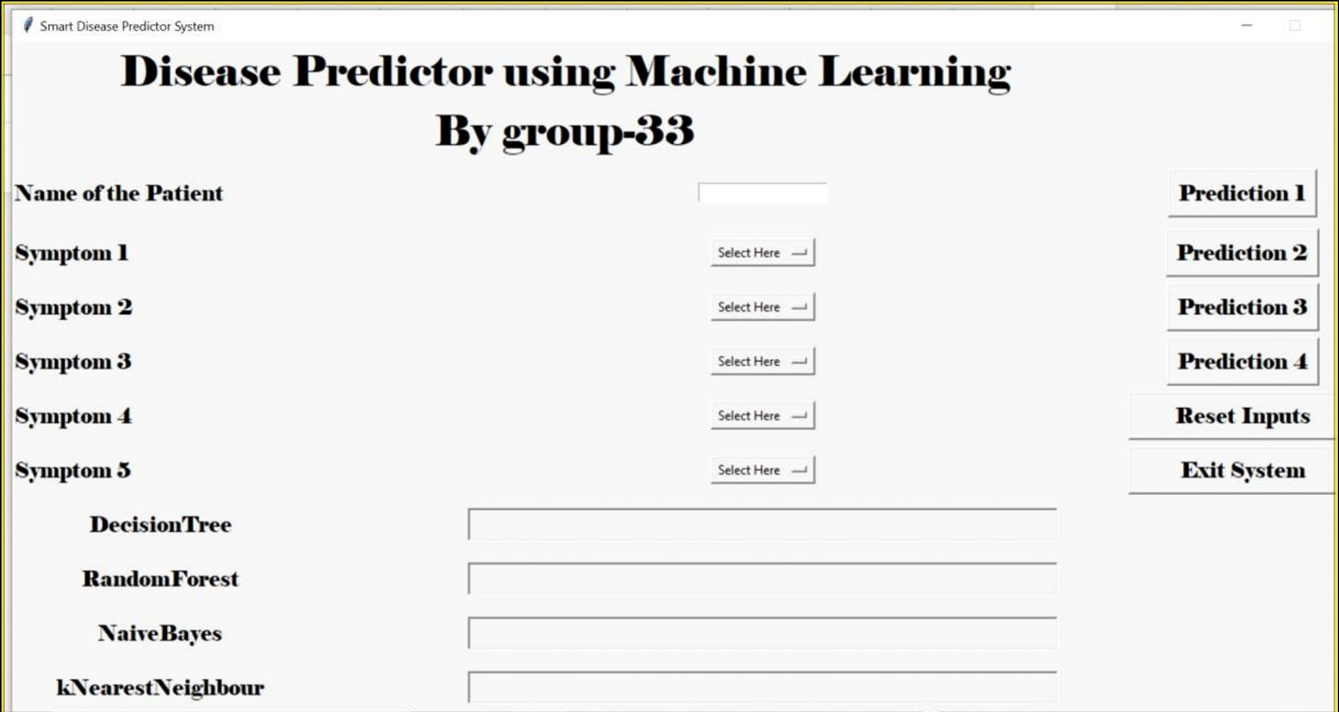


Figure 7 User GUI interface

Root.title() is used to set the the title as Smart Disease Predictor System



Figure 8 Label is used to add heading and contributors section.



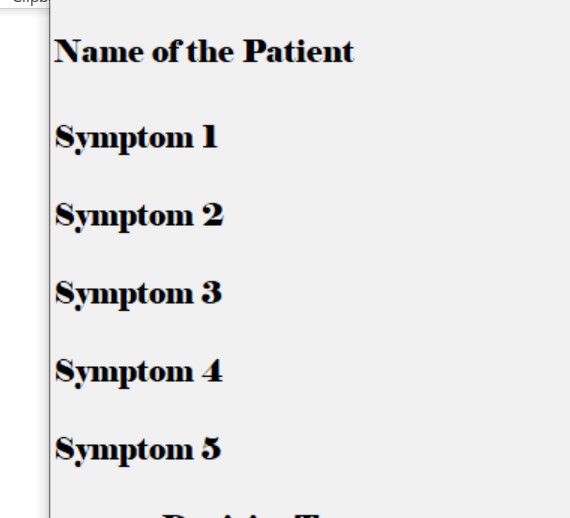
 Labels are further used for different section

Figure 9 OptionMenu is used to create drop down menu



Figure 10 Button used to predict the outcome

Buttons are used to give functionalities and predict the outcome of models also two utilitybuttons namely exit and rest are also created.

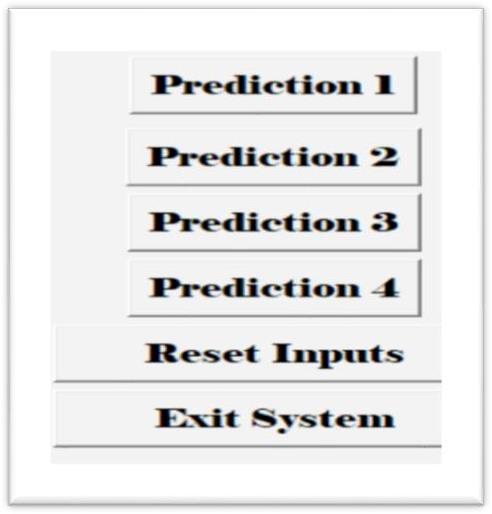


Figure 11 Text is used to show output of the prediction using blank space



Figure 12 Messagebox are used at three different places, one- to restrain the to enter name

,



Figure 13 two- to ask for at least two symptoms

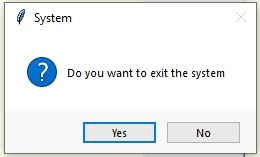


Figure 14 three- to confirm to exit system

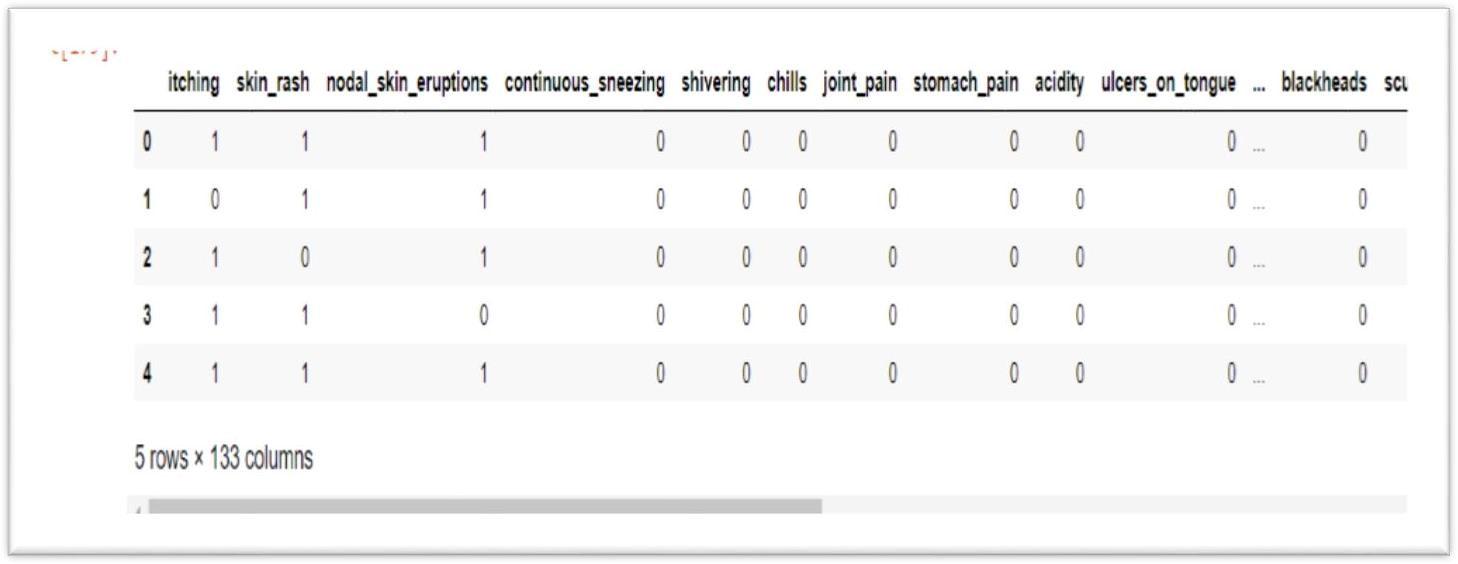


Figure 15 This is the output produced which contains the initial five rows of the dataframe df.

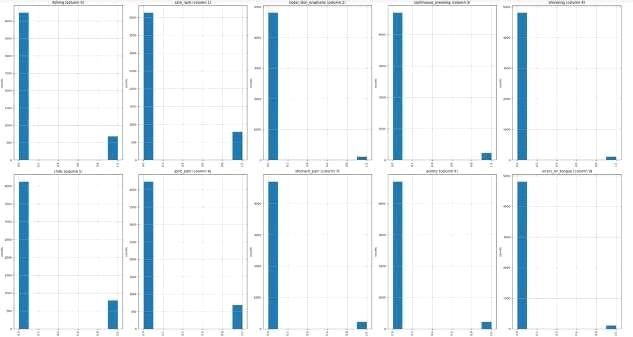


Figure16 Functions like plotpercolumbdistribution() plotScatterMatrix() is used to visualize the data.

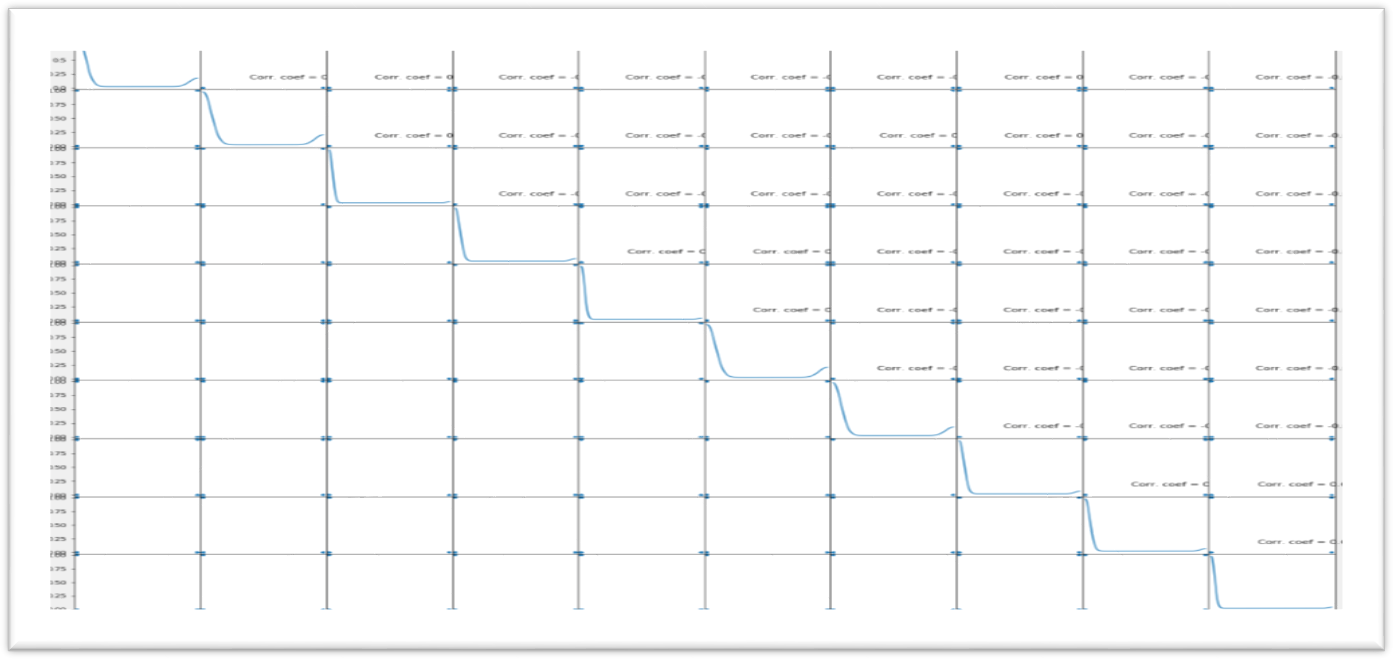


Figure 17 Output for the distribution graph of the columns of training.csv file.

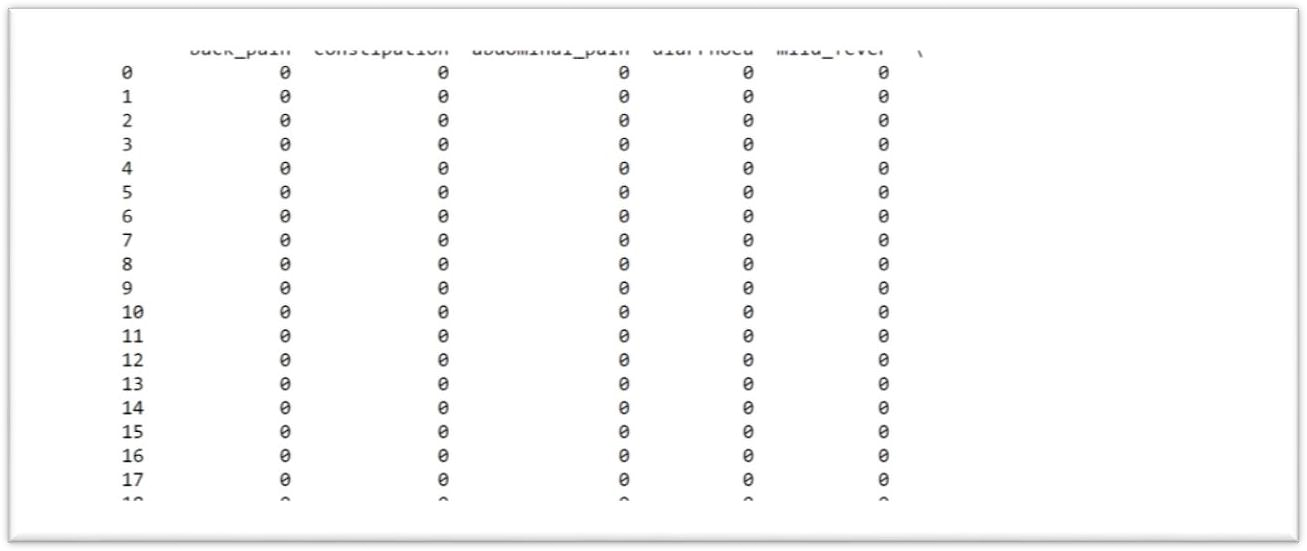


Figure 18 Output for the print(X)

Output for the print(X) in which different symptoms has the values ‘0’ or ‘1’ according to their presence in the particular diseases

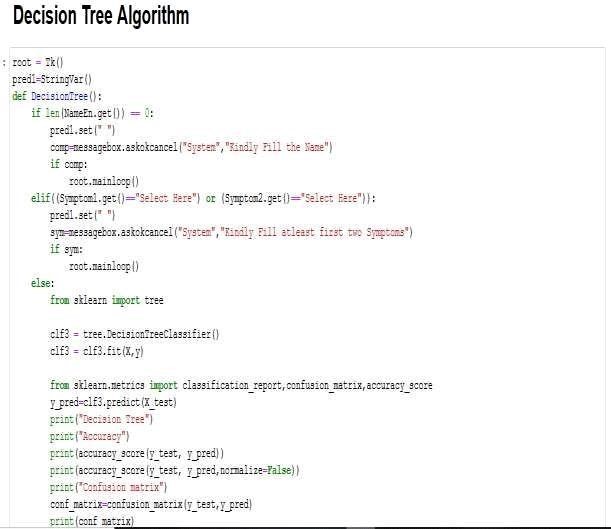


Figure 19 Function like scatterplt and scatterinp are used to compare input to training d

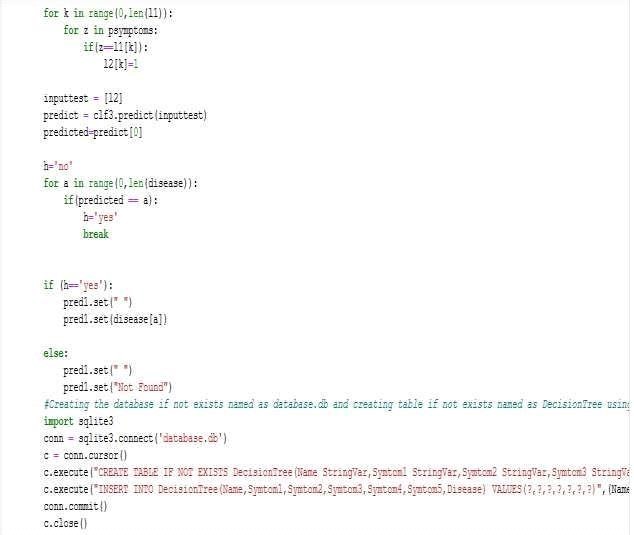


Figure 20 Algorithm of decision tree and database storage.

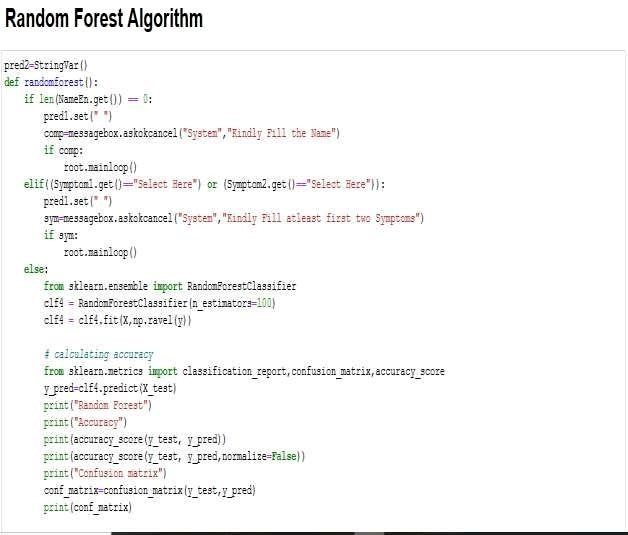


Figure 21 Algorithm of random forest classifier.



Figure 22 Algorithm of K nearest neighbour

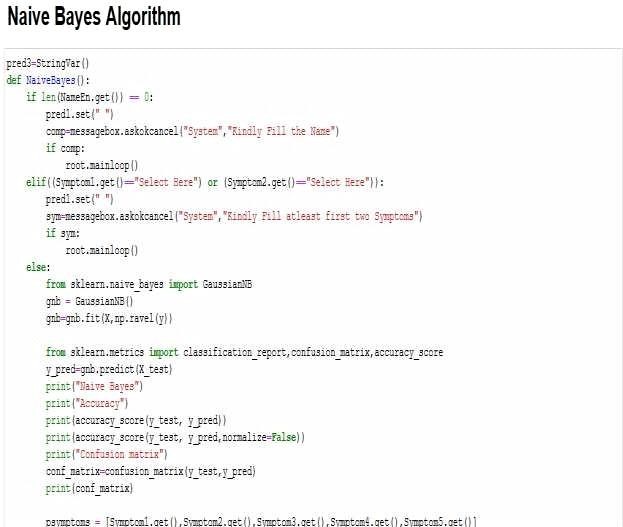


Figure 23 Algorithm of naïve bayes classifier

All these classifier is connected to database and GUI to function seamlessly.

Code of GUI to set initial values of labels.

Figure 24 Code of message box.

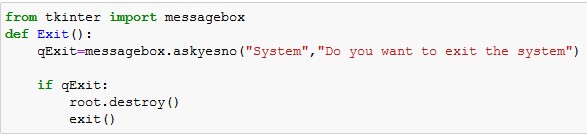


Figure 25 Code of message Box

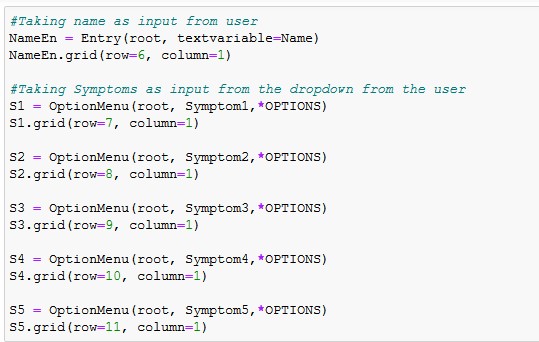


Figure 26 Code of option menu



figure 27 Code of buttons



Figure28 Output for the print(y) in which different disease has values according to their symptoms.

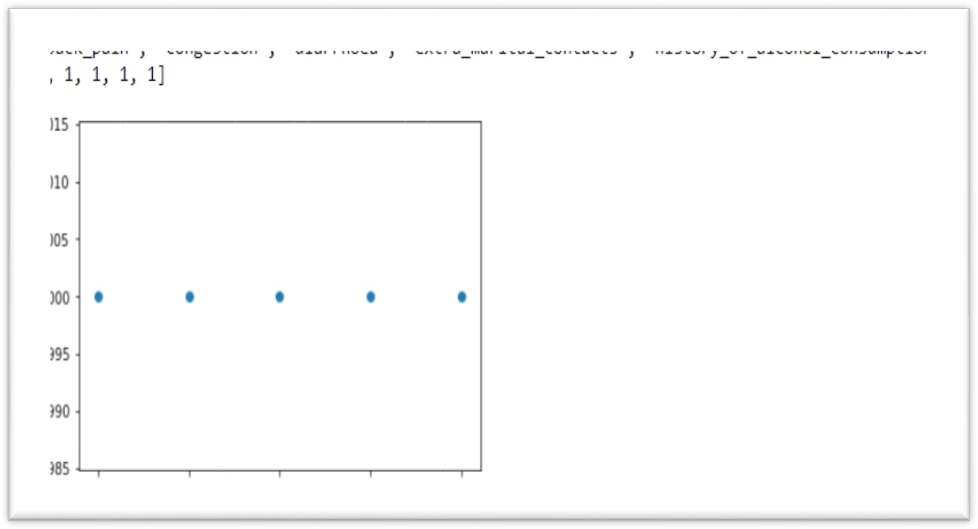


Figure 29 The scatterplot for the symptoms which are given by the user as input

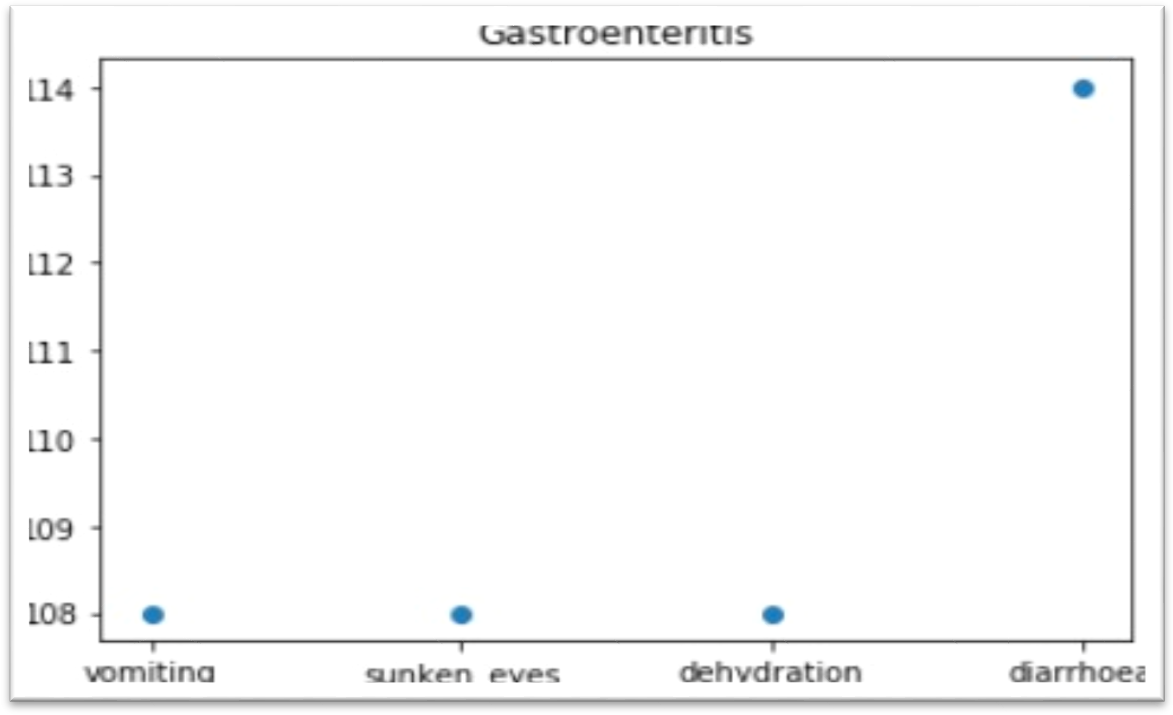


Figure 31 The scatter plot for the disease on the basis of symptoms which given by the user as input

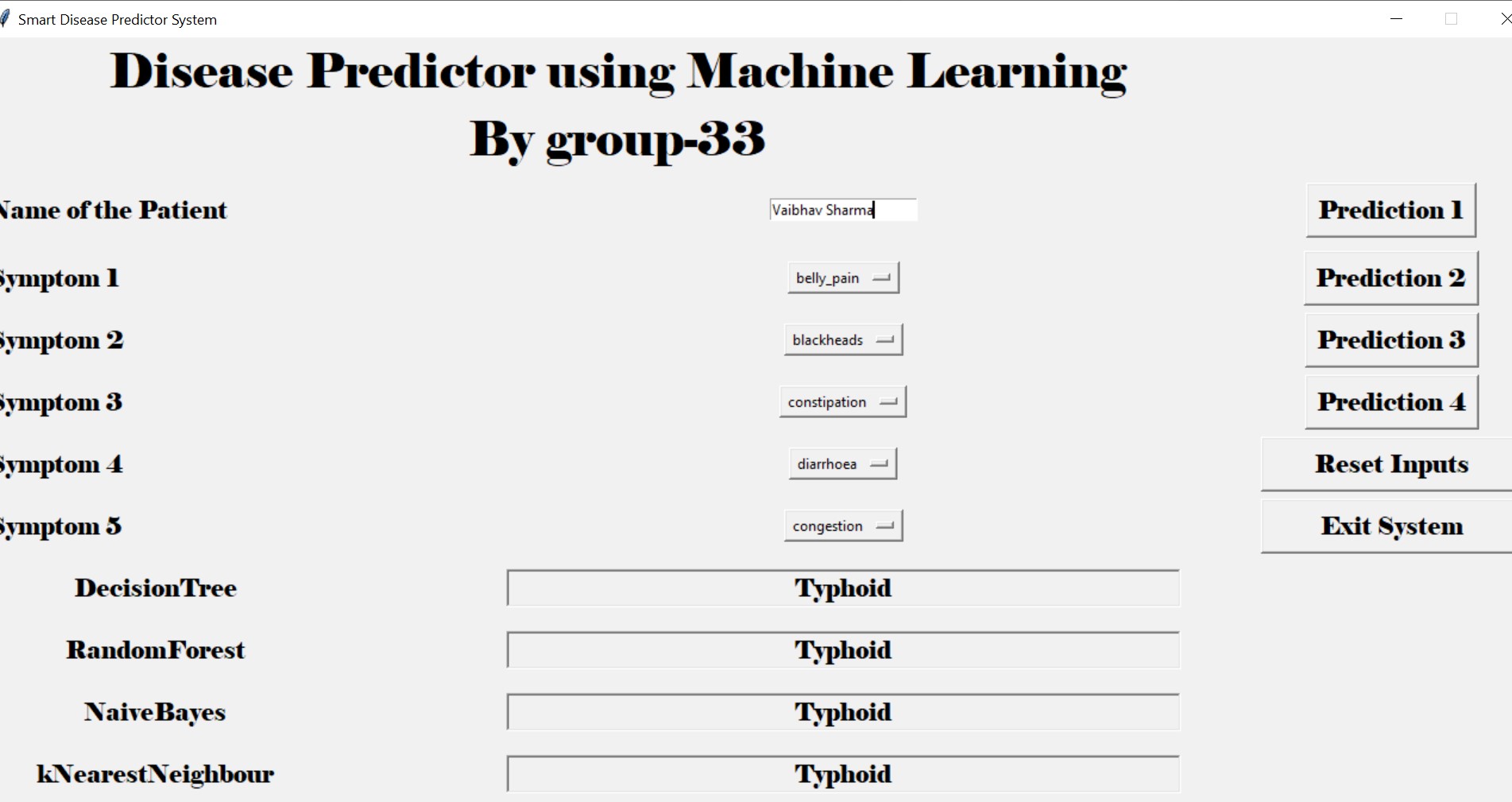


Figure 32 Final output shown to the user

## Database Collection

Dataset for this project was collected from a study of 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>

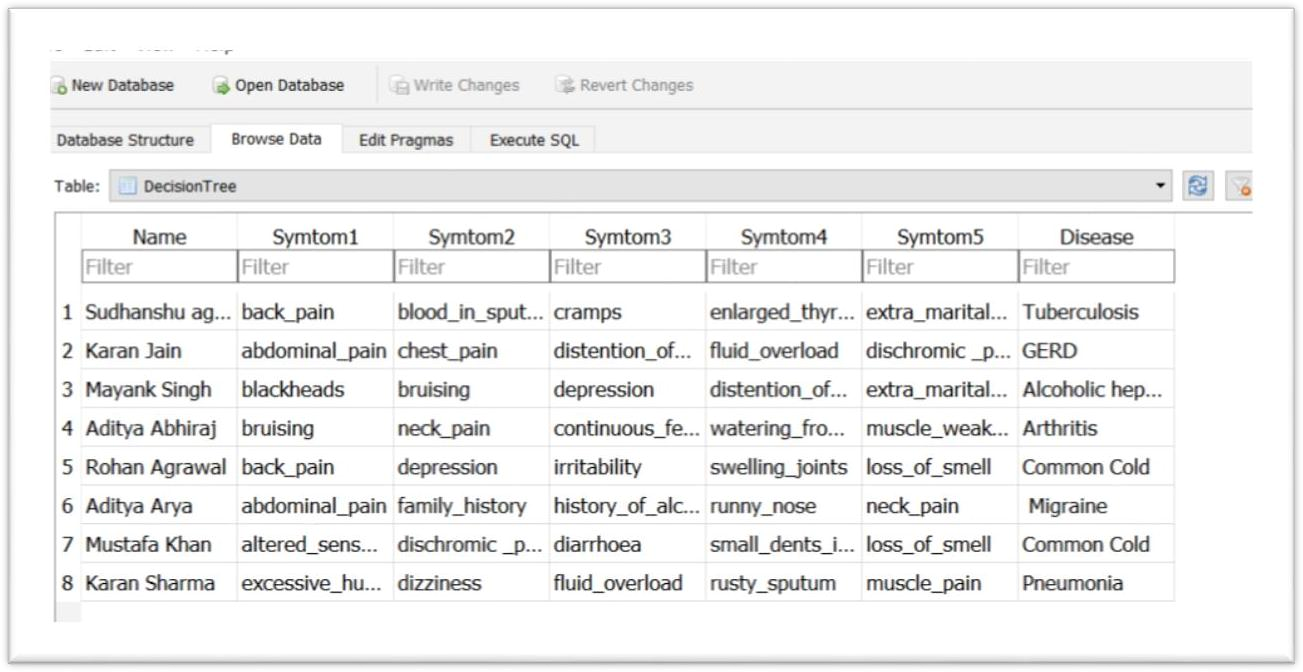


Figure 33 The database Created using Sqlite3

# CHAPTER 5

## Conclusion

We set out to create a system which can predict disease based on 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 algorithms to do so. On an average we achieved accuracy of ~94%. 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. Our system also has an easy-to-use interface. It also has various visual representation of data collected and results achieved.

**REFERENCES**

* Aditya Tomar, “Disease Prediction System using data mining techniques”, in International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016.
* Dr. Srinivasan, K.Pavya, “A study on data mining prediction techniques in healthcare sector”, in International Research Journal of Engineering and Technology (IRJET), March- 2016.
* 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.
* Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, “Predicting Disease by Using Data Mining Based on Healthcare Information System” , in IEEE 2012.
* 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.
* Computational Intelligence and Communication Technology (IEEE-CICT 2017) Implementing WEKA for medical data classification and early disease prediction. “3rd IEEE International Conference on"

[7]2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge

Management (ABLAZE-2015) “Predictions in Heart Disease Using Techniques of Data Mining”.

* Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin “Analysis of Data

Mining Techniques for Heart Disease Prediction”

* Dr. M.S. Shashidhara, M. Giri, Girija D.K “Data mining approach for prediction of fibroid Disease using Neural Networks,”
* Uma Ojha, Dr. Savita Goel. “Study on prediction of Breast cancer recurrence using Data mining techniques”
* Jackins, Vimal, Kaliappan and Lee, “Ai based smart prediction of clinical disease using random forest classifier and naïve byes” 2020
* .<http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html>