### A PROJECT REPORT ON

**“Disease prediction based on symptoms”**

### Submitted

**in partial fulfillment of the requirements for the Degree of**

### MASTERS OF COMPUTER APPLICATIONS

### by

**ROHIT NAUTIYAL**

### (Enrollment Number : PV-D1901019) Under the Supervision of

**MS POONAM VERMA**

**(Assistant Professor)**



### SCHOOL OF COMPUTING GRAPHIC ERA HILL UNIVERSITY

**DEHRADUN**

**(2019-2022)**

**CERTIFICATE**

I hereby declare that the work which is being presented in the project entitled, “**Disease prediction based on symptoms**” has been carried out by **ROHIT NAUTIYAL** for the partial fulﬁlment of the requirements for the award of the MCA, submitted in the School of Computing, **GRAPHIC ERA HILL UNIVERSITY** is an authentic record of our own **Ms. POONAM VERMA**. I further declare that the matter embodied in this project has not been submitted by us for the award of any other degree.

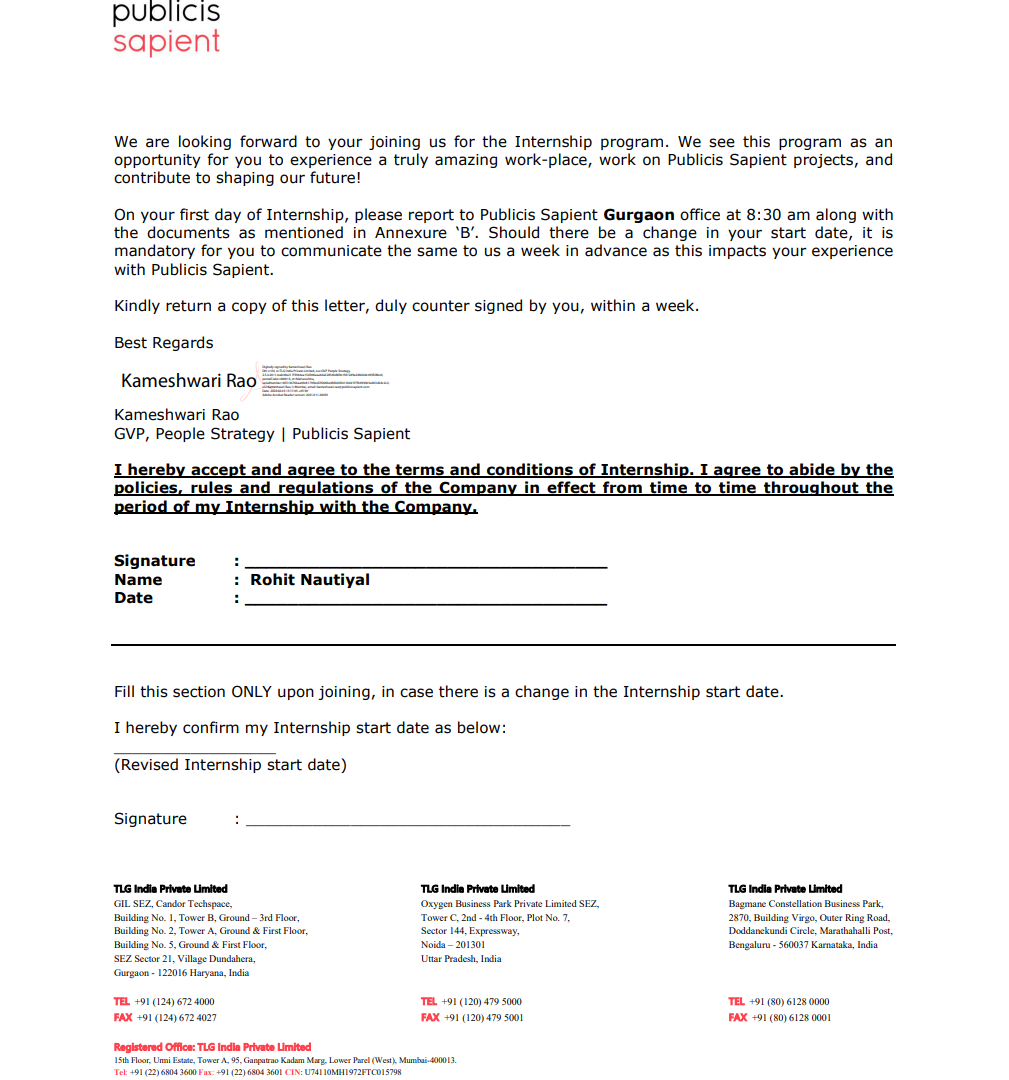
### PROJECT INCHARGE

MS POONAM VERMA

**(**Assistant Professor)

**OFFER LETTER**



****

**ACKNOWLEDGEMENT**

I take the opportunity to express my gratitude to all of them who in some or other way helped me to accomplish this challenging project in “**Disease prediction based on symptoms**”.

I owe a great many thanks to Graphic Era Hill University and my project guide **MS POONAM VERMA** who has been a constant support and guidance throughout the making of my project and for monitoring my project with attention and care. She has taken the pains to evaluate the project and make necessary corrections in the report as needed. I am really thankful for her kind and supportive nature. Her inspiring nature has always made my work easy.

# ABSTRACT

The wide adaptation of computer-based technology in the health care industry resulted in the accumulation of electronic data. Due to the substantial amounts of data, medical doctors are facing challenges to analyze symptoms accurately and identify diseases at an early stage. However, supervised machine learning (ML) algorithms have showcased significant potential in surpassing standard systems for disease diagnosis and aiding medical experts in the early detection of high-risk diseases. In this literature, the aim is to recognize trends across various types of supervised ML models in disease detection through the examination of performance metrics. The most prominently discussed supervised ML algorithms were Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbor (KNN). As per findings, Support Vector Machine (SVM) is the most adequate at detecting kidney diseases and Parkinson's disease. The Logistic Regression (LR) performed highly at the prediction of heart diseases. Finally, Random Forest (RF), and Convolutional Neural Networks (CNN) predicted in precision breast diseases and common diseases, respectively.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **1.OFFER LETTER** | ……………………………………… | 1 |
| **2. Acknowledgement** …………………………….. | | 2 |
| **3. Abstract** ………………………………… | | 3 |
| **4. Chapters** ……………………………………………………….. | | 4 |

CHAPTER 1. **PROJECT DESCRIPTION** 9

CHAPTER 2. **INTRODUCTION** 14

2.1 Introduction

* 1. Database Collection
  2. Library used

CHAPTER 3. **MACHINE LEARNING** 19

* 1. Introduction
  2. Features
  3. Importance
  4. Classification
  5. Prerequisites

CHAPTER 4. **ARTIFICIAL INTELLIGENCE** 24

* 1. Introduction
  2. Goals
  3. Advantage/Disadvantage
  4. Prerequisites

CHAPTER 5. **MODELS** 32

* 1. **Decision Tree Algorithm**

5.1.1 Types

5.1.2 Importance

5.1.3 Advantage/Disadvantage

* 1. **Random Forest Algorithm**

5.2.1 Application

5.2.2 Advantage/Disadvantage

* 1. **K Nearest Neighbor**

5.3.1 Application

5.3.2 Advantage/Disadvantage

* 1. **Naïve Bayes Algorithm**

5.4.1 Application

5.4.2 Advantage/Disadvantage

CHAPTER 5. **GUI**  40

CHAPTER 6. **CONCLUSION**  53

CHAPTER 7. **REFERENCES**  54

Disease prediction based on symptoms.

# Project Description

Health information needs are also changing the information seeking behavior and can be observed around the globe. Challenges faced by many people are looking online for health information regarding diseases, diagnoses and different treatments. If a recommendation system can be made for doctors and medicine while using review mining will save a lot of time. In this type of system, the user face problem in understanding the heterogeneous medical vocabulary as the users are laymen. User is confused because a large amount of medical information on different mediums are available. The idea behind recommender system is to adapt to cope with the special requirements of the health domain related with users.

**# Machine Learning Project on Disease Predictions using Symptoms**

**# Problem Statement**

Health information needs are also changing the information seeking behavior and can be

observed around the globe. Challenges faced by many people are looking online for health

information regarding diseases, diagnoses and different treatments. If a recommendation system

can be made for doctors and medicine while using review mining will save a lot of time. In this

type of system, the user face problem in understanding the heterogeneous medical vocabulary as

the users are laymen. User is confused because a large amount of medical information on

different mediums are available.The idea behind recommender system is to adapt to cope with the special requirements of the health domain related with users.

**# Explanation of Files**

**## training.csv**

\* This is the main dataset which has been used in this project. This dataset consist of mainly two columns "Disease" and "Symptoms" but this dataset is preprocessed so it helps in easily clasifying the data. This dataset is used to train our model.

## testing.csv

\* This is the dataset which has been used to test our model so that we can know the accuracy of our model. this dataset is predefined with output.

## Compiled\_Report.pdf

\* This is the complete explaination of Each and every module implemented in software. You can refer to the pdf for the detailed explaination of every part.

**## PythonCodeOfAlgorithm.py**

This is the file which consist of dataset and there are various differnt algorithms used for training of our model which are as follows:

\* Decision Tree

\* Random Forest

\* K Nearest Neighbors

\* Naive Bayes

These four algorithms is used to train our model and all gives an accuracy of over 90

**## Database**

The database used in this project is "SQLite" whose name is database.db which consist of four tables in which we have shown the results of four different algorithms. we are saving the results of users with their names for future preferences.

**## GUI.py**

This is the file which is used to create the interface of our system. GUI stands for Graphical User Interface and to create it we have used Tkinter which gives a software kind of view to our project where user can directly interact with the system by entering the symptoms of dieases and he/she will get the disease through various algorithms.

**## Project\_ML.ipynb**

This is the Jupyter notebook which consist of complete code. This is used to explain the working of each and every module used in the project.

**## GUI.jpeg**

This file contains the screenshot of the built GUI which shows the working of the system

**# Working with GUI**

**## Step 1:**

Enter the name in the provided space in front of the label as "Name of the Patient". It is the mandatory field which user have to enter in order to get result.

**## Step 2:**

Select 5 Symptoms from the dropdown menu which are labelled as Symptom 1, Symptom 2, Symptom 3, Symptom 4, Symptom 5 respectively. If user is not aware of 5 symptoms then it is mandatory for him to enter at least 2 starting systems, otherwise the result will not come and a message box will pop up for the same

**## Step 3:**

As per user interest, he/she can predict the disease using different algorithms such as Decision tree algorithm, Random forest algorithm, Naive bayes algorithm and K-Nearest neighbor. According to algorithm click on buttons:</br>

Press Prediction 1 for Decision tree algorithm</br>

Press Prediction 2 for Random forest algorithm</br>

Press Prediction 3 for Naive bayes algorithm</br>

Press Prediction 4 for K-Nearest neighbour</br>

(User can predict the disease using more than one algorithm at a time)

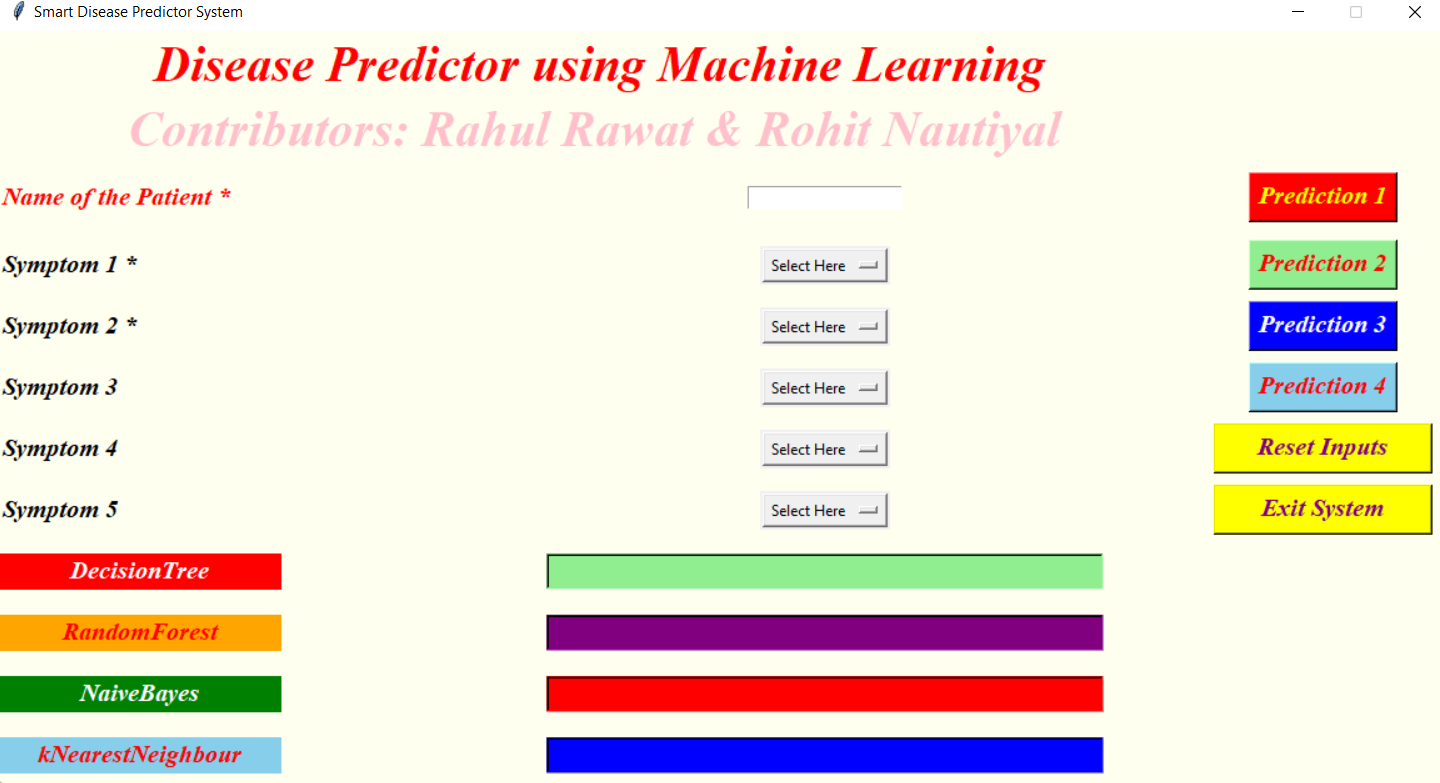
**## Step 4:**

Disease Recommendation will be available in front of the labels of algorithm of user's choice.

**## Step 5:**

Click on "Reset" button to predict the disease for any other patient or Press "Exit System" button to come out of the GUI.

**# A picture of GUI Interface**



**Introduction**

With the rise in number of patient and disease every year medical system is overloaded and with time have become overpriced in many countries. Most of the disease involves a consultation with doctors to get treated. With sufficient data prediction of disease by an algorithm can be very easy and cheap. Prediction of disease by looking at the symptoms is an integral part of treatment. In our project we have tried 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 visualized the result of our study and this project.

# 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>

# 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 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.

1. 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.

1. 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.

1. 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.

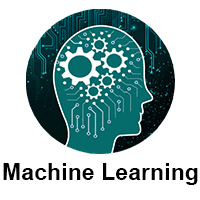
In this project we have used sklearn to get advantage of inbuilt classification algorithms like decision tree, random forest classifier, KNN and naïve Bayes. We have also used inbuilt cross validation and visualization features such as classification report, confusion matrix and accuracy score.

**RELATED WORK**

# RELATED WORK The work presented in focuses on disease prediction from clinical data provided by New York - Presbyterian Hospital. As these are clinical data, automated disease prediction is relatively different and easier than predicting from user text input. It is observed that input from common user contains less number of clinical terms. That means, matching the symptom names from user text input with system database has much more complexity. [4] emphasizes on prediction of potential infectious disease outbreaks from online text sources. Which is also a specialized source where explicit medical terms are used. A lot of effort has been put on to predict specific diseases [7], [8]. For instance, authors in [7] focus on predicting coronary heart diseases by mining text. There are also quite a number of research works that have been done in recent years on healthcare forums. [9] is such a work where natural language processing is used to rate and analyze user comments in order to predict diseases and extract rare side effects of drugs. This s ystem took into ac-count suggestions provided by different users on comment sections in disease analysis. Healthcare websites such as isabelhealthcare.com, mayo-clinic.org, patient.co.uk, are providing disease prediction based on user input ([5], [10] and [6]). [10] uses jargon-laden interface (I.e. users need to navigate through a longlist of symptoms). From user‟s point of view, it is a cumbersome task and the process is time consuming as well. Moreover, if a certain symptom is not found by the users, they are compelled to skip that symptom which is not desired at all. [5], [6] take guided input from user. However, they rely on mere symptom-disease relationship framework ([11], [12]) and use full text database [23]. Upon user input, these systems start looking for exact word match in the database from each input line. Thus it does not allow linguistic diversity. E.g. if the database does not contain a symptom‟s synonym used by a user, it will not be able to match the input perfectly. If the input contains more non-technical terms than expected, its performance degrades significantly. The framework used is very much rigid and confined to specific input types.

# CONTRIBUTION

# In this paper, the contribution includes proposing a new disease prediction framework (ADPS) that takes into account symptom names as well as other vital parameters (to be described in section 5) to improve disease prediction accuracy and proposing techniques (to be described in section 5) to allow greater linguistic diversity so that users do not feel uncomfortable while giving input.

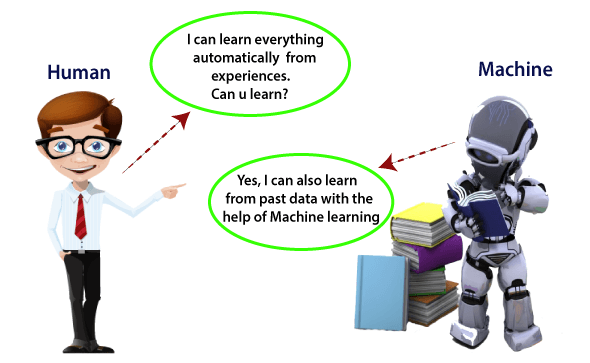


Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as **Supervised**, **Unsupervised**, and **Reinforcement** learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

## What is Machine Learning

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.



**Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.**

## Features of Machine Learning:

* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.

**Importance of Machine Learning:**

* Rapid increment in the production of data
* Solving complex problems, which are difficult for a human
* Decision making in various sector including finance
* Finding hidden patterns and extracting useful information from data.

## Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

### 1) Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

* **Classification**
* **Regression**

### 2) Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

* **Clustering**
* **Association**

### 3) Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

## Prerequisites

Before learning machine learning, you must have the basic knowledge of followings so that you can easily understand the concepts of machine learning:

* Fundamental knowledge of probability and linear algebra.
* The ability to code in any computer language, especially in Python language.
* Knowledge of Calculus, especially derivatives of single variable and multivariate functions.

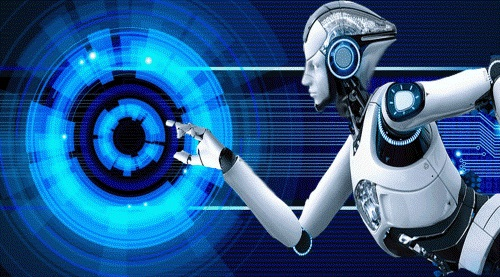
# Artificial Intelligence



In today's world, technology is growing very fast, and we are getting in touch with different new technologies day by day.

Here, one of the booming technologies of computer science is Artificial Intelligence which is ready to create a new revolution in the world by making intelligent machines.The Artificial Intelligence is now all around us. It is currently working with a variety of subfields, ranging from general to specific, such as self-driving cars, playing chess, proving theorems, playing music, Painting, etc.

AI is one of the fascinating and universal fields of Computer science which has a great scope in future. AI holds a tendency to cause a machine to work as a human.



Artificial Intelligence is composed of two words **Artificial** and **Intelligence**, where Artificial defines *"man-made,"* and intelligence defines *"thinking power"*, hence AI means *"a man-made thinking power."*

## Goals of Artificial Intelligence

Following are the main goals of Artificial Intelligence:

1. Replicate human intelligence
2. Solve Knowledge-intensive tasks
3. An intelligent connection of perception and action
4. Building a machine which can perform tasks that requires human intelligence such as:
   * Proving a theorem
   * Playing chess
   * Plan some surgical operation
   * Driving a car in traffic
5. Creating some system which can exhibit intelligent behavior, learn new things by itself, demonstrate, explain, and can advise to its user.

## Advantages of Artificial Intelligence

Following are some main advantages of Artificial Intelligence:

* **High Accuracy with less errors:** AI machines or systems are prone to less errors and high accuracy as it takes decisions as per pre-experience or information.
* **High-Speed:** AI systems can be of very high-speed and fast-decision making, because of that AI systems can beat a chess champion in the Chess game.
* **High reliability:** AI machines are highly reliable and can perform the same action multiple times with high accuracy.
* **Useful for risky areas:** AI machines can be helpful in situations such as defusing a bomb, exploring the ocean floor, where to employ a human can be risky.
* **Digital Assistant:** AI can be very useful to provide digital assistant to the users such as AI technology is currently used by various E-commerce websites to show the products as per customer requirement.
* **Useful as a public utility:** AI can be very useful for public utilities such as a self-driving car which can make our journey safer and hassle-free, facial recognition for security purpose, Natural language processing to communicate with the human in human-language, etc.

## Disadvantages of Artificial Intelligence

Every technology has some disadvantages, and thesame goes for Artificial intelligence. Being so advantageous technology still, it has some disadvantages which we need to keep in our mind while creating an AI system. Following are the disadvantages of AI:

* **High Cost:** The hardware and software requirement of AI is very costly as it requires lots of maintenance to meet current world requirements.
* **Can't think out of the box:** Even we are making smarter machines with AI, but still they cannot work out of the box, as the robot will only do that work for which they are trained, or programmed.
* **No feelings and emotions:** AI machines can be an outstanding performer, but still it does not have the feeling so it cannot make any kind of emotional attachment with human, and may sometime be harmful for users if the proper care is not taken.
* **Increase dependency on machines:** With the increment of technology, people are getting more dependent on devices and hence they are losing their mental capabilities.
* **No Original Creativity:** As humans are so creative and can imagine some new ideas but still AI machines cannot beat this power of human intelligence and cannot be creative and imaginative.

## Prerequisite

Before learning about Artificial Intelligence, you must have the fundamental knowledge of following so that you can understand the concepts easily:

* Any computer language such as C, C++, Java, Python, etc.(knowledge of Python will be an advantage)
* Knowledge of essential Mathematics such as derivatives, probability theory, etc.

## Audience

Our AI tutorial is designed specifically for beginners and also included some high-level concepts for professionals.

**MODELS**

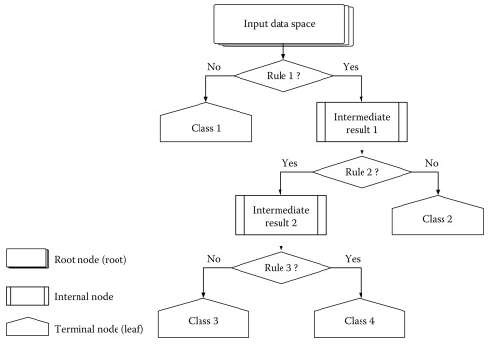
There are four different kind of models present in our project to predict the disease these are

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

**Decision Tree Algorithm**

**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 dew to its high adaptability. It is also capable of engaging problems of higher dimensionality. It mainly consists of three parts root, nodes and leaf.

Roots consists of attribute which has most effect on the outcome, leaf tests for value of certain attribute and leaf gives out the output of tree.



Decision tree is the first prediction method we have used in our project. It gives us an accuracy of ~95%.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**Types of Decision Trees**

Types of decision trees are based on the type of target variable we have. It can be of two types:

Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.

Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

Example:- Let’s say we have a problem to predict whether a customer will pay his renewal premium with an insurance company (yes/ no). Here we know that the income of customers is a significant variable but the insurance company does not have income details for all customers. Now, as we know this is an important variable, then we can build a decision tree to predict customer income based on occupation, product, and various other variables. In this case, we are predicting values for the continuous variables.

**Important Terminology related to Decision Trees**

**Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

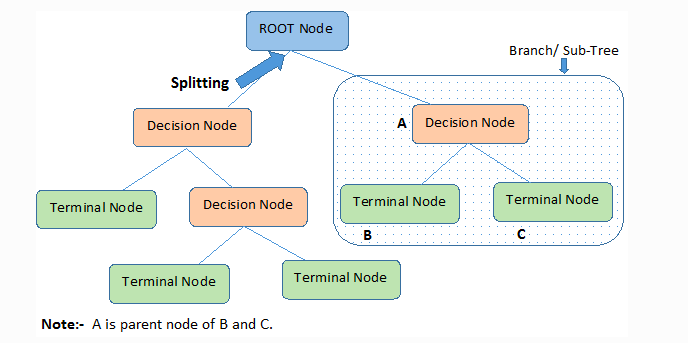
Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

**Leaf / Terminal Node:** Nodes do not split is called Leaf or Terminal node.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

**Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example.

Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node.

## Advantages of the Decision Tree

* It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
* It can be very useful for solving decision-related problems.
* It helps to think about all the possible outcomes for a problem.
* There is less requirement of data cleaning compared to other algorithms.

## Disadvantages of the Decision Tree

* The decision tree contains lots of layers, which makes it complex.
* It may have an overfitting issue, which can be resolved using the **Random Forest algorithm.**
* For more class labels, the computational complexity of the decision tree may increase.

**Random Forest Algorithm**

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

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

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

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

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables*** as in the case of regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.

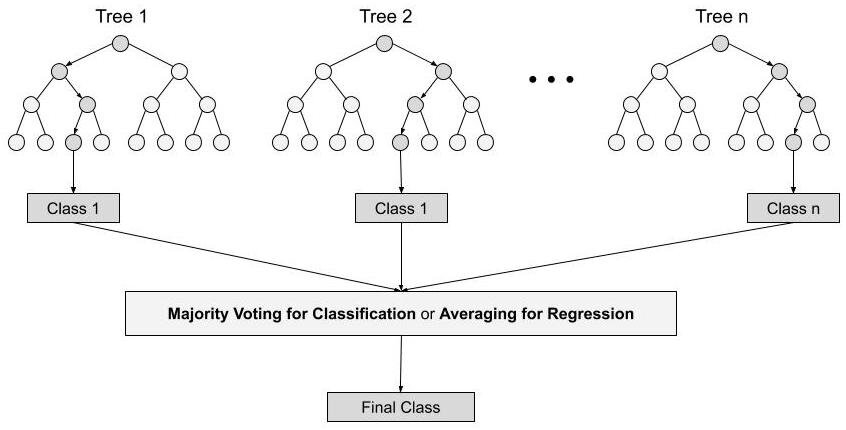
**Steps involved in random forest algorithm:**

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.



For example:  consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.



## Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

## Advantages of Random Forest

* Random Forest is capable of performing both Classification and Regression tasks.
* It is capable of handling large datasets with high dimensionality.
* It enhances the accuracy of the model and prevents the overfitting issue.

## Disadvantages of Random Forest

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks

**K Nearest Neighbor**

**K Nearest Neighbor** 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.

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

* K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
* **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

## Advantages of KNN Algorithm:

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

## Disadvantages of KNN Algorithm:

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.

**Naïve Bayes algorithm**

**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.

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

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
* It is mainly used in *text classification* that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
* Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

### Advantages of Naïve Bayes Classifier:

* Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
* It can be used for Binary as well as Multi-class Classifications.
* It performs well in Multi-class predictions as compared to the other Algorithms.
* It is the most popular choice for **text classification problems**.

### Disadvantages of Naïve Bayes Classifier:

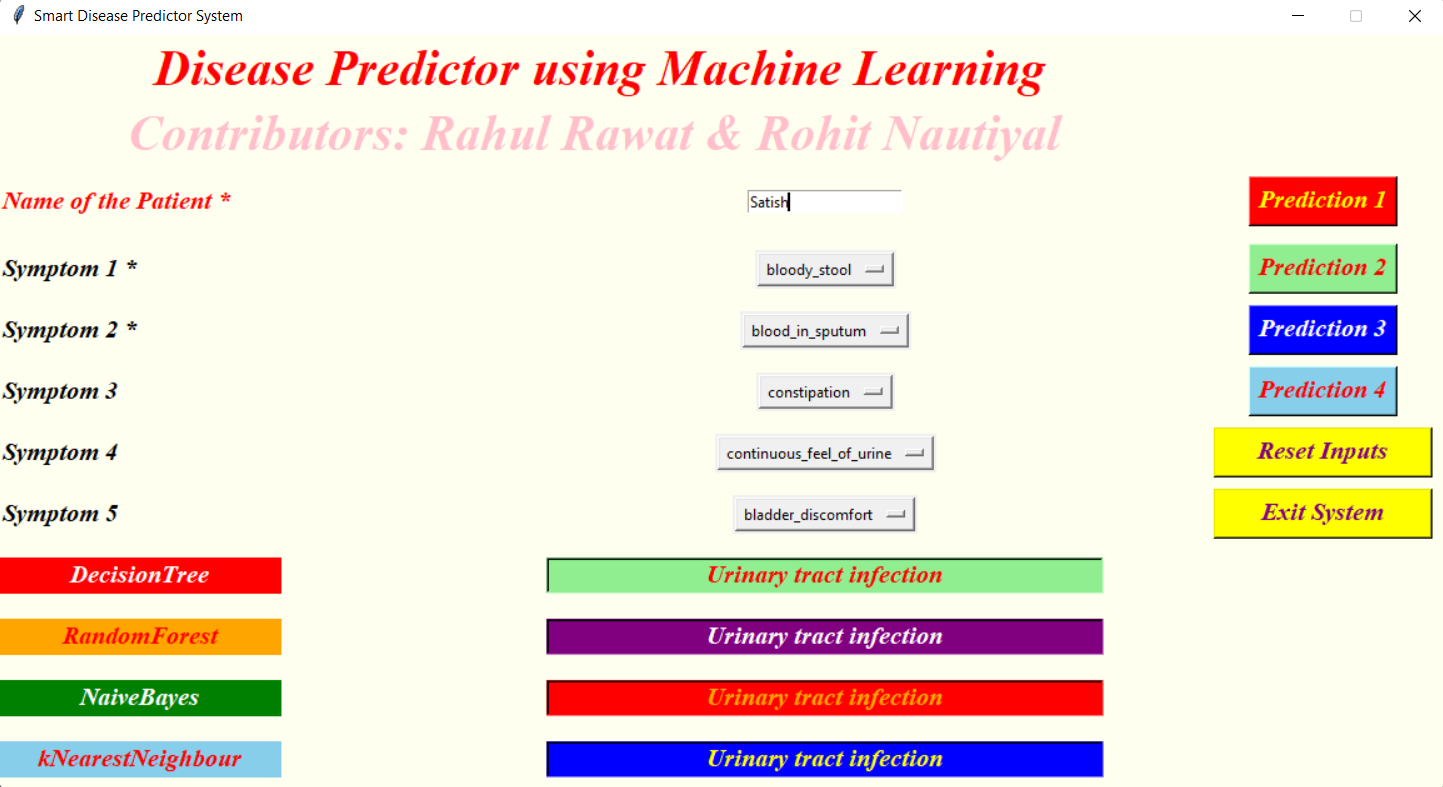
* Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

### Applications of Naïve Bayes Classifier:

* It is used for **Credit Scoring**.
* It is used in **medical data classification**.
* It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner.
* It is used in Text classification such as **Spam filtering** and **Sentiment analysis**

**GUI**

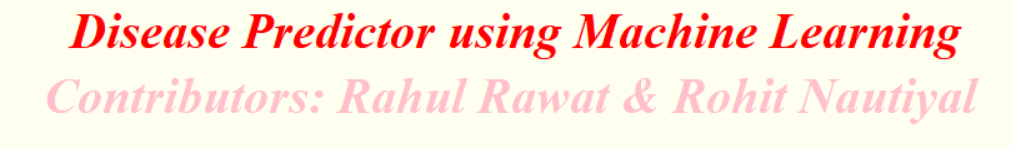
GUI made for this project is a simple tkinter GUI consisting of labels, messagebox, button, text, title and option menu



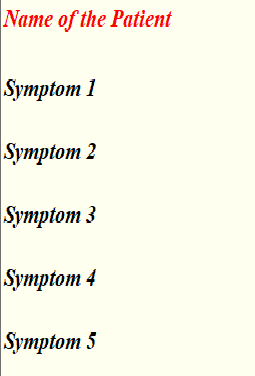
Root.title() is used to set the the title as Smart Disease Predictor System



Label is used to add heading and contributors section.



Labels are further used for different sections



OptionMenu is used to create drop down menu

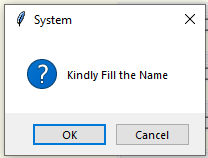


Buttons are used to give functionalities and predict the out come of models also two utility buttons namely exit and rest are also created.

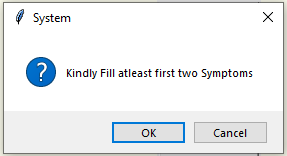


Text is used to show output of the prediction using blank space.

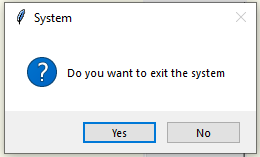
Messagebox are used at three different places, one- to restrain the to enter name



two- to ask for at least two symptoms,



three- to confirm to exit system.

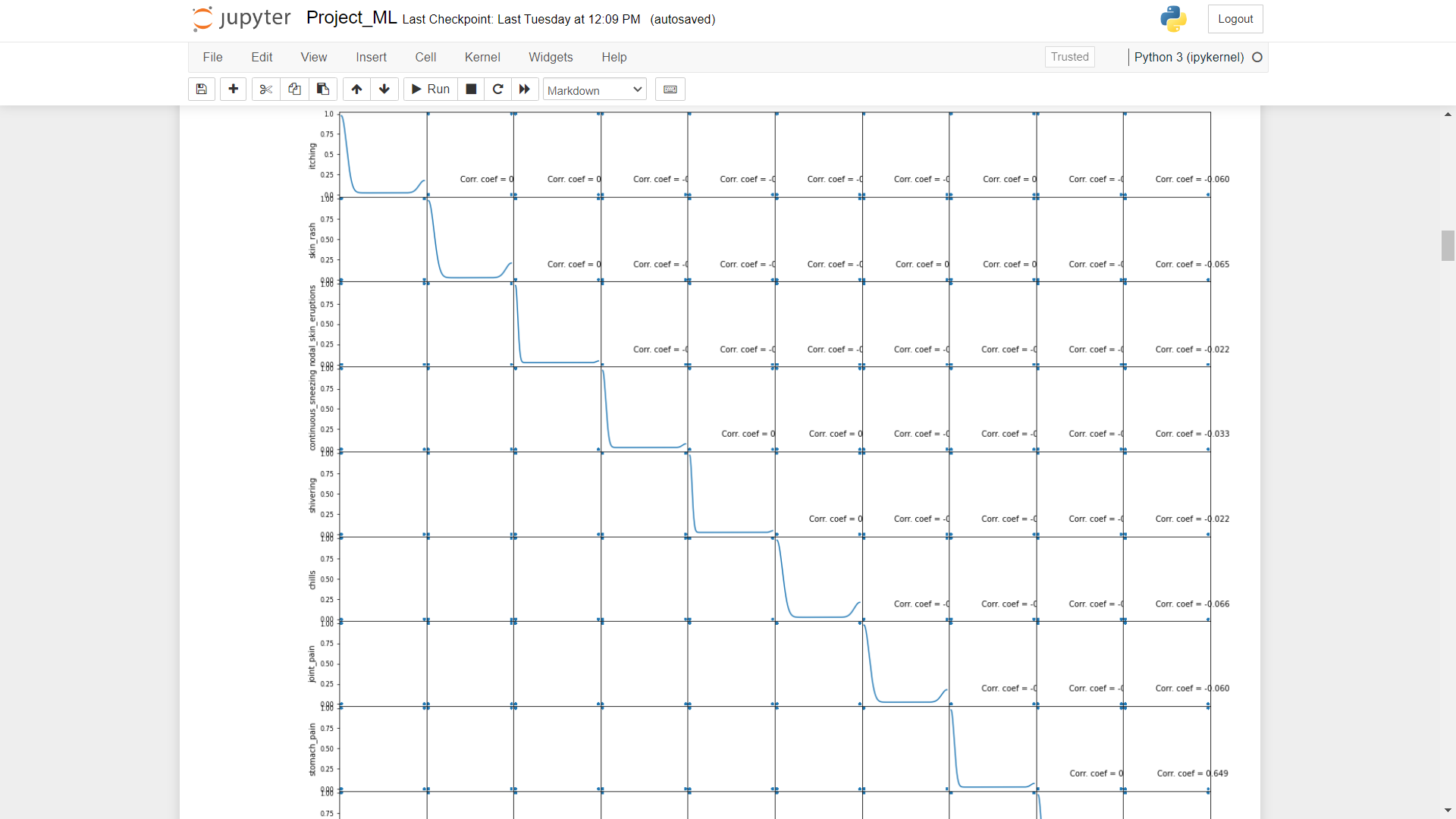


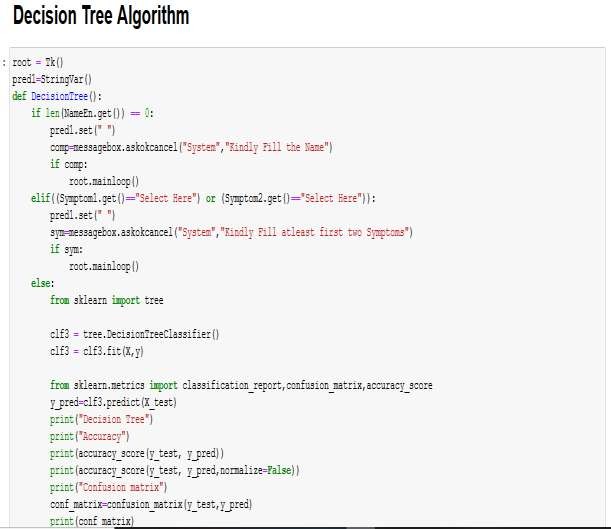
# Modules

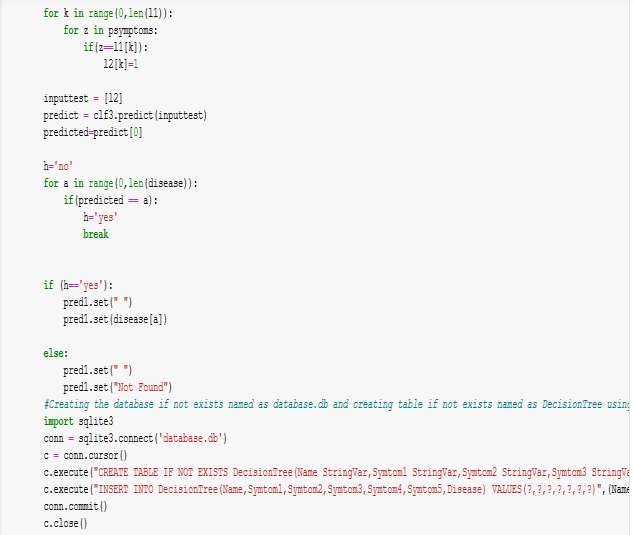
Functions like plotpercolumbdistribution() plotScatterMatrix() is used to viaualize the data.



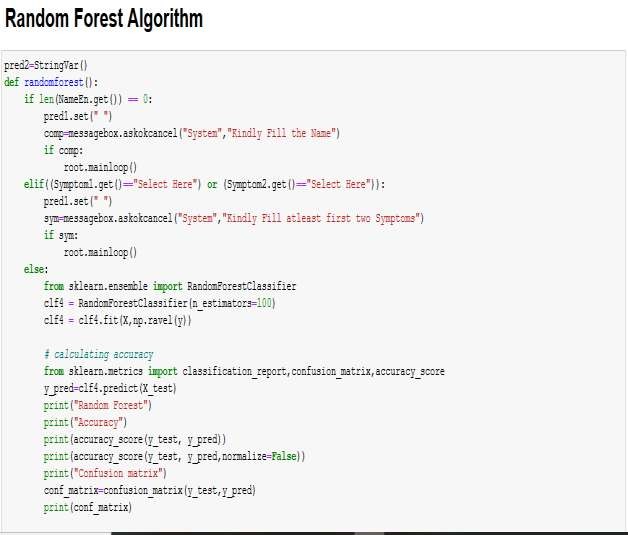
Function like scatterplt and scatterinp are used to compare input to training data.





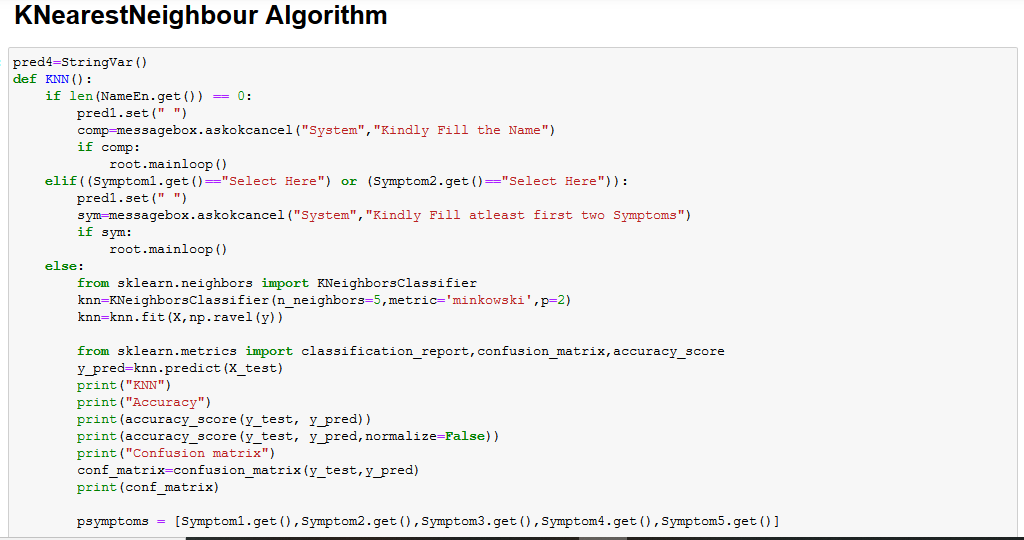


**Algorithm of decision tree and database storage.**

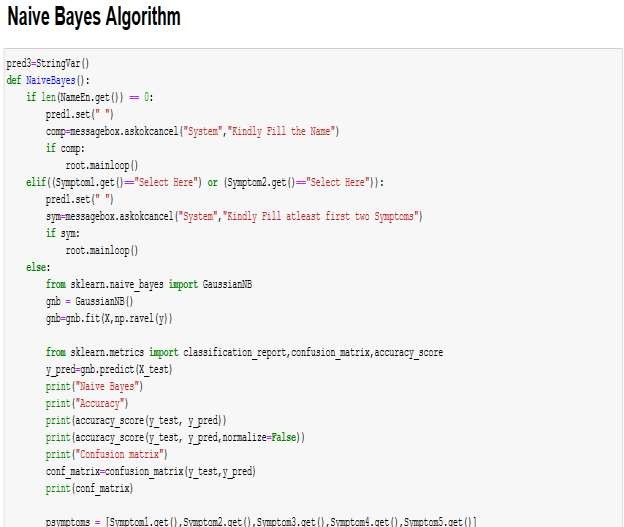


**Algorithm of random forest classifier.**

* .

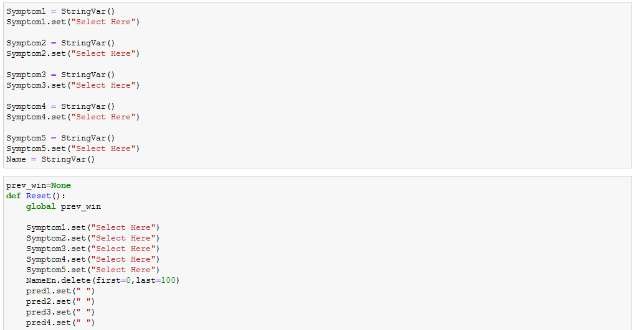


**Algorithm of K nearest neighbor.**

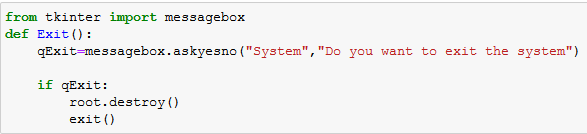


**Algorithm of naïve bayes classifier**

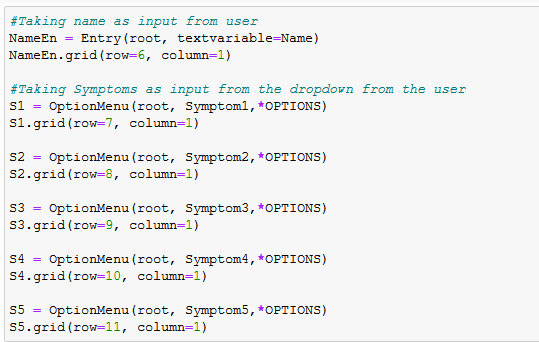
All these classifier is connected to database and GUI to function seamlessly.



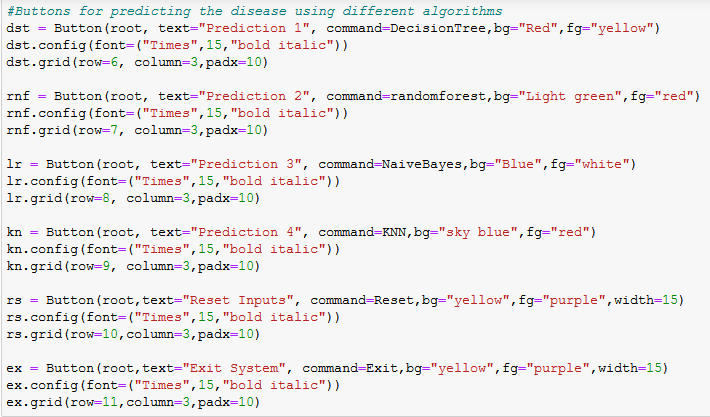
**Code of GUI to set initial values of labels**.



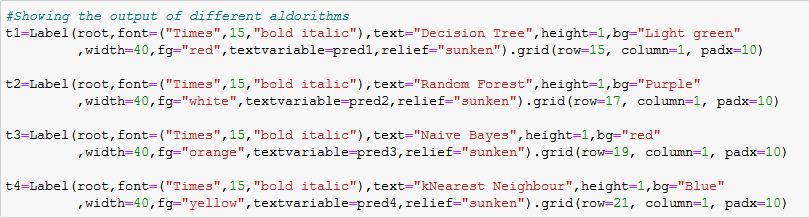
**Code of message box.**



**Code of option menu**



**Code of buttons.**



**Code of result display.**

# Conclusions

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 ~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.

Technology has ushered numerous ways to drive mankind towards a better

world, a better life. Mankind will be better off if technology is blended

into our lifestyle. People rely on technology to find out solutions for problems

they cannot solve by themselves. Health related issues are one of those

where automated help can greatly benefit the attention seeker as the person is

getting the necessary information by just a few clicks. In this work, we show

that „Automated Disease Prediction System‟ can help people who are

facing difficulties, better understand their physical condition by

predicting potential diseases. We also show that our framework enables

the system to perform significantly better than existing ones. Having said that,

our system accuracy can be increased further as there is space left for

improvement. Like the decision tree and parent tree generation is a cumbersome

task but it is a continuous process, same goes with the enrichment of

the database. It will get better and better over time and accuracy of disease

prediction will also be on the rise.

# References

1. <http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html>
2. Pew Research centre health fact sheet :

www.pewinternet.org/fact-sheets/health-fact-sheet.

1. http://edition.cnn.com/2012/05/04/tech/socialmedia/

facebook-lies-privacy [Accessed 07/06/2015]

1. Xiaoyan Wang, Amy Chused, Nomie Elhadad, Carol

Friedman, and Marianthi Markatou : “Automated