



DISEASE PREDICTION WITH SYMPTOMS USING MACHINE LEARNING

A MINI PROJECT REPORT

Submitted by

SARANAMALAR V (1919103100)
SRI HARI R (1919103118)

In partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

In

ELECTRONICS AND COMMUNICATION ENGINEERING

SONA COLLEGE OF TECHNOLOGY, SALEM 636 005

(AUTONOMOUS)

ANNA UNIVERSITY: CHENNAI 600 025

JUNE 2022

SONA COLLEGE OF TECHNOLOGY, SALEM
(AUTONOMOUS)

BONAFIDE CERTIFICATE

Certified that this project report titled “**Disease Prediction with Symptoms using Machine Learning**” is the bonafide work of “ **SARANAMALAR V (1919103100)** and **SRI HARI R(1919103118)**” who carried out the work under my supervision.

SIGNATURE

Dr.R.S. SABEENIAN

HEAD OF THE DEPARTMENT

Professor,
Department of ECE,
Sona College of Technology,
Salem-636005

SIGNATURE

Prof.N.S.YOGANATHAN

SUPERVISOR

Assistant Professor,
Department of ECE,
Sona College of Technology,
Salem-636005

Submitted for the Project Viva-Voce examination held on _____

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

At this pleasing moment of having successfully completed our mini project, we wish to convey our sincere thanks and gratitude to our parents and the college management our chairman **Shri C. VALLIAPPA**, and our vice chairman **Shri. CHOCKO VALLIAPPA** and **Shri. THYAGU VALLIAPPA** who provided all the facilities to us.

We would like to express our sincere thanks to our principal, **Dr.S.R.R. SENTHIL KUMAR**, for motivating us to do our project and for offering adequate duration for completing our project.

We are also immensely grateful to our Head of the Department **Dr.R.S. SABEENIAN, M.E., Ph.D.**, for his constructive suggestion and encouragement during our project.

With deep sense of gratitude, we extend our earnest and sincere thanks to our project guide and co-ordinator **Prof.N.S.YOGANATHAN** Assistant Professor, Department of ECE for his kind guidance and encouragement during this project.

We also express our debt thanks to our Teaching and Non-teaching staffs Department of Electronics and Communication Engineering, Sona College of Technology.

Finally, we take this opportunity to extend our deep appreciation to our family and friends, for all that they meant to us during the crucial times of the completion of our project.

ABSTRACT

In this project, Disease Prediction using Machine Learning is the system that will predict the diseases from the symptoms which are given by the users. The system processes the symptoms provided by the patients as input and gives the output as the probability of the disease. Machine Learning algorithms is used in the prediction of the disease. All used algorithms are supervised machine learning algorithm. The probability of the disease is calculated by the 4 Algorithms . The accurate analysis of medical database benefits in early prediction of disease, patient care and community services. The techniques of machine learning have been successfully employed in various applications including Disease prediction. The ultimate aim of developing classifier system using machine learning algorithms is to immensely help to solve the health related issues by assisting the physicians to predict and diagnose diseases at an early stage. A Sample data of 4920 patients' records diagnosed with 40 diseases is selected for analysis. A class label was composed of 40 diseases. 95 of 132 independent variables(symptoms) closely related to diseases selected and optimized. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree , Random Forest , K-Nearest Neighbours and Naïve Bayes classifier. This paper presents the comparative study of the results of the above algorithms. With an increase in medical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. It gives an accuracy rate of 94%. Totally 5 symptoms and a minimum of 2 symptoms can be given as input. If the user give at least 2 symptoms as input, the system will predict the disease associated with the following symptoms and show to the respective user

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO.
	ABSTRACT	iv
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	viii
1	INTRODUCTION	1
	1.1 Introduction to Machine Learning	1
	1.2 How does Machine Learning work?	
	1.3 Machine Learning Methods	2
	1.3.1 Decision Tree	2
	1.3.2 Random Forest	3
	1.3.3 KNN	3
	1.3.4 Naive Bayes	3
2	LITERATURE SURVEY	4
3	MODELS	6
4	DECISION TREE	8
	4.1 Important Terminologies Related to Decision Tree	9
	4.2 How Do decision Tree Work?	9
	4.3 Advantages and Disadvantages	11
5	RANDOM FOREST	12
	5.1 Ensemble Uses Two Types Of Methods	12
	5.2 Important Features of Random Forest	13

	5.3 Advantages and Disadvantages of Random Forest Algorithm	14
	5.3.1 Advantages	14
	5.3.2 Disadvantages	15
6	NAÏVE BAYES	16
	6.1 Why is it called Naïve Bayes?	16
	6.2 Bayes Theorem	17
	6.3 Types of Naïve Bayes Classifier	17
	6.4 Advantages of Naive Bayes classifier	19
	6.5 Disadvantages of Naive Bayes classifier	19
7	K-NEAREST NEIGHBOR	20
	7.1 How does KNN work?	20
	7.2 Advantages of KNN algorithm	22
	7.3 Disadvantages of KNN algorithm	23
8	DATASETS	24
	8.1 Diseases	24
	8.2 Symptoms	25
9	LIBRARY USED	28
10	GUI	31
11	RESULTS AND DISCUSSION	35
12	CONCLUSION	41
	REFERENCE	42

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.1	Block Diagram of Machine Learning Processing	2
4.1	Flow Chart of Decision Tree	9
6.1	Gaussian Distribution of Naïve Byes Classifier	18
7.1	Graphical Representation of Data points	21
7.2	Graphical Representation of Euclidean Distance	21
7.3	New Data Point	22
8.1	Training Datasets	24
10.1	GUI Interface	31
10.2	Heading Label	31
10.3	Section Label	32
10.4	Option Menu	32
10.5	Buttons	33
10.6	Output	33
10.7	Message box	34
11.1	Execution Output	35
11.2	Execution Output	36
11.3	Execution Output	37
11.4	Execution Output	38
11.5	Execution Output	39
11.6	Execution Output	40

LIST OF ABBREVIATIONS

ML	Machine Learning
RF	Random Forest
KNN	K-Nearest Neighbor
GUI	Graphical User Interface
GERD	Gastroesophageal Reflux Disease
AIDS	Acquired Immune Deficiency Syndrom

CHAPTER 1

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.

1.1 Introduction of 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**.

1.2 How does Machine Learning work:

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

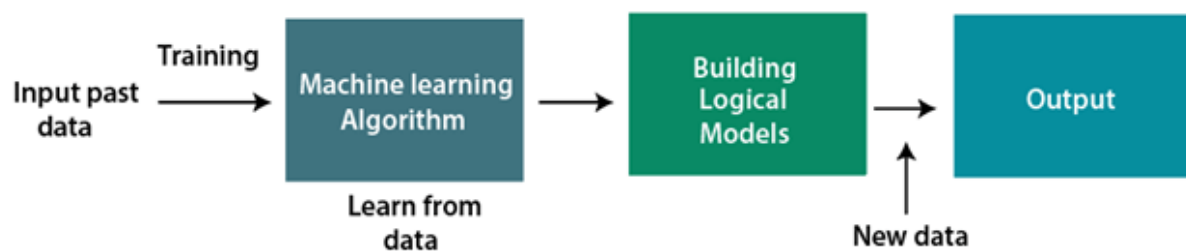


FIG 1.1 BLOCK DIAGRAM OF MACHINE LEARNING PROCESSING

1.3 Machine Learning Methods:

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are **supervised learning** [1] which trains algorithms based on example input and output data that is labeled by humans, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data. Let's explore these methods in more detail.

1.3.1 DECISION TREE:

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. 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.

1.3.2 RANDOM FOREST ALGORITHM :

Random Forest is essentially a collection of Decision trees. In this algorithm, firstly the random samples are selected from a given dataset. Then a decision tree is constructed for every dataset. Then the prediction result is obtained from every dataset. Voting is performed for every predicted result. Finally the most voted prediction result is selected as final prediction result.

1.3.3 KNN:

K Nearest Neighbour is a supervised learning algorithm. It is a basic yet essential algorithm. It finds extensive use in pattern finding and data mining. It works by finding a pattern in data which links data to results and it improves upon the pattern recognition with every iteration.

1.3.4 NAIVE BAYES:

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.

CHAPTER 2

LITERATURE SURVEY

Marimuthu et al. [1] aimed to predict heart diseases using supervised ML techniques. The authors structured the attributes of data as gender, age, chest pain, gender, target and slope.

Vahid et al. [2], where the Logistic Regression outperformed other techniques such as ANN, SVM, and Adaboost. The studies excelled in conducting an extensive analysis on the ML models.

A. S. Nagdive et al.[3], “Comparative study of machine learning algorithms for breast cancer prediction.

Dahiwade et al. [4] proposed a ML based system that predicts common diseases. The symptoms dataset was imported from the UCI ML depository, where it contained symptoms of many common diseases.

Aditi Gavhane et.[5] al proposed a paper “Prediction of Heart Disease Using Machine Learning”, in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.

MIN CHEN et al, [6] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

Lambodar Jena et al, [7] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively

Dhomse Kanchan B. et al, [8] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for Heart disease

Senthilkumar Mohan et al, [9] focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%.

Rashmi G Saboji et al, [10] tried to find a scalable solution that can predict heart disease utilizing classification mining and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifier but Random Forest gives more accurate results with accuracy 98%.

CHAPTER 3

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 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. 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.

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.

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 pattern recognition with every iteration.

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.

CHAPTER 4

DECISION TREE

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:^[1]

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

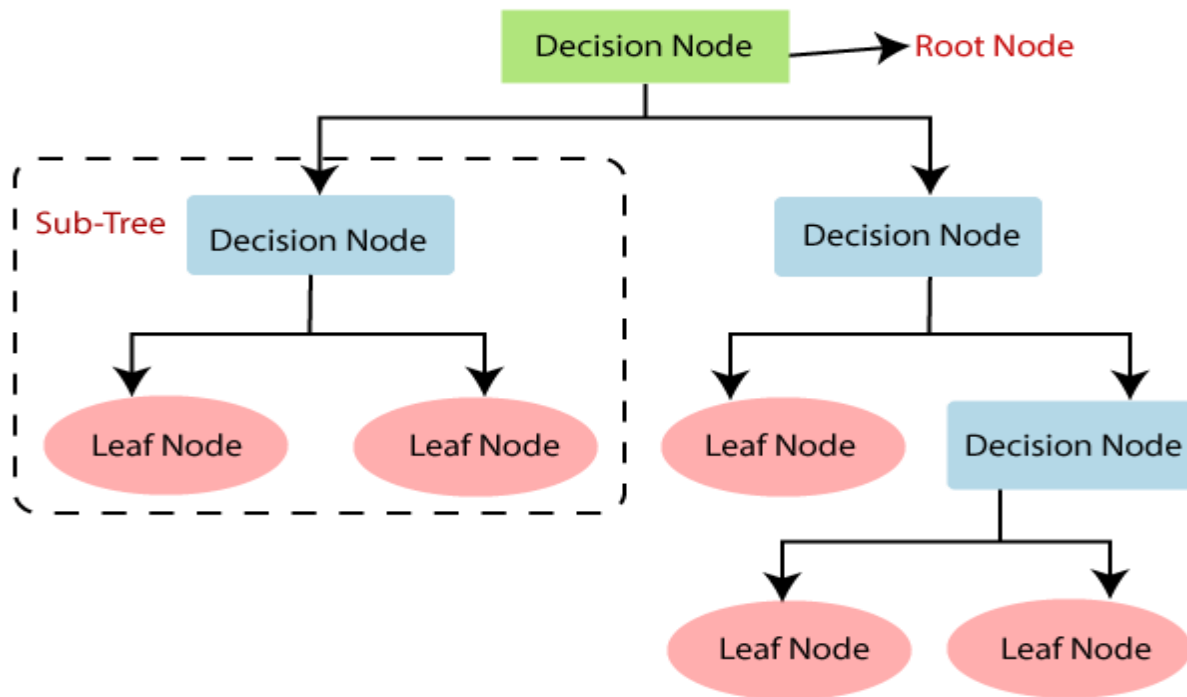


FIG 4.1 FLOWCHART OF DECISION TREE

4.1 Important Terminology related to Decision Trees

1. **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
4. **Leaf / Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
6. **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.

7. **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.

4.2 How do Decision Trees work?

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes.[3] In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

The algorithm selection is also based on the type of target variables. Let us look at some algorithms used in Decision Trees:

ID3 → (extension of D3)

C4.5 → (successor of ID3)

CART → (Classification And Regression Tree)

CHAID → (Chi-square automatic interaction detection Performs multi-level splits when computing classification trees)

MARS → (multivariate adaptive regression splines)

The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking. A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment.

4.3 Advantages and Disadvantages: Among decision support tools, decision trees (and influence diagrams) have several advantages. Decision trees:

- Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
- Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
- Help determine worst, best, and expected values for different scenarios.
- Use a white box model. If a given result is provided by a model.
- Can be combined with other decision techniques.
- The action of more than one decision-maker can be considered.

4.4 Disadvantages of decision trees:

- They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
- They are often relatively inaccurate. Many other predictors perform better with similar data. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree.
- For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favour of those attributes with more levels.
- Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.

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

CHAPTER 5

RANDOM FOREST

Random Forest Algorithm is a supervised learning algorithm used for both classification and regression. This algorithm works on 4 basic [2]

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.

5.1 Working of Random Forest Algorithm

Before understanding the working of the random forest we must look into the ensemble technique. *Ensemble* simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

5.1.1 Ensemble uses two types of methods:

1. **Bagging** – It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. **Boosting** – It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

Bagging:

Bagging, also known as ***Bootstrap Aggregation*** is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as ***row sampling***. This step of row sampling with replacement is called ***bootstrap***. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as ***aggregation***.

5.2 Importance features of Random Forest:

- 1. Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- 2. Immune to the curse of dimensionality-** Since each tree does not consider all the features, the feature space is reduced.
- 3. Parallelization-** Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- 4. Train-Test split-** In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

5. Stability- Stability arises because the result is based on majority voting/averaging.

5.3 Advantages and Disadvantages of Random Forest Algorithm:

5.3.1 Advantages :

1. It can be used in classification and regression problems.
2. It solves the problem of overfitting as output is based on majority voting or averaging.
3. It performs well even if the data contains null/missing values.
4. Each decision tree created is independent of the other thus it shows the property of parallelization.
5. It is highly stable as the average answers given by a large number of trees are taken.
6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
7. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
8. We don't have to segregate data into train and test as there will always be 30% of the data which is not seen by the decision tree made out of bootstrap.

5.3.2 Disadvantages:

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.
2. Training time is more compared to other models due to its complexity. Whenever it has to make a prediction each decision tree has to generate output for the given input data.

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

CHAPTER 6

NAIVE BAYES

- Naive Bayes algorithm is a supervised learning algorithm, which is based on **Bayes** and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naive Bayes Classifier is one of the simple and most effective Classification algorithms[7] 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 theorem .**
- Some popular examples of Naive Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles.**

6.1 Why is it called Naive Bayes?

The Naive Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- **Naïve:** It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

6.2 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:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

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.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

6.3 Types of Naive Bayes Classifiers

1. Multinomial Naive Bayes Classifier

Feature vectors represent the frequencies with which certain events have been generated by a **multinomial distribution**. This is the event model typically used for document classification.

2. Bernoulli Naive Bayes Classifier:

In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks,[8] where binary term occurrence (i.e. a word occurs in a document or not) features are used rather than term frequencies (i.e. frequency of a word in the document).

3. Gaussian Naive Bayes Classifier:

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution** (Normal distribution). When plotted, it gives a bell-shaped curve which is symmetric about the mean of the feature values

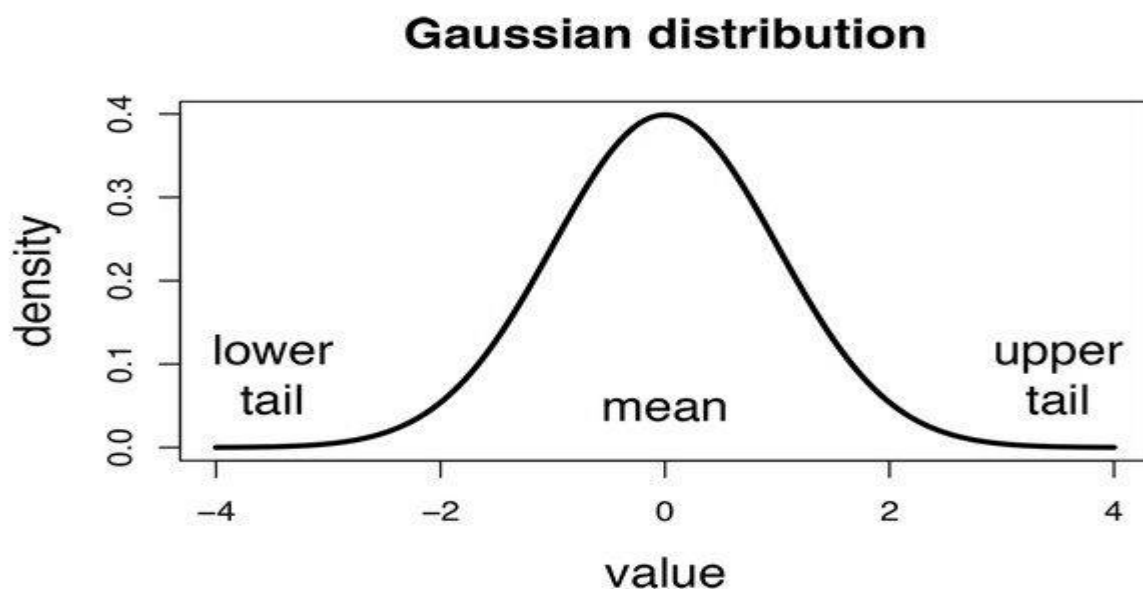


FIG 6.1 GAUSSIAN DISTRIBUTION OF NAIVE BAYES CLASSIFIER

6.4 Advantages of Naive 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**.

6.5 Disadvantages of Naive Bayes Classifier:

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

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

CHAPTER 7

K-NEAREST NEIGHBOUR

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 pattern recognition with every iteration.

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

7.1 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 neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, 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 neighbor is maximum.
- **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

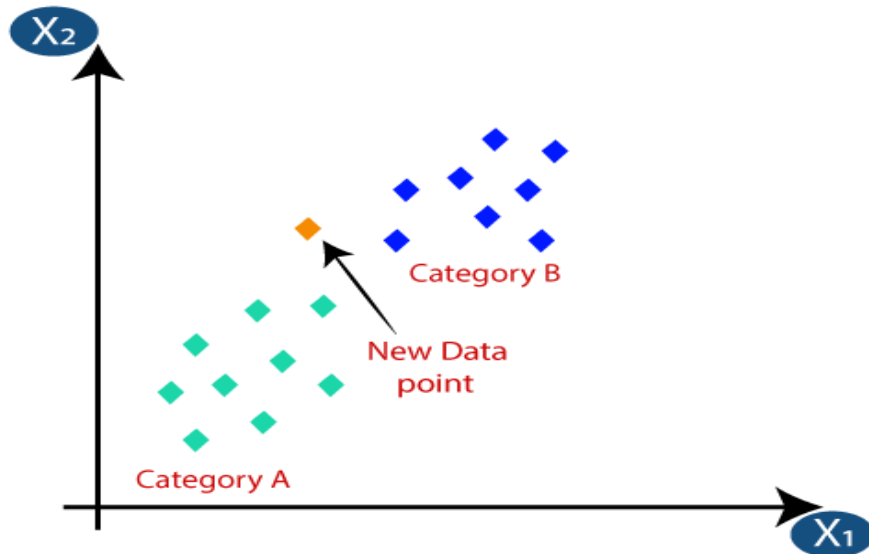
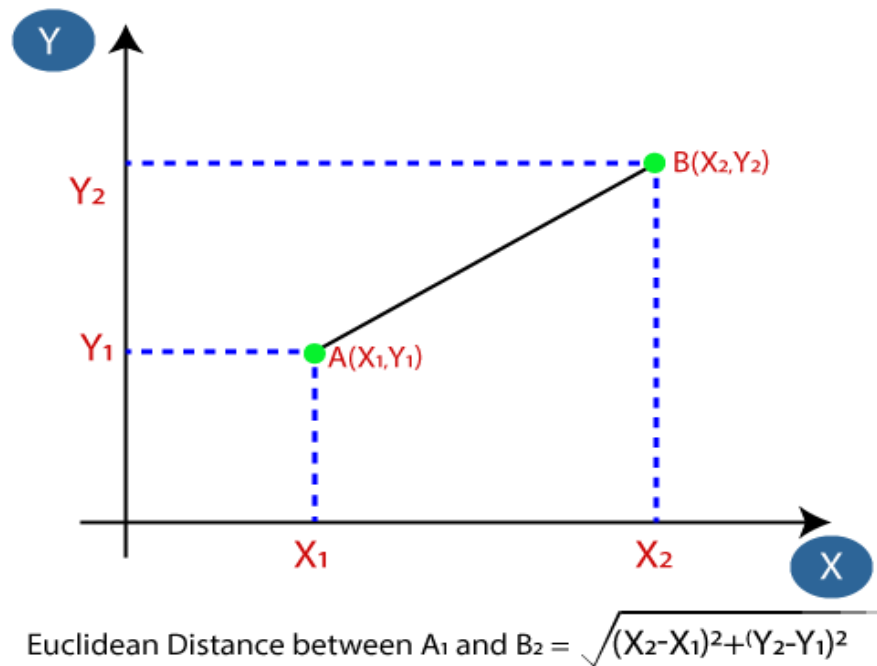


FIG 7.1 GRAPHICAL REPRESENTATION OF DATA POINTS

- Firstly, we will choose the number of neighbors, so we will choose the $k=5$.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



○ FIG 7.2 GRAPHICAL REPRESENTATION OF EUCLIDEAN DISTANCE

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:

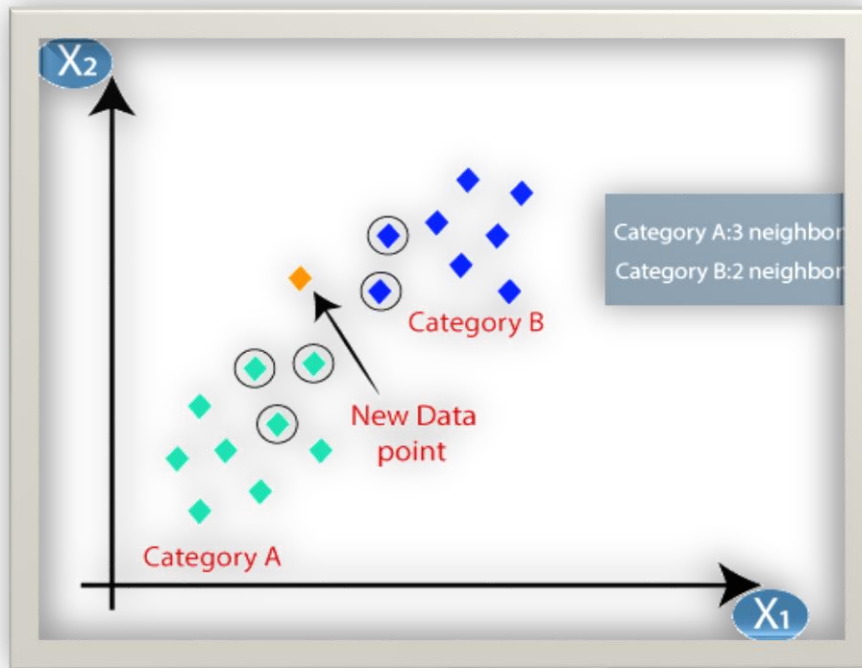


FIG 7.3 NEW DATA POINTS

- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.[5]

7.2 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.

7.3 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.

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

CHAPTER 8

DATASETS

Our data sets consists of symtoms as labels and diseases as objects. 80% of datas was used for training and remaining 20% was used as testing datas.

EC21	X	✓	f _x	Allergy																	
	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED
1	coma	stomach_distention	history_of	fluid_over	blood_in	prominen	palpitatio	painful_w	pus_filled	blackhead	scurring	skin_peel	silver_like	small_der	inflamma	blister	red_sore	yellow_cr	prognosis		
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Fungal infection	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 Allergy	
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 GERD	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 GERD	

training

+

FIG 8.1 TRAINING DATASETS

8.1 DISEASES:

1. Fungal infection
2. Allergy
3. GERD
4. Chronic cholestasis
5. Drug Reaction
6. Peptic ulcer disease
7. AIDS
8. Diabetes
9. Gastroenteritis
10. Bronchial Asthma

- 11.Hypertension
- 12.Migraine
- 13.Cervical spondylosis
- 14.Paralysis (brain haemorrhage)
- 15.Jaundice
- 16.Malaria
- 17.Chicken pox
- 18.Dengue
- 19.Typhoid
- 20.hepatitis A
- 21.Hepatitis B
- 22.Hepatitis C
- 23.Hepatitis D
- 24.Hepatitis E
- 25.Alcoholic hepatitis
- 26.Tuberculosis
- 27.Common Cold
- 28.Pneumonia
- 29.Dimorphic haemorrhoids(piles)
- 30.Heart attack
- 31.Varicose veins
- 32.Hypothyroidism
- 33.Hyperthyroidism
- 34.Hypoglycaemia
- 35.Osteoarthritis
- 36.Arthritis
- 37.(vertigo) Paroxysmal Positional Vertigo
- 38.Acne
- 39.Urinary tract infection
- 40.Psoriasis
- 41.Impetigo

8.2 SYMPTOMS:

1. Backpain
2. Constipation
3. Adominalpain
4. Diarrhoea
5. Mildfever
6. Yellowurine
7. Yellowing of eyes
8. Acute liver failure
9. Fluid overload
- 10.Swelling of stomach

- 11.Swelled lymph nodes
- 12.Malaise
- 13.Blurred and distorted vision
- 14.Phlegm
- 15.Throat irritation
- 16.Redness of eyes
- 17.Sinus pressure
- 18.Runny nose
- 19.Congestion
- 20.Chest pain
- 21.Weakness in limbs
- 22.Fast heart rate
- 23.Pain during bowel movements
- 24.Pain in anal region
- 25.Bloody stool
- 26.Irritation in anus
- 27.Neck pain
- 28.Dizziness
- 29.cramps
- 30.30.Bruising
- 31.Obesity
- 32.Swollen legs
- 33.Puffy face and eyes
- 34.enlarged thyroid
- 35.brittle nails
- 36.swollen extremities
- 37.excessive hunger
- 38.extra marital contacts
- 39.drying and tingling lips
- 40.slurred speech
- 41.knee pain
- 42.hip joint pain
- 43.muscle weakness
- 44.stiff neck
- 45.swelling joints
- 46.movement stiffness
- 47.spinning movements
- 48.loss of balance
- 49.unsteadiness
- 50.weakness of one body side
- 51.loss of smell
- 52.bladder discomfort
- 53.foul smell of urine

54.continuous feel of urine
55.passage of gases
56.internal itching
57.toxic look (typhos)
58.depression
59.irritability
60.muscle pain
61.altered sensorium
62.red spots over body
63.belly pain
64.abnormal menstruation
65.dischromic patches
66.watering from eyes
67.increased appetite
68.polyuria
69.family history
70.mucoid sputum
71.rusty sputum
72.lack of concentration
73.visual disturbances
74.receiving blood transfusion
75.receiving unsterile injections
76.coma
77.stomach bleeding
78.distention of abdomen
79.history of alcohol consumption
80.fluid overload
81.blood in sputum
82.prominent veins on calf
83.palpitations
84.painful walking
85.pus filled pimples
86.blackheads
87.scurring
88.skin peeling
89.silver like dusting
90.small dents in nails
91.inflammatory nails
92.blister
93.red sore around nose
94.yellow crust ooze

CHAPTER 9

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.

2. **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.

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 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.

4. **sklearn:** Sklearn is an open source python library with implements a huge range of machinelearning, 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.

CHAPTER 10

GUI

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

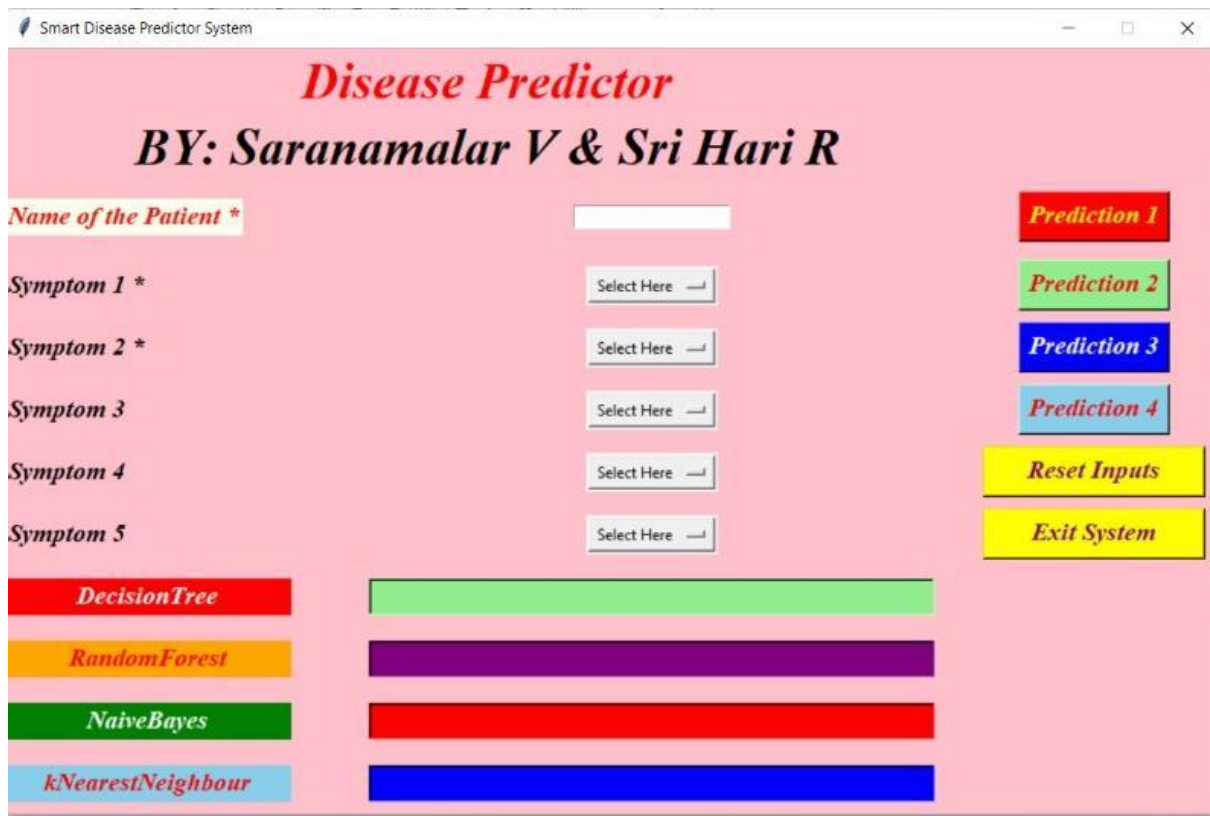


FIG 10.1 GUI INTERFACE

Label is used to add heading and contributors section

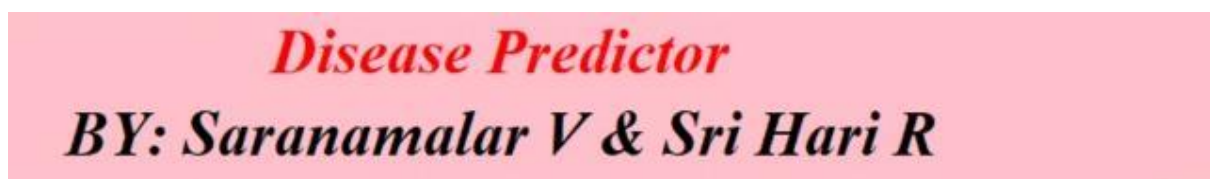
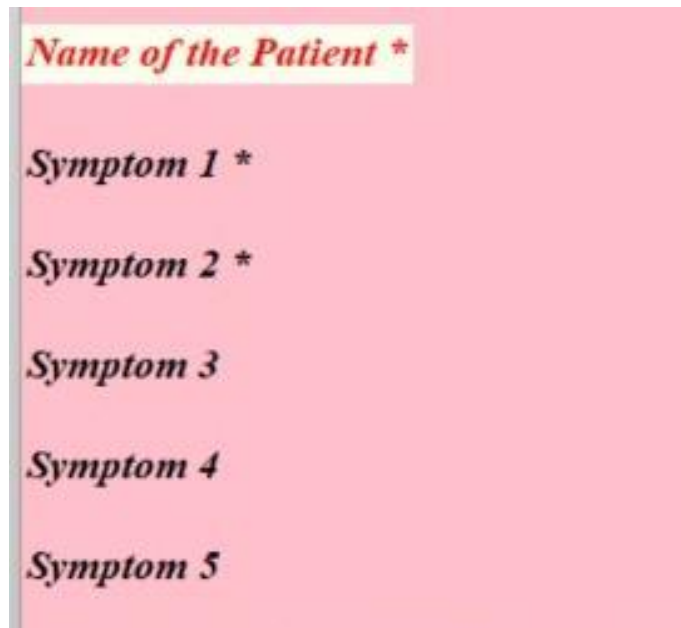


FIG 10.2 HEADING LABEL

Labels are further used for different sections



*Name of the Patient **

*Symptom 1 **

*Symptom 2 **

Symptom 3

Symptom 4

Symptom 5

FIG 10.3 SECTION LABEL

OptionMenu is used to create drop down menu



Select Here ▾

Select Here ▾

Select Here ▾

Select Here ▾

Select Here ▾

FIG 10.4 OPTION MENU

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



FIG 10.5 BUTTONS

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

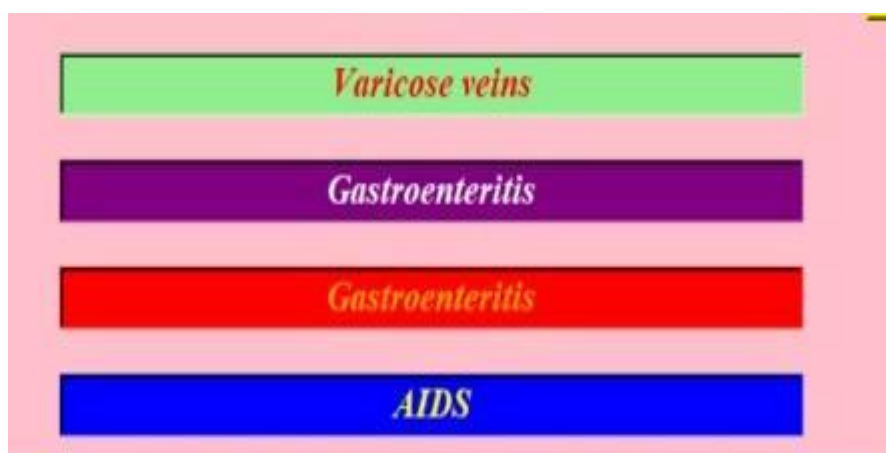


FIG 10.6 OUTPUT

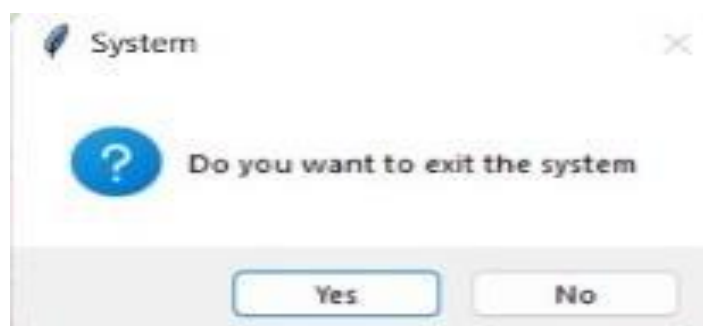
Messagebox are used at three different places, one- to restrain the to enter name



two- to ask for at least two symptoms,



three- to confirm to exit system



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

CHAPTER 11

RESULTS AND DISCUSSIONS

PATIENT NAME: John

SYMPTOMS: 1.Abdominal pain
2.Mild Fever
3.Yellow Urine

The screenshot displays a web application titled "Smart Disease Predictor System". The main heading is "Disease Predictor" by "Saranamalar V & Sri Hari R". The interface includes input fields for patient name, five symptoms, and a section for model predictions. The patient name is "John". The symptoms entered are "abdominal_pain", "mild_fever", and "yellow_urine". The prediction section shows four colored buttons labeled "Prediction 1" through "Prediction 4", all of which display the word "Jaundice". There are also "Reset Inputs" and "Exit System" buttons. At the bottom, a table lists four machine learning models: DecisionTree, RandomForest, NaiveBayes, and kNearestNeighbour, each with a corresponding prediction of "Jaundice".

Model	Prediction
DecisionTree	Jaundice
RandomForest	Jaundice
NaiveBayes	Jaundice
kNearestNeighbour	Jaundice

FIG 11.1 EXECUTION OUTPUT

All the four algorithms Decision Tree, Random Forest, NaiveBayes and KNN have showed the same result JAUNDICE.

PATIENT NAME: Regina

SYMPTOMS: 1.Loss of smell
2.Runny Nose
3.Chest Pain

The screenshot displays a web-based application titled "Disease Predictor" by Saranamalar V & Sri Hari R. The interface includes input fields for patient name, five symptoms, and a dropdown for model selection. On the right, there are buttons for "Prediction 1" through "Prediction 4", "Reset Inputs", and "Exit System". Below the inputs, a table shows the predicted diseases for each model.

Model	Predicted Disease
DecisionTree	Common Cold
RandomForest	Common Cold
NaiveBayes	GERD
kNearestNeighbour	Heart attack

FIG 11.2 EXECUTION OUTPUT

The prediction of decision tree is common cold the prediction of random forest is common cold the prediction of NaiveBayes is GERD and the prediction of KNN is heart attack out of all KNN predicted the correct disease that is heart attack.

PATIENT NAME: Rachel Green

SYMPTOMS: 1.Dizziness

2.Loss of balance

3.Lack of concentration

FIG 11.3 EXECUTION OUTPUT

The prediction of decision tree is common cold, the prediction of random forest is hypertension, the prediction of NaiveBayes is hypertension, and the prediction of KNN is hypertension ,out of all random forest, NaiveBayes and KNN predicted the disease correctly that is hypertension.

PATIENT NAME: Regina Phalange

SYMPTOMS: 1.Malaise
2.Swelled lymph nodes
3.Mild fever

The screenshot displays a web-based application titled "Disease Predictor" by "Saranamalar V & Sri Hari R". The interface includes input fields for patient name and five symptoms, a list of machine learning models, and a column of prediction results. The patient name is "regina phalange". The symptoms entered are "malaise", "swelled_lymph_nodes", "mild_fever", and two "Select Here" options. The models listed are DecisionTree, RandomForest, NaiveBayes, and kNearestNeighbour. The predictions for these models are "Common Cold", "Common Cold", "Chicken pox", and "Chicken pox" respectively. On the right side, there are buttons for "Prediction 1" through "Prediction 4", "Reset Inputs", and "Exit System".

Input Field	Value	Prediction
Name of the Patient *	regina phalange	Prediction 1
Symptom 1 *	malaise	Prediction 2
Symptom 2 *	swelled_lymph_nodes	Prediction 3
Symptom 3	mild_fever	Prediction 4
Symptom 4	Select Here	Reset Inputs
Symptom 5	Select Here	Exit System
DecisionTree	Common Cold	
RandomForest	Common Cold	
NaiveBayes	Chicken pox	
kNearestNeighbour	Chicken pox	

FIG 11.4 EXECUTION OUTPUT

The prediction of decision tree is common cold, the prediction of random forest is common cold the prediction of NaiveBayes is chicken pox and the prediction of KNN is chicken pox out of all NaiveBayes and KNN predicted the disease correctly that is chicken pox.

PATIENT NAME: Phoebe Buffay

SYMPTOMS: 1. Altered sensorium
2. Weakness of one body side

The screenshot displays the 'Smart Disease Predictor System' window. The title bar reads 'Smart Disease Predictor System'. The main header area is pink and contains the text 'Disease Predictor' in red and 'BY: Saranamalar V & Sri Hari R' in black. Below this, there are input fields for patient information and symptoms. The 'Name of the Patient' field contains 'phoebe buffay'. The 'Symptom 1' field contains 'altered_sensorium' and 'Symptom 2' contains 'weakness_of_one_body_side'. 'Symptom 3' and 'Symptom 4' have 'Select Here' buttons, and 'Symptom 5' has a 'Select Here' button. To the right of the input fields are four prediction buttons: 'Prediction 1' (red), 'Prediction 2' (green), 'Prediction 3' (blue), and 'Prediction 4' (light blue). Below these are two yellow buttons: 'Reset Inputs' and 'Exit System'. At the bottom, there are four colored boxes representing different machine learning models and their predictions: 'DecisionTree' (red) predicts 'Paralysis (brain hemorrhage)' (green), 'RandomForest' (orange) predicts 'Paralysis (brain hemorrhage)' (purple), 'NaiveBayes' (green) predicts 'Paralysis (brain hemorrhage)' (red), and 'kNearestNeighbour' (light blue) predicts 'Paralysis (brain hemorrhage)' (blue).

Model	Prediction
DecisionTree	Paralysis (brain hemorrhage)
RandomForest	Paralysis (brain hemorrhage)
NaiveBayes	Paralysis (brain hemorrhage)
kNearestNeighbour	Paralysis (brain hemorrhage)

FIG 11.5 EXECUTION OUTPUT

The prediction of decision tree is paralysis, The prediction of random forest is paralysis, The prediction of NaiveBayes is paralysis , The prediction of KNN is paralysis, Out of all Decision tree, Random forest, NaiveBayes all predicted the disease correctly that is paralysis.

PATIENT NAME: Joey Tribbiani

SYMPTOMS:

1. Abdominal pain
2. Diarrhoea
3. Constipation
4. Toxic look (typhos)
5. Mild fever

The screenshot shows a web application titled "Smart Disease Predictor System". The main heading is "Disease Predictor" by "Saranamalar V & Sri Hari R". The interface includes input fields for patient name and five symptoms, a list of machine learning models, and a section for predictions.

Input Field	Value
Name of the Patient *	Joey Tribbiani
Symptom 1 *	abdominal_pain
Symptom 2 *	diarrhoea
Symptom 3	constipation
Symptom 4	toxic_look_(typhos)
Symptom 5	mild_fever

Model	Prediction
DecisionTree	Paralysis (brain hemorrhage)
RandomForest	Typhoid
NaiveBayes	Typhoid
kNearestNeighbour	Typhoid

Additional buttons on the right include "Prediction 1" (red), "Prediction 2" (green), "Prediction 3" (blue), "Prediction 4" (light blue), "Reset Inputs" (yellow), and "Exit System" (yellow).

FIG 11.6 EXECUTION OUTPUT

The prediction of Decision tree is paralysis, The prediction of Random forest is typhoid, The prediction of NaiveBayes is typhoid, The prediction of KNN is typhoid, Out of all Random forest, NaiveBayes and KNN predicted the disease correctly that is typhoid.

CHAPTER 12

CONCLUSION

We set out to create a system which can predict disease on the basis of symptoms given to it. Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff. We were successful in creating such a system and use 4 different algorithm to do so. On an average we achieved accuracy of ~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

- [1] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach,” *International Journal of Computer Applications*, vol. 181, no. 18, pp. 20–25, 2018
- [2] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 204– 207.
- [3] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, “Comparative study of machine learning algorithms for breast cancer prediction,” *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, pp. 796–801, 2020
- [4] A. S. Nagdive, “Comparative study of machine learning algorithms for breast cancer prediction,” *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology-*, pp. 1287- 2017.
- [5] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning” *IEEE Xplore ISBN: 978-1-5386-0965-1*, pp. 1275-1278, 2018.
- [6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities” *IEEE Access*, vol. 5, no. 1, pp. 8869–8879, 2017.

[7] Lambodar Jena and Ramakrushna Swain, “Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers” IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017

[8] Dhomse Kanchan B. and Mahale Kishor M., “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis” IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2018.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.

[10] Rashmi G Saboji and Prem Kumar Ramesh, “A Scalable Solution for Heart Disease Prediction using Classification Mining Technique” IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.