DISEASE PREDICTOR

MINOR PROJECT II

Submitted by:

ANKIT RAJ(9917103080)
MAHENDRA SINGH(9917103170)
ASHEESH MITTAL(9917103180)
SHUBH PANDEY(9917103181)

Under the supervision of
Dr. DEVPRIYA SONI

Department of CSE/IT
Jaypee Institute of Information Technology University, Noida

MAY 2020

ACKNOWLEDGEMENT

We would like to place on record my deep sense of gratitude to Dr Devpriya Soni , CSE Faculty, Jaypee Institute of Information Technology, India for her generous guidance, help and useful suggestions.

I also wish to extend my thanks to classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Name of Students:
ANKIT RAJ (9917103080)
MAHENDRA SINGH (9917103170)
ASHEESH MITTAL (9917103180)
SHUBH PANDEY (9917103181)

Abstract

The successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. The medical environment is still information rich but knowledge weak. There is a wealth of data possible within the medical systems. However, there is a lack of powerful analysis tools to identify hidden relationships and trends in data. Disease is a term that assigns to a large number of heath care conditions related to the body. These medical conditions describe the unexpected health conditions that directly control all the body parts. God has granted each and every individual a beautiful gift called life, so it is our responsibility to live our life to fullest and try to stay safe from the dangers of the world. So we have developed a model with the help of machine learning algorithms like decision tree, random forest ,KNN ,naïve Bayes and Stacking Classifier which take into account the symptoms felt by person and according to that symptoms it predicts the disease which the person can be suffering from. It saves time as well as makes it easy to get a warning about your health before it's too late.

Table of Contents

Topic	Page No.
Acknowledgement	i
Abstract	ii
List of Figures	iii
List of Abbreviations	iv
Chapter 1: Introduction	
1.1 Introduction	1
1.2 Problem statement	1
1.3 Objective	1
Chapter 2: Background	2
Chapter 3: Requirement Analysis	3
Chapter 4: Detailed Design	4
4.1 Methodology	4
4.2 Data set	4
4.3 Algorithm used	4
4.3.1 Decision Tree	4
4.3.2 Random Forest Classifier	5
4.3.3 Naïve Bayes	5
4.3.4 KNN	6
4.3.5 Support Vector Machine	7
4.3.6 Stacking Classifier	7
4.4 System Design	8
4.4.1 State Diagram	8
4.4.2 Class Diagram	8
4.4.3 Sequential Diagram	9
4.5 Project Pipeline	9
Chapter 5: Implementation	10
5.1 Tools used	10
5.2 Data Extraction	11
5.3 Data Analysis	11
5.4 Algorithm Implemented	12
5.5 Web Platform Development	15
Chapter 6: Result and testing	16
Chapter 7: Conclusion	17
Chapter 8 : Gantt chart	18
Chapter 9: References	19

List of figures

Figure	Title	Name of figures	Page
1.	Figure 4.1	Decision Tree	4
2.	Figure 4.2	Random Forest Classier	5
3.	Figure 4.3	Naïve Bayes	6
4.	Figure 4.4	KNN	6
5.	Figure 4.5	Support Vector Machine	7
6.	Figure 4.6	Stacking Classifier	7
7.	Figure 4.7	State Diagram	8
8.	Figure 4.8	Class Diagram	8
9.	Figure 4.9	Sequential Diagram	9
10.	Figure 4.10	Project Pipeline	9
11.	Figure 5.1	Data Set	10
12.	Figure 5.2	Support Vector Machine	13
13.	Figure 5.3	Stacking Classifier	13
14.	Figure 5.4	The Home Page	14
15.	Figure 5.5	The Login Page	14
16.	Figure 5.6	The Symptoms Check Box Page	15
17.	Figure 8.1	Gantt Chart	16

CHAPTER 1: INTRODUCTION

1. 1 Introduction

At present, when one suffers from particular disease, then the person has to visit to doctor which is time consuming and costly too. Also if the user is out of reach of doctor and hospitals it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using a automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other related Disease Prediction System using data mining techniques that analyzes the risk level of the patient .

Disease Predictor is a web based application that predicts the disease of the user with respect to the symptoms given by the user. Disease Prediction system has data sets collected from different health related sites. With the help of Disease Predictor the user will be able to know the probability of the disease with the given symptoms.

1.2 Problem Statement

The purpose of this system is to provide prediction for the general and more commonly occurring disease that when unchecked can turn into fatal disease. This allows user to share their symptoms regarding their health. Here we use some data analysis techniques to predict whether a person is suffering from a disease or not.

1.3 Objective

The core objective of this project is to implement data mining techniques Decision Tree , Random Forest , Naïve Bayes , Support vector machine, KNN and Stacking Classifier that classifies the disease as per the input of the user and to develop web interface platform for the prediction of the disease .This project aims to provide a web platform to predict the occurrences of disease on the basis of various symptoms. The user can select various symptoms and can find the diseases with their probabilistic figures.

CHAPTER 2 : BACKGROUND STUDY

2.1 SUMMARY OF RESEARCH PAPERS

PAPER 1

Title of paper: Disease Prediction Using Machine Learning

Author : Akash C. Jamgade, Prof. S. D. Zade

Summary : In this paper we not only use structured data, but also the text data of the patient based on the proposed k-mean algorithm. To find it out it combined both data, and the accuracy rate can be reached upto 95%. None of the existing system and work is focused on using both the data types in the field of medical big data analytics. It proposes a K-mean clustering algorithm for both structured and unstructured data.

PAPER 2

Title of paper: Disease Prediction System

Author : Sarthak Khurana, Atishay Jain, Shikhar Kataria ,Kunal Bhasin , Sunny Arora

Summary : This paper is using python as a platform to run the machine learning algorithms. Firstly it decided the problem and then getting the dataset to work on .Then it visualized the data with the help of scatter plot or any different plot and saw it on an excel file by doing this it checks the redundancy in the data i.e. outliers, missing values etc. Then it treats the data by replacing the missing values, as python is a case sensitive programming language it transformed all the letters into capital. Creating dummy variables to sort the data into mutually exclusive categories also the no of dummy variables should be less than the no of categories of a qualitative variable. Also it didn't did the mistake of replacing the missing values with mean of that variable but by doing so we can miss very important variations in the data.

PAPER 3

Title of paper: Multi Disease Prediction using Data Mining Techniques

Author : K.Gomathi , Dr. D. Shanmuga Priyaa

Summary: This paper analyzes data mining techniques which can be used for predicting different types of diseases. This paper reviewed the research papers which mainly concentrate on predicting heart disease, Diabetes and Breast cancer etc. In this study two different data mining classification techniques was used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier.

CHAPTER 3 : REQUIREMENT ANALYSIS

3.1 Software and Hardware Requirements

This project uses Python Version 3 environment. For this project to be executed the following hardware and software requirements should meet:

1. Python (v3 or more) IDE/Anaconda (jupyter)
2. 8 GB RAM or more
3. Intel i3 Core processor or more
4. Visual Studio Code

3.2 Functional Requirements

★ The project must provide a detail inside into comparison between different method of classification of techniques that are employed to determine best out of available techniques.

★ Project will have the ability to distinguish between a better algorithm that can be computed for better Disease prediction by employing certain techniques.

★ Project will outcome some prediction percentage and that is used as a base to distinguish better algorithm among all.

3.3 Non-functional Requirements

★ The project shall provide an accurate percentage out of the algorithm employed.

★ The code that is used compute percentage should be accurate enough to give the correct measurements outcome.

★ Data size should be kept to its best suitable for outcome analysis.

★ Project should be error free and should be able to complete in the stipulated time of project.

CHAPTER 4 : DETAILED DESIGN

4.1 METHODOLOGY

Disease Prediction is implemented using Decision tree , Random forest , Naïve Bayes , Support vector machine, KNN and Stacking Classifier.

4.2 Data Set

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The symptoms of the disease are collected from different health related websites.

4.3 Algorithm Used

4.3.1 Decision Tree

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.
In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

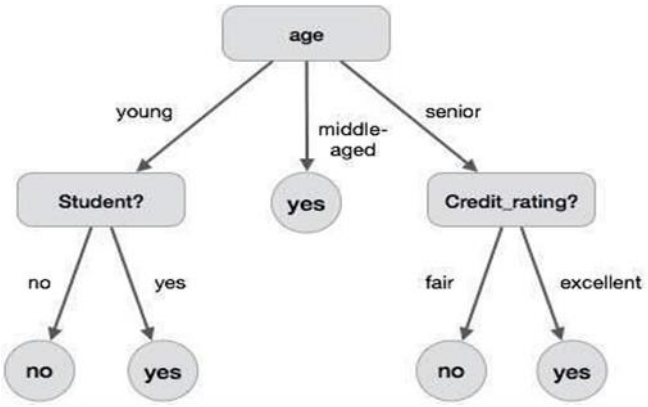


Figure 4.1 Decision tree

4.3.2 Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:
A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.
Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

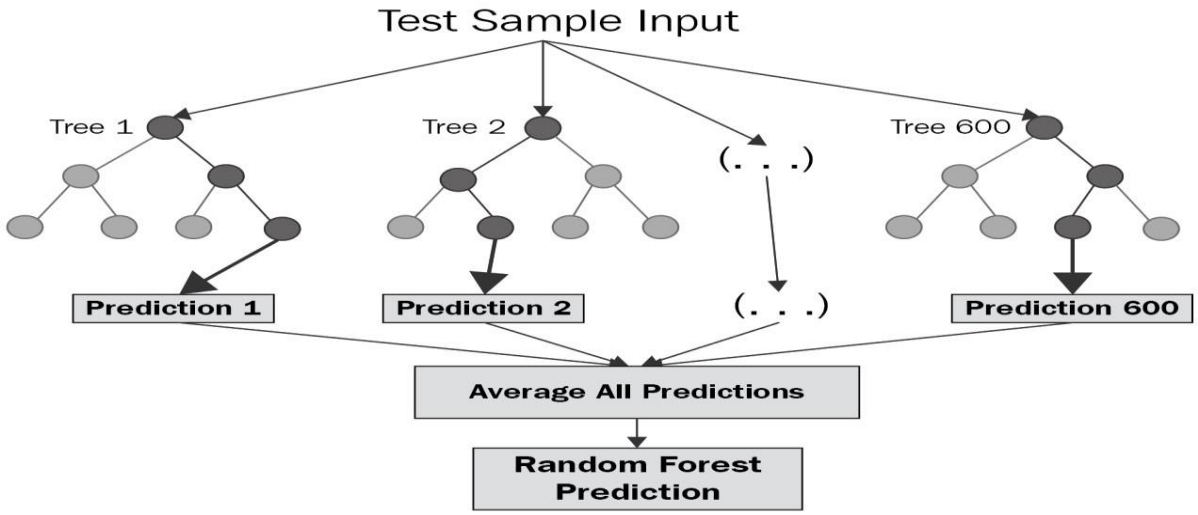


Figure 4.2 Random Forest

4.3.3 Naïve Bayes

The Naive Bayesian classifier is based on Bayes theorem with independence assumptions between predictors. A Naive Bayesian model is simple to build, with no difficult iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

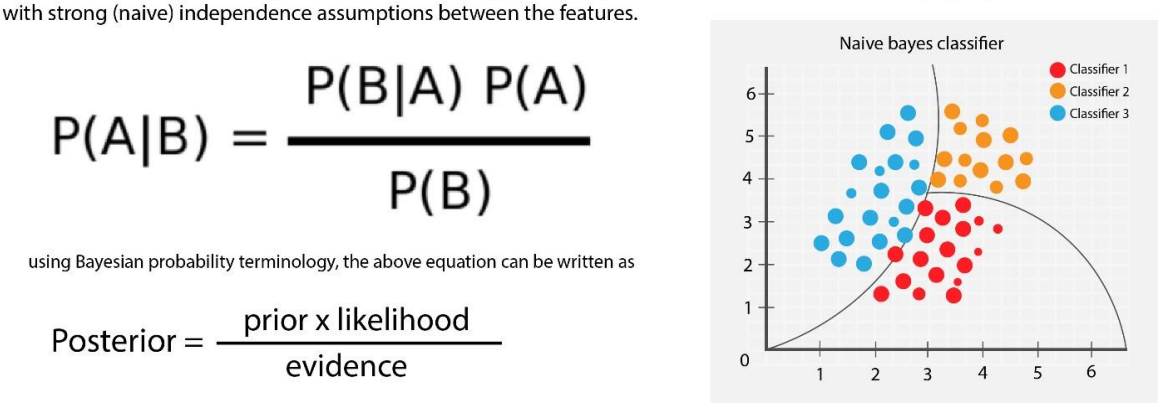


Figure 4.3 Naïve Bayes

4.3.4 KNN

KNN is a basic lazy and nonparametric classifier. KNN is preferred when every one of the features are persistent. KNN is likewise called as case-based reasoning and has been utilized in numerous applications like statistical estimation, pattern recognition. Classification is distinguishing the closest neighbour to decide the class of an unknown sample. KNN is favoured over other classification algorithms due to its high merging velocity and straightforwardness. KNN characterization has two phases:

a) Find the k number of examples in the dataset that is nearest to instance S

b) These k number of examples at that point vote to decide the class of instance S

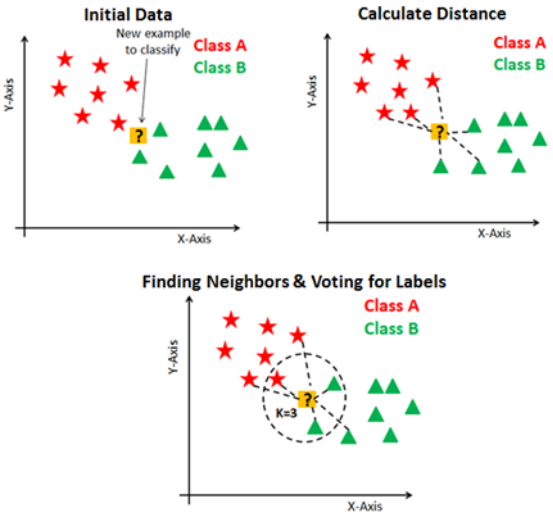


Figure 4.4 KNN

4.4.4 Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

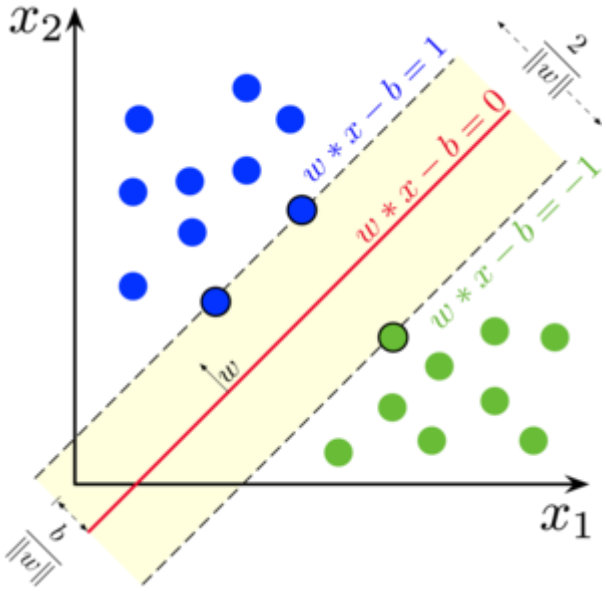


Figure 4.5 Support vector Machine

4.4.5 Stacking Classifier

Stacking is a method where a single training dataset is given to multiple models and trained. The training set is further divided using k-fold validation and the resultant model is formed. Here each model indicates a different algorithm used. In this we have used the combination SVM ,random forest and logstic regression. In this this have made an input table with the help of random forest and SVM algorithim then we have used this table as an input for logistic regression to predictc the answer**.**

4.4 System Design

4.4.1 State Diagram

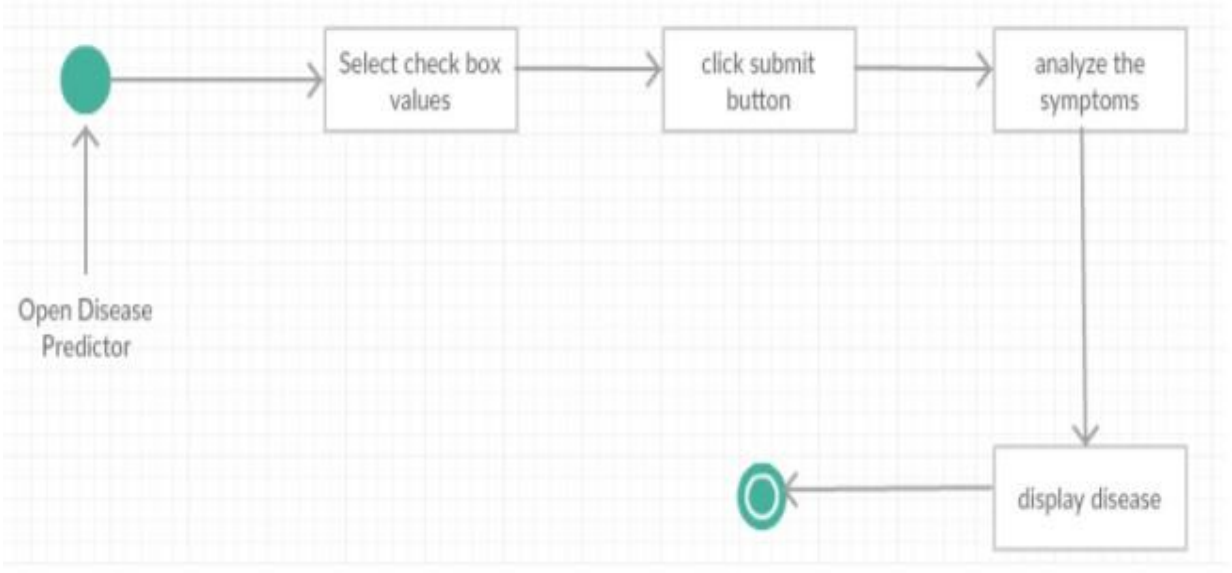


Figure 4.7 State Diagram

4.4.2 Class Diagram

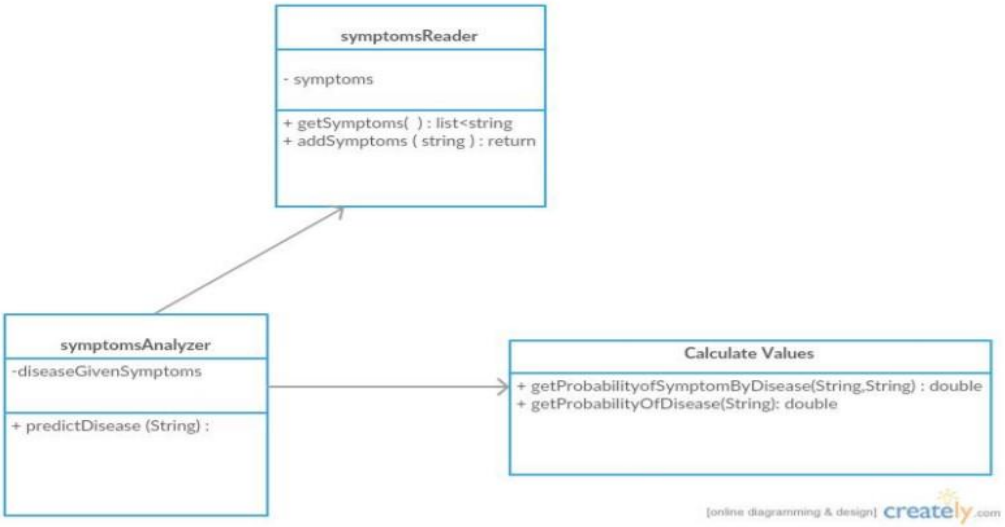


Figure 4.8 Class Diagram

4.4.3 Sequential Diagram

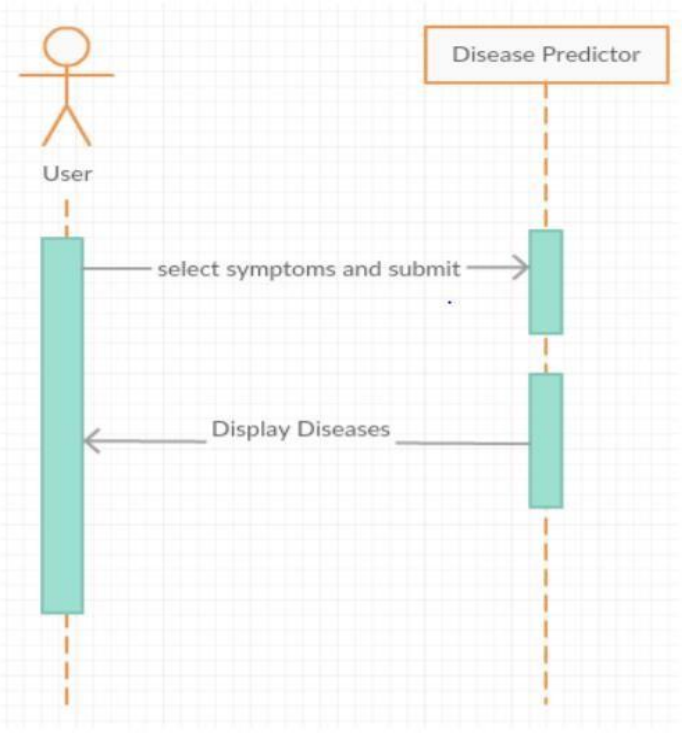


Figure 4.9 Sequential Diagram

4.5 Project Pipeline

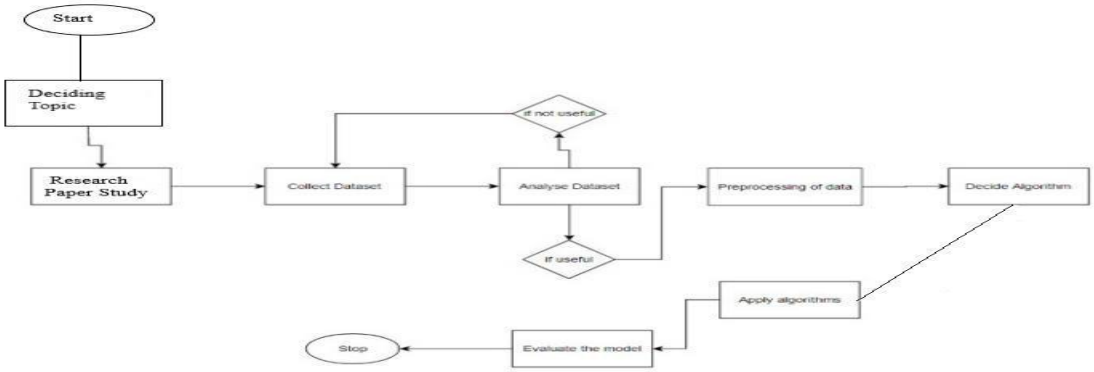


Figure 4.10 Project Pipeline

CHAPTER 5 : IMPLEMENTATION

5.1 TOOLS USED :

PYTHON : KNN ,NB ,decision tree ,Random Forest, SVM and Stacking Classifier implementation is carried out in python. This language supports both object oriented and functional programming and is compatible with machine learning methods. It enables to express concepts in fewer lines of codes as compared to languages such as C++ or Java. It has many well documented and easy to use libraries that can help accomplish lot of different programming tasks. We used the following libraries :

SCIKIT-LEARN : It features various classification, regression and clustering algorithms including SVMs, random forests and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

MATPLOTLIB : It is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. We can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

SEBORN: Seaborn is a python data visualization library based on matplotlib.it provide a high level interface for drawing attractive and information statistical graphics.

PANDAS: In computer programming, pandas is a software library written for python programming language for data manipulating and analysis. In particular, it offer data structure and operation for manipulating numerical tables and time series.

HTML/CSS: CSS is used to define the cell padding of table cells, the style, thickness, and colour of a table's border, and the padding around images or other objects. CSS gives Web developers more exact control over how Web pages will look than HTML does.

FLASK: Flask is used to connect our Machine Learning Code to Front End.

5.2 Data Extraction

This is the first and the most important step this includes data collection. In this step the real time heart disease data is collected. This is also known as Data collection step. We collected this dataset from UCI.

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads
0	1	1	1	0	0	0	0	0	0	0	...	0
1	0	1	1	0	0	0	0	0	0	0	...	0
2	1	0	1	0	0	0	0	0	0	0	...	0
3	1	1	0	0	0	0	0	0	0	0	...	0
4	1	1	1	0	0	0	0	0	0	0	...	0
5	0	1	1	0	0	0	0	0	0	0	...	0
6	1	0	1	0	0	0	0	0	0	0	...	0
7	1	1	0	0	0	0	0	0	0	0	...	0
8	1	1	1	0	0	0	0	0	0	0	...	0
9	1	1	1	0	0	0	0	0	0	0	...	0

Figure 5.1 Data set

5.3 Data Analysis

In this step , we have analysed our dataset by graphic representation of data. To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable.

5.4 Algorithm Implementation

5.4.1 Decision Tree

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer (DAC)approach. Tree pruning is performed to remove the irrelevant samples on D.

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij}$$

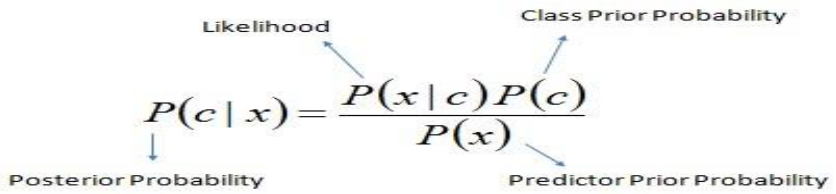
5.4.2 Random Forest

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, X={x1, x2, x3, ...,xn} with responses Y={x1, x2, x3, ..., xn} which repeats the bagging from b= 1to B. The unseen samples x0is made by averaging the predictions from every individual trees on x0:

$$j = \frac{1}{B} \sum_{b=1}^{B} fb\,(x')$$

5.4.3 Naïve Bayes

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing some n features (independent variables), it assigns to this instance probabilities for each of K possible outcomes or classes. Using Bayes Theorem , the conditional probability can be decomposed as



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

5.4.4 KNN

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data
   3.1 Calculate the distance between the query example and the current example from the data.
   3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order).
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

5.4.5 Support Vector Machine

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

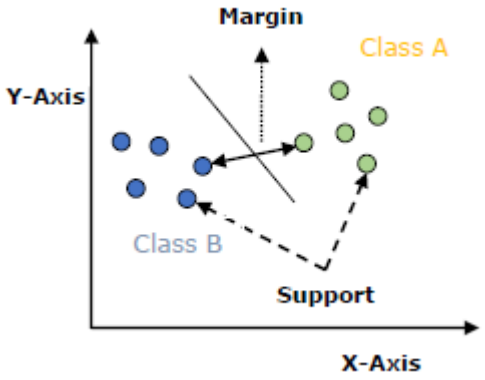


Figure 5.2 Support Vector Machine

5.4.6 Stacking classifer

Stacking is a method where a single training dataset is given to multiple models and trained. The training set is further divided using k-fold validation and the resultant model is formed. Here each model indicates a different algorithm used. In this we have used the combination svm,random forest and logstic regression. In this this have made an input table with the help of random forest and SVM algorithim then we have used this table as an input for logistic regression to predictc the answer.

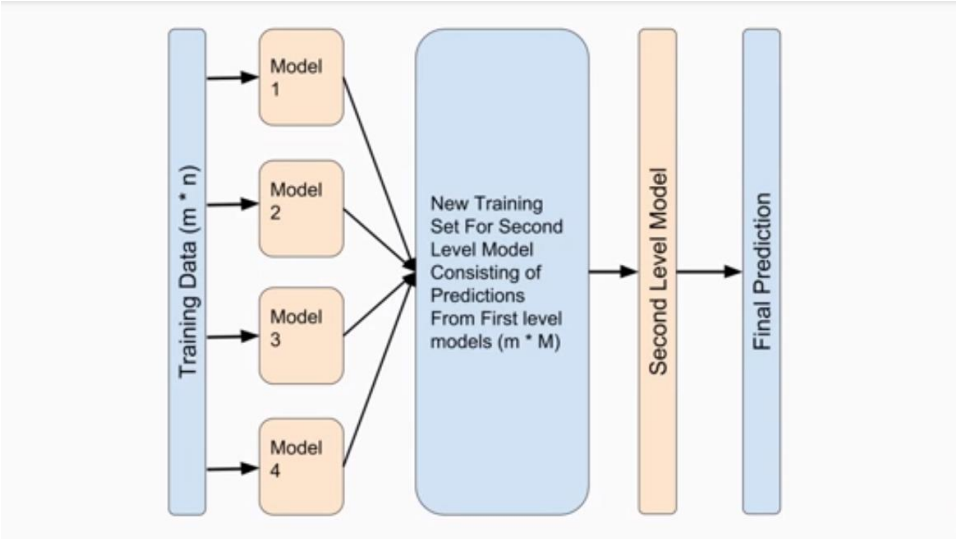


Figure 5.3 Stacking Classifier

5.5 Web Platform Development

Through the Webpage User can Login into their account and mark their symptoms on the page to get the desired result.

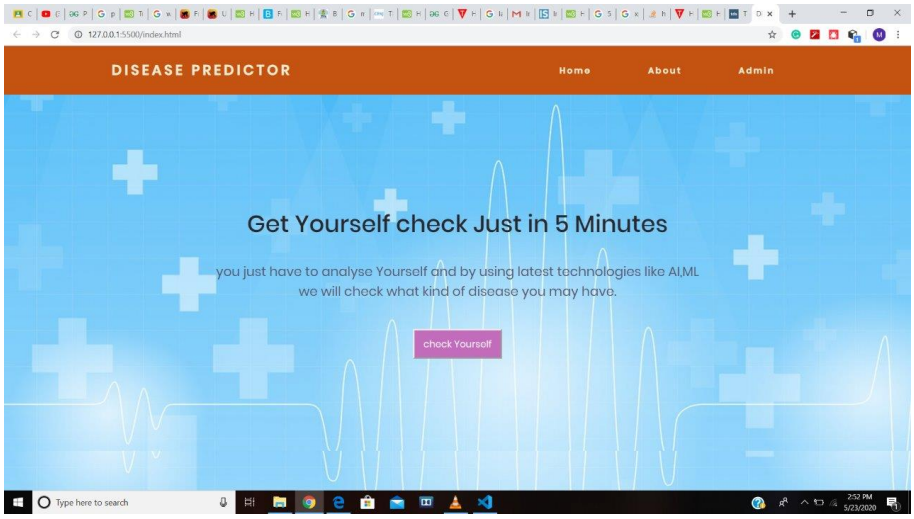


Figure 5.4 The Home Page

Figure 5.4 Shows The homepage of the website that the patients are required to visit. By clicking on the Check yourself button the user can move forward to the login page.

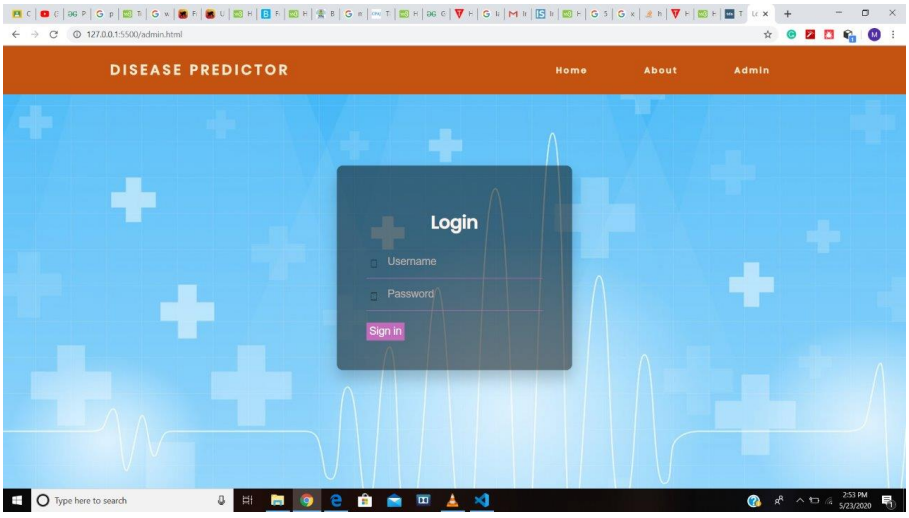


Figure 5.5 The Login Page

Figure 5.5 Shows the Login Page. The user can enter the personal details to login and go to the Symptoms page.

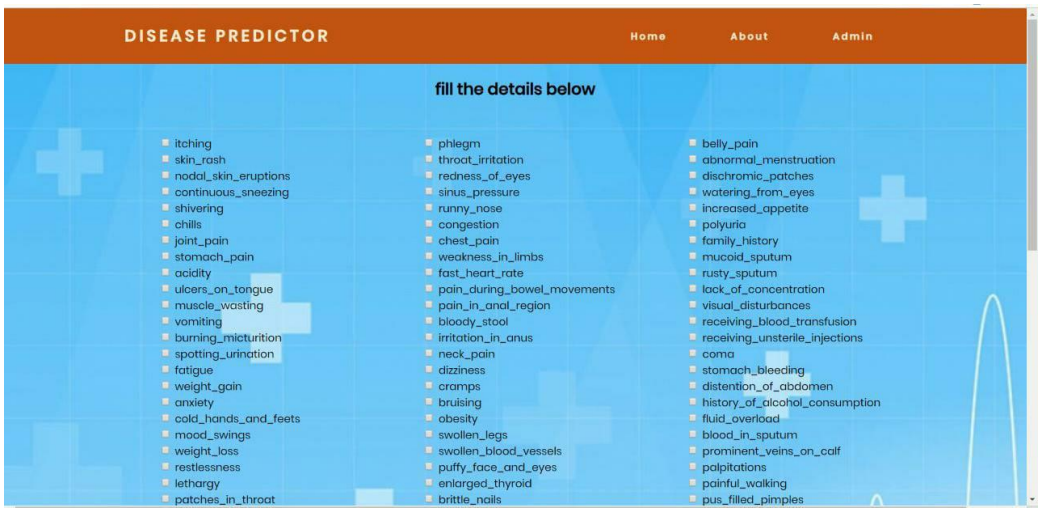


Figure 5.6 The Disease Symptoms Checkbox Page

Figure 5.6 shows The disease symptoms Checkbox page where the user can enter the details to get the output of the disease.

CHAPTER 6 : RESULT AND TESTING

The prediction accuracy of the various Data mining techniques are:

1.Decision Tree: 81.7%

2.Random Forest: 98.5%

3.KNN: 92%

4.SVM: 96.5%

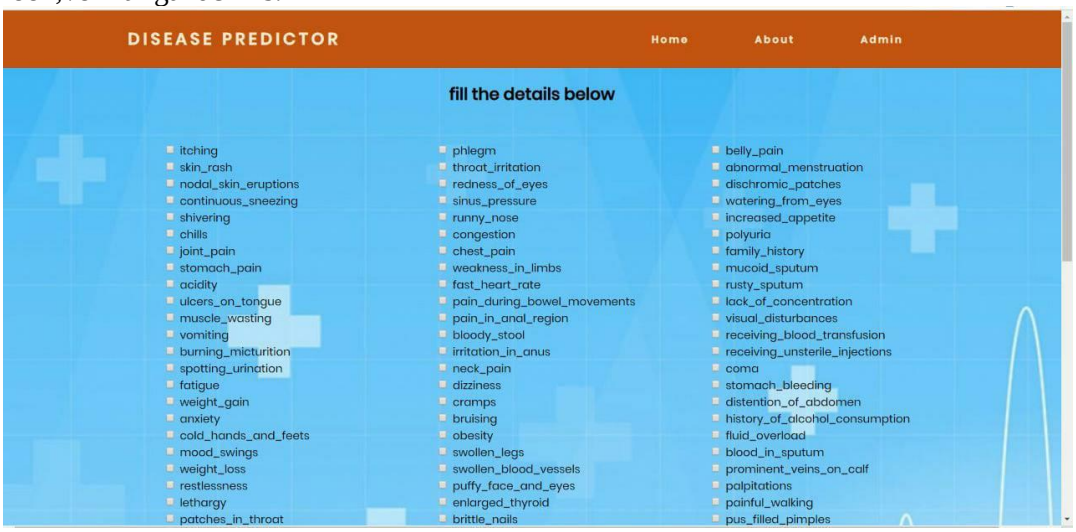
5.Naive Bayes: 97.5%

6.Stacking Classifier: 98.5%

As we can see that Random Forest and Stacking Classifier have the best accuracy for prediction but here we are using Random Forest Technique because of space complexity issue.

The test case designed for the project is discussed below:

Test case 1: Submit the symptoms from the list.
The symptoms are : abdominal pain, nausea ,high fever , fatigue, constipation, belly pain , toxic look,vomitingandchills.



Expected result: According to the symptoms submitted the expected disease is Typhoid.

Test Case 2: Submit the symptoms from the list.
The symptoms are : Visual disturbances, depression, irritability , stiff neck, blurred vision, Acidity ,excessive hunger and headache.

Expected result: According to the symptoms submitted the expected disease is Migraine.

CHAPTER 7 : CONCLUSION

The ultimate goal is to facilitate coordinated and well-informed health care systems capable of ensuring maximum patient satisfaction. In developing nations, predictive analytics are the next big idea in medicine –the next evolution in statistics – and roles will change as a result. Patients can get to become higher knowing and can get to assume a lot of responsibility for his or her own care, if they are to make use of the information derived. Physician roles can probably modification to a lot of an advisor than head, who will advise, warn and help individual patients. Perhaps time with individual patients can increase and physicians will another time have the time to create positive and lasting relationships with their patients. In this study five different data mining classification techniques were used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications

CHAPTER 8 : GANTT CHART

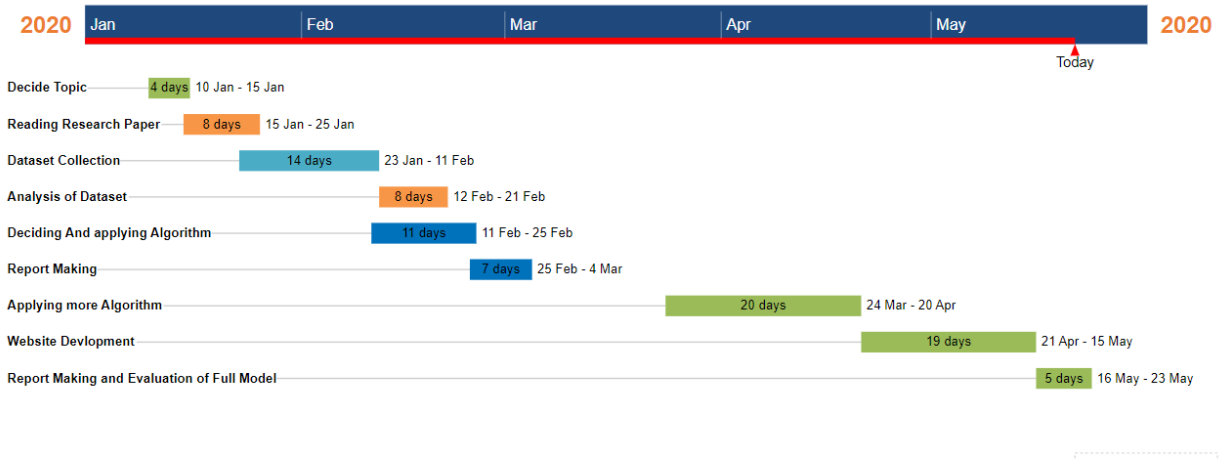


Figure 8.1 Gantt Chart

CHAPTER 9 : REFERENCES

[1] Akash C. Jamgade, Prof. S. D. Zade ,"Disease prediction using Machine learning", International Research Journal of Engineering and Technology (IRJET) ,Volume 6,Issue 5,May 2019

[2] Sarthak Khurana, Atishay Jain, Shikhar Kataria,Kunal Bhasin, Sunny Arora, "Disease prediction system", International Research Journal of Engineering and Technology (IRJET) , Volume 6,Issue 5,May 2019

[3] K.Gomathi*, Dr. D. Shanmuga Priyaa," Multi Disease Prediction using Data Mining Techniques", http://www.publishingindia.com , Sep 2017