DISEASE PREDICTOR

A REPORT

submitted by

P.JYOSHNA
17MIS1153

in partial fulfilment for the award

of

M. Tech. Software Engineering (Integrated)

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

September 2020

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

DECLARATION

I hereby declare that the project entitled **"Disease Predictor"** submitted by me to the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai 600127 in partial fulfilment of the requirements for the award of the degree of **Master of Technology - Software Engineering (Integrated)** is a record of bonafide work carried out by me**.** I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or university.

Signature

P.Jyoshna
17MIS1153

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

CERTIFICATE

The project report entitled "**Disease Predictor**" is prepared and submitted by **P.Jyoshna** **(Register No: 17MIS1153)**. It has been found satisfactory in terms of scope, quality and presentation as partial fulfilment of the requirements for the award of the degree of **Master of Technology – Software Engineering (Integrated)** in Vellore Institute of Technology, Chennai, India.

Examined by:

Examiner I **Examiner II**

ACKNOWLEDGEMENT

I wish to thank those who were involved in the successful completion of my internship at Triculin Technologies Pvt. Ltd., starting from the Senior Information Security Analyst , for giving me the opportunity and freedom to learn as per my interests; the head of the project team for being a constant support and guidance; the project lead for my internship Vanga Mohana Aditya Reddy, for providing me with the necessary resources; and the entire staff of the company for their support and positivity which made my internship a worthwhile experience.

I would also like to thank my parents, for being my motivation to take up this internship; and last, but not the least, the faculty and management at Vellore Institute of Technology (VIT), Chennai, for providing me with such an avenue to help realize how interesting it is to work in today's industry.

It is my proud privilege to express my profound gratitude to the Dean of SCOPE, Dr. Jagadeesh Kannan R and Head of the Department Dr. Asnath Victy Phamila and the Associate Dean of SCOPE,Dr.Geetha S. for providing me this valuable opportunity to have industrial exposure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CARE- Collaborative Assessment and Recommendation Engine

ICD- International Classification Of Disease.

NB- Naïve Bayes

HTML- HyperText Markup Language

CSS- Cascading Style Sheets

ABSTRACT

"Disease Prediction" system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output

Disease Prediction is done by implementing the Naïve Bayes Classifier. Naïve Bayes Classifier calculates the probability of the disease. Therefore, average prediction accuracy probability 60% is obtained.

Keywords: Predictive Modeling, Naïve Bayes Classifier

CHAPTER 1 INTRODUCTION

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 Heart 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.

As the use of internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to internet than hospitals and doctors. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24 hours.

Problem Statement

There are many tools related to disease prediction. But particularly heart related diseases have been analyzed and risk level is generated. But generally there are no such tools that are used for prediction of general diseases. So Disease Predictor helps for the prediction of the general diseases.

Objective

General Objective

-To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

Specific Objective

-To develop web interface platform for the prediction of the disease.

Scope and Limitations

Scope

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.

Limitations

The limitations of this project are:

a. Disease Predictor does not recommend medications of the disease.

b. Past history of the disease has not been considered

1.4 Outline of Document

Preliminary Section	• Title Page • Abstract • Table of Contents • List of figures and Tables
Introduction Section	• Background of Research • Statements of Problems • Objectives

Requirement Analysis and Feasibility Analysis	• Literature Review • Requirement Analysis • Feasibility Analysis
System Design	• Methodology • System Design • Implementation and Testing
Maintainace and Support	• Maintenance • Support
Conclusion and Recommendation	• Conclusion • Recommendation

CHAPTER 2 REQUIREMENT ANALYSIS AND FEASIBILITY ANALYSIS

Literature Review

K.M. Al-Aidaroos, A.A. Bakar and Z. Othman have conducted the research for the best medical diagnosis mining technique. For this authors compared Naïve Baeyes with five other classifiers i.e. Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN) and a simple rule-based algorithm (ZeroR). For this, 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007) were selected for evaluating the performance of all algorithms. In the experiment it was found that NB outperforms the other algorithms in 8 out of 15 data sets so it was concluded that the predictive accuracy results in Naïve Baeyes is better than other techniques.

Table 1- Predictive Accuracy of Bayes and other Technique

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echoeardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	84.97	78.45	83.18	78.21	81.81	54.76
Postooerative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	8\15	5\15	0\15	2\15	1\15	1\15

(Al-Aidaroos, Bakar, & Othman, 2012)

Darcy A. Davis, Nitesh V. Chawla, NicholasBlumm, Nicholas Christakis, Albert-Laszlo Barabasi have found that global treatment of chronic disease is neither time or cost efficient. So the authors conducted this research to predict future disease risk. For this CARE was used (which relies only on a patient's medical history using ICD- 9-CM codes in order to predict future diseases risks). CARE combines collaborative filtering methods with clustering to predict each patient's greatest disease risks based on their own medical history and that of similar patients. Authors have also described an Iterative version, ICARE, which incorporates ensemble concepts for improved performance. These novel systems require no specialized information and provide predictions for medical conditions of all kinds in a single run. The impressive future disease coverage of ICARE represents more accurate early warnings for thousands of diseases, some even years in advance. Applied to full potential, the CARE framework can be used explore a broader disease

histories, suggest previously unconsidered concerns, and facilitating discussion about early testing and prevention.

(A.Davis, V.Chawla, Blumm, Christakis, & Barbasi, 2008)

JyotiSoni, Ujma Ansari, Dipesh Sharma and SunitaSoni have done this research research paper into provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering is not performing well.

(JyotiSoni, Ansari, Sharma, & Soni, 2011)

Shadab Adam Pattekari and AsmaParveen have conducted a research using Naïve Bayes Algorithm to predict the heart diseases where user provides the data which is compared with trained set of values. So from this research, patients were able to provide their basic information which is compared with the data and the heart disease is predicted.

(Adam & Parveen, 2012)

M.A.NisharaBanu, B Gomathy used medical data mining techniques like association rule mining, classification, clustering I to analyze the different kinds of heart based problems. Decision tree is made to illustrate every possible outcome of a decision. Different rules are made to get the best outcome. In this research age , sex, smoking, overweight, alcohol intake, blood sugar, hear rate, blood pressure are the parameters used for making the decisions. Risk level for different parameters are stored with their id's ranging (1-8). ID lesser than of 1 of weight contains the normal level of prediction and higher ID other than 1 comprise the higher risk levels .K-means clustering technique is used to study the pattern in the dataset. The algorithm clusters informations into k groups. Each point in the dataset is assigned to the closed cluster. Each cluster center is recomputed as the average of the points in that cluster.

(NisharBanu, MA; Gomathy, B;, 2013)

Requirement Analysis

Functional requirements

a. Predict disease with the given symptoms.
b. Compare the given symptoms with the input datasets

Non-functional requirements

a. Display the list of symptoms where user can select the symptoms.

b. Naïve Bayes Classifier is used to classify the data sets.

Feasibility Analysis

Technical feasibility

The project is technically feasible as it can be built using the existing available technologies. It is a web based applications that uses Grails Framework. The technology required by Disease Predictor is available and hence it is technically feasible.

Economic feasibility

The project is economically feasible as the cost of the project is involved only in the hosting of the project. As the data samples increases, which consume more time and processing power. In that case better processor might be needed.

Operational feasibility

The project is operationally feasible as the user having basic knowledge about computer and Internet. Disease Predictor is based on client-server architecture where client is users and server is the machine where datasets are stored.

CHAPTER 3 SYSTEM DESIGN

Methodology

Disease Prediction has been already implemented using different techniques like Neural Network, decision tree and Naïve Byes algorithm. Particularly heart related disease is mostly analyzed. From the analysis it was found that Naïve Bayes is more accurate than other techniques. So, Disease Predictor also uses Naïve Bayes for the prediction of different diseases.

Data collection

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.

Algorithm implemented

The algorithm implemented in this project is Naïve Bayes Classifier.

Naïve Bayes classifier depends on Bayes Theorem

Equation 1:

$$P(Y|X1,\ldots\ldots\ldots\ldots,X_{n)} = \frac{P(Y)P(X1,\ldots\ldots,X_n|Y)}{P(X1,\ldots\ldots X_n)}$$

Where,

Y is the class variable

$X_1, X_2, \ldots\ldots\ldots, X_n$ are the dependent features

From equation 1 we get equation 2 as:

$$P(\text{Disease}|\text{symptom}_1, \text{symptom}_2, \ldots\ldots, \text{symptom}_n) = \frac{P(\text{Disease})P(\text{symptom}_1, \ldots\ldots, \text{symptom}_n|\text{Disease})}{P(\text{symptom}_1, \text{symptom}_2, \ldots\ldots\ldots\ldots \text{Symptom}_n)}$$

Using the naive independence assumption :

$P(\text{symptom}_1, \ldots\ldots, \text{symptom}_n|\text{Disease}) = P(\text{Symptom}_i|\text{Disease})$

Where i= 1, 2,……,n

Equation 3:

$$P(\text{Disease}|\text{symptom}_1, \text{symptom}_2, \ldots\ldots, \text{symptom}_n) = \frac{P(\text{Disease})P(\text{Symptom}_i|\text{Disease})}{P(\text{symptom}_1, \text{symptom}_2, \ldots\ldots\ldots\ldots \text{Symptom}_n)}$$

So the relation becomes:

Equation 4:

$$P(\text{Disease}|\text{symptom}_1, \text{symptom}_2, \ldots\ldots, \text{symptom}_n) = \frac{P(\text{Disease})\prod_{i=1}^{n} P(\text{Symptom}_i|\text{Disease})}{P(\text{symptom}_1, \text{symptom}_2, \ldots\ldots\ldots\ldots \text{Symptom}_n)}$$

Since $P(\text{symptom}_1, \text{symptom}_2, \ldots\ldots\ldots\ldots \text{Symptom}_n)$ is constant, we can use the following classification rule:

$$P(\text{Disease}|\text{symptom}_1, \text{symptom}_2, \ldots\ldots, \text{symptom}_n) = P(\text{Disease}) \prod_{i=1}^{n} P(\text{Symptom}_i|\text{Disease})$$

$P(\text{Disease}|\text{symptom}_1, \text{symptom}_2, \ldots\ldots, \text{symptom}_n) \propto$

$P(\text{Disease}) \prod_{i=1}^{n} P(\text{symptomi}|\text{Disease})$

$$\hat{Y} = \text{ARG MAX } P(\text{Disease}) \prod_{i=1}^{n} P(\text{Symptom}_i|\text{Disease})$$

The value $P(\text{Symptom}_i|\text{Disease})$ of can be calculated by using multinomial Naïve Bayes which is given by:

$$P(symptom_i|Disease) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Where:

N_{yi}= Frequency of same disease in the dataset

N_y= Total symptoms of the particular disease

n= total symptoms in the dataset

α=1, known as Laplace Smoothing

The value of P(Disease) can be calculated by using Laplace Law of Succession which is given by:

$$\text{P(Disease)} = \frac{N(Disease) + 1}{N + 2}$$

Where,

N (Disease) = Frequency of the same disease in the dataset

N= Total disease in the dataset

System Design

Class Diagram

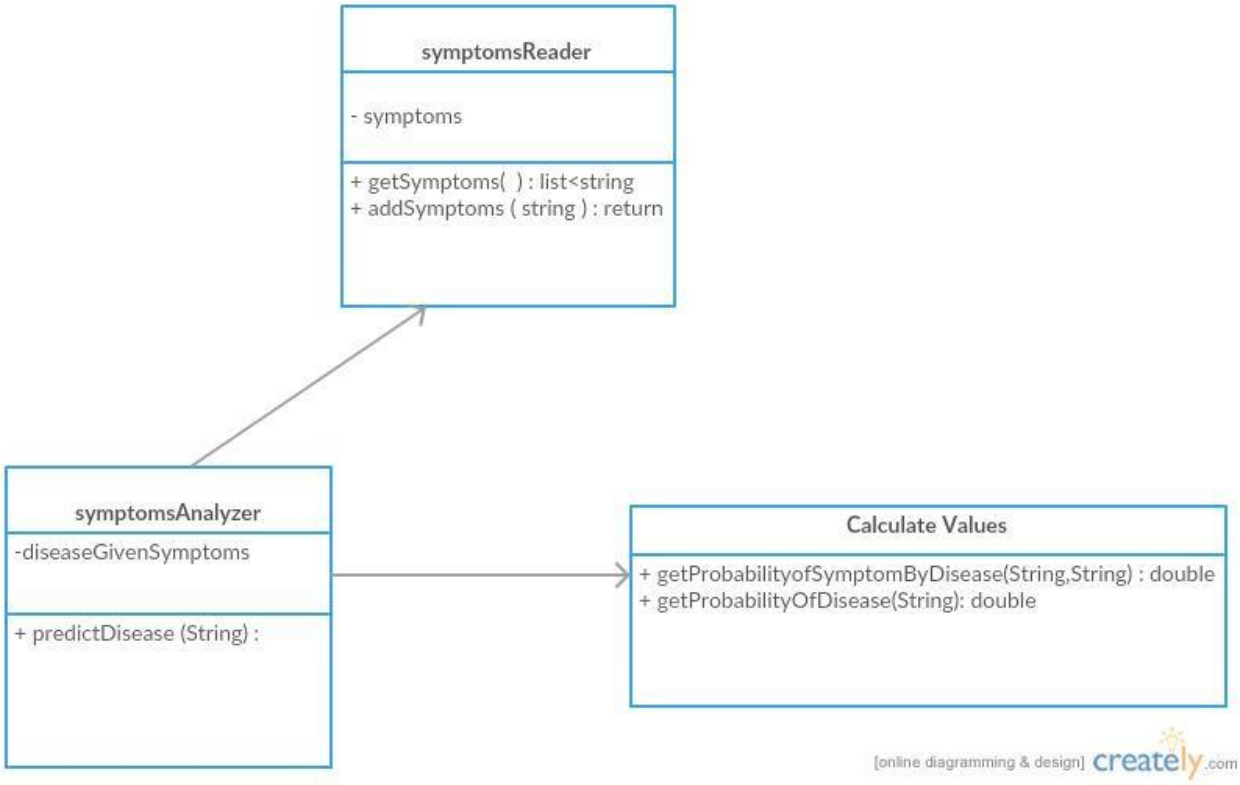


Figure 1 - Class Diagram

It explain the classes used in the Disease Predictor. There are three classes used in total,

Symptoms Reader: Reads the user input and creates the list of symptoms

Symptoms Analyzer: According to symptoms parameter displays the subjective result.

Calculate Values: Calculates the probabilistic model of the diseases.

State diagram

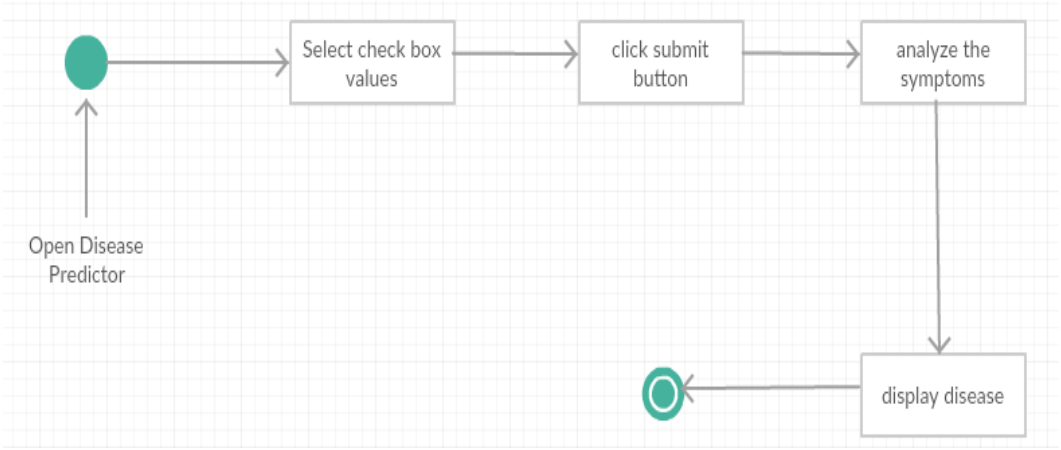


Figure 2- State Diagram

It explains different state of the system. First the user opens Disease Predictor. The user selects the symptoms. When finished selecting symptoms the user submits the symptoms. Disease Predictor analyzes the symptoms and displays the result.

Sequence diagram

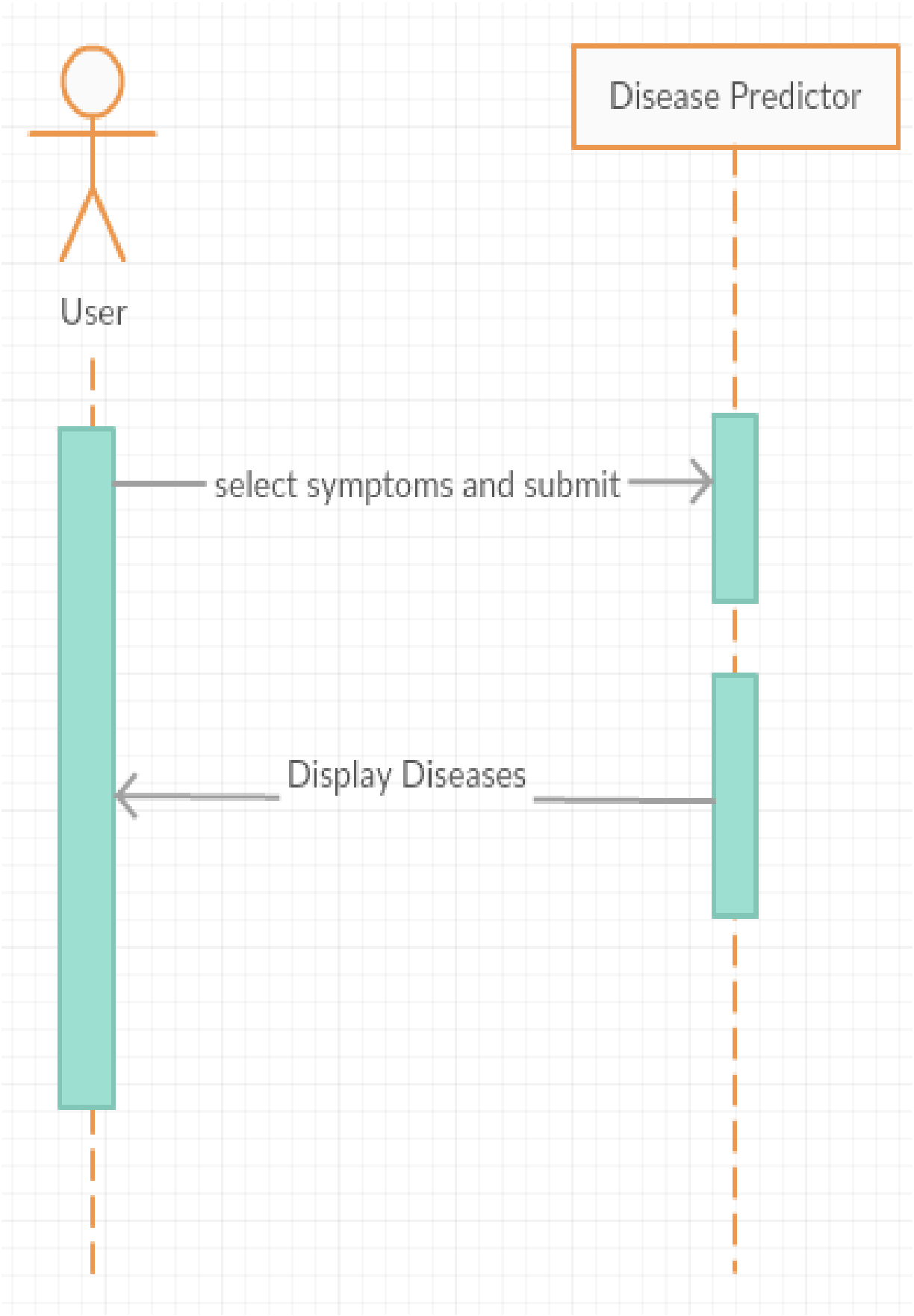


Figure 3- Sequence Diagram

It explains the sequence of the Disease Predictor. Initially system shows the symptoms to be selected. The user selects the symptoms and submits to the system .The Disease Predictor predicts and display the result

CHAPTER 4 IMPLEMENTATION AND TESTING

Implementation

Disease Predictor is the ability to predict the disease that has been provided to the system. For disease prediction, we need to implement the naïve Byes Classifier.

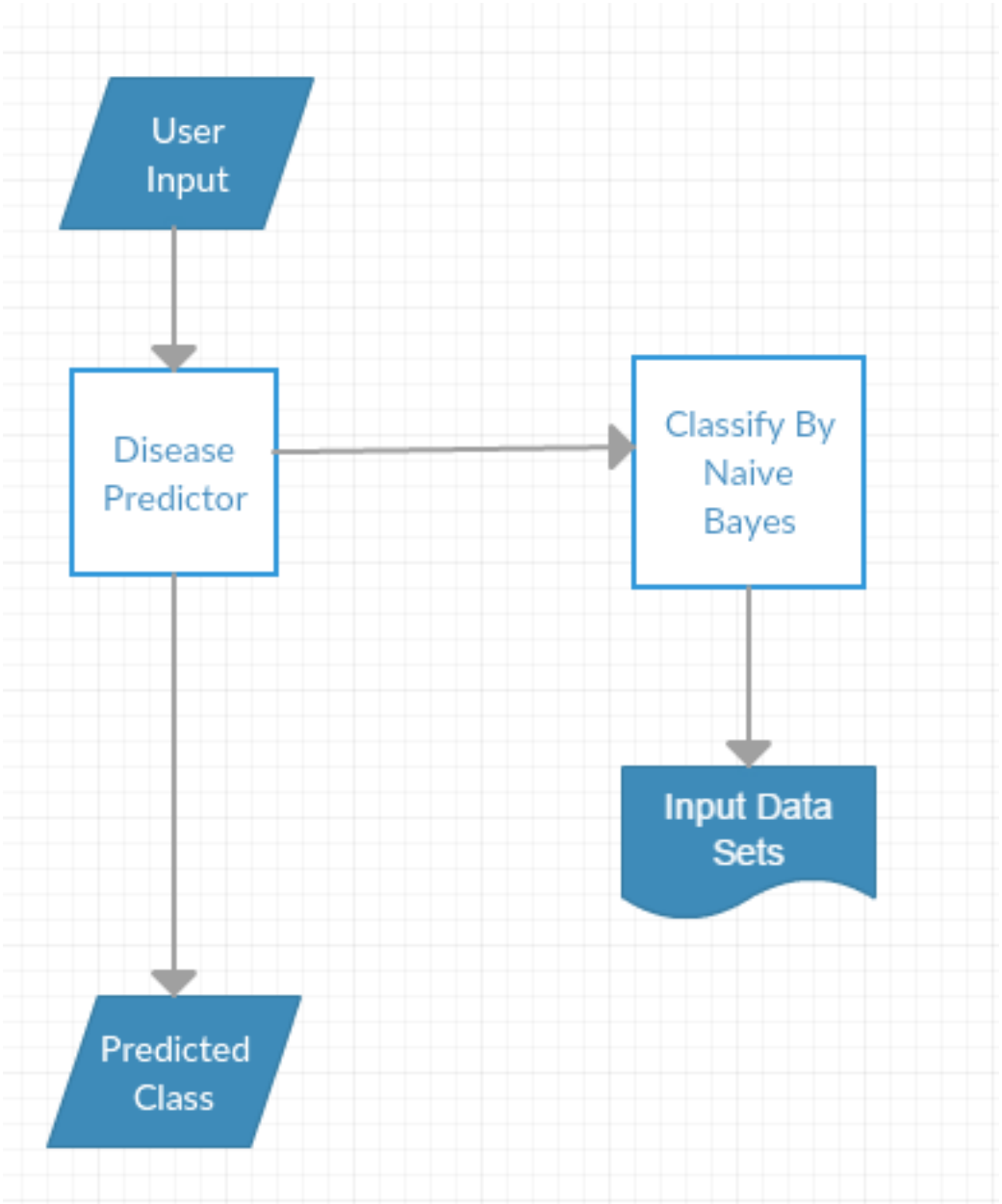


Figure 4- Workflow

As shown in the figure the input data sets are classified using Naïve Bayes classifier. The sample input data sets is shown below

Table 2- Sample Data Sets

Symptoms	Disease
Runny nose ,Sore throat ,Cough ,Congestion, body aches, headache ,Sneezing , fever	Common cold
Fever ,profuse sweating ,headache ,nausea ,vomiting ,diarrhea ,anemia ,muscle pain ,convulsions ,coma bloody stools ,shaking chill	Malaria
poor appetite ,abdominal pain ,headaches ,generalized aches and pains ,fever ,lethargy ,intestinal bleeding or perforation ,diarrhea , constipation	Typhoid

Naïve Bayes classifier uses the following rule to classify the datasets:

$$\hat{Y} = \text{ARG MAX } P(\text{Disease}) \prod_{i=1}^{n} P(\text{symptom}_i | \text{Disease})$$

User gives input to the system. The input consists of symptoms. The user marks the symptoms due to which the user is feeling unwell.

1. ☐ Fever
2. ☐ Cough
3. ☐ Vomiting

The "Disease Predictor" system predicts the disease according to the input data sets and calculates the probability of the disease.

The sample output is given as:

Table 3- Sample Output

Disease Name	Probability
Typhoid	0.5%
Malaria	0.3%
Flu	0.333%

Tools Used

1. HTML is used to display content in the browser.
2. CSS is used to properly align the HTML content.
3. Grails framework is used for developing the application.
4. Creately is used for constructing figures.

Description

The major classes in the application are:

SymptomsReader

This class is the run first when the user wants for disease prediction
Input: User selects the symptoms from the list.
Output: The selected symptoms are put in the list

SymptomsAnalyzer

Input: Takes the user input i.e. symptoms.
Output: Predicts the disease

CalculateValues

Here the actual mathematical computation takes place.

4.3 Testing

The test case designed for the project is discussed below:

Test Case- I: Submit the symptoms from the list	
Precondition: The application is open.	
Assumptions: The symptoms for the disease are available	
Test steps:	1. Select the checkbox from the list 2. Select submit
Expected Result: The symptoms selected should be submitted and further analyzed to calculate the probability of the disease.	

CHAPTER 5 MAINTENANCE AND SUPPORT

Corrective Maintenance

In case of any bugs left in the system, the bugs and issues will be fixed for smooth running of the application. The accuracy of the system can be further improved with other algorithms if needed.

Adaptive Maintenance

The features in the application can be added such as history of the disease can be kept in the log. The available list of symptoms can also be added for covering more number of diseases.

CHAPTER 6 CONCLUSION AND RECOMMENDATION

Conclusion

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. Average prediction accuracy probability of 55% is obtained. Disease Predictor was successfully implemented using grails framework.

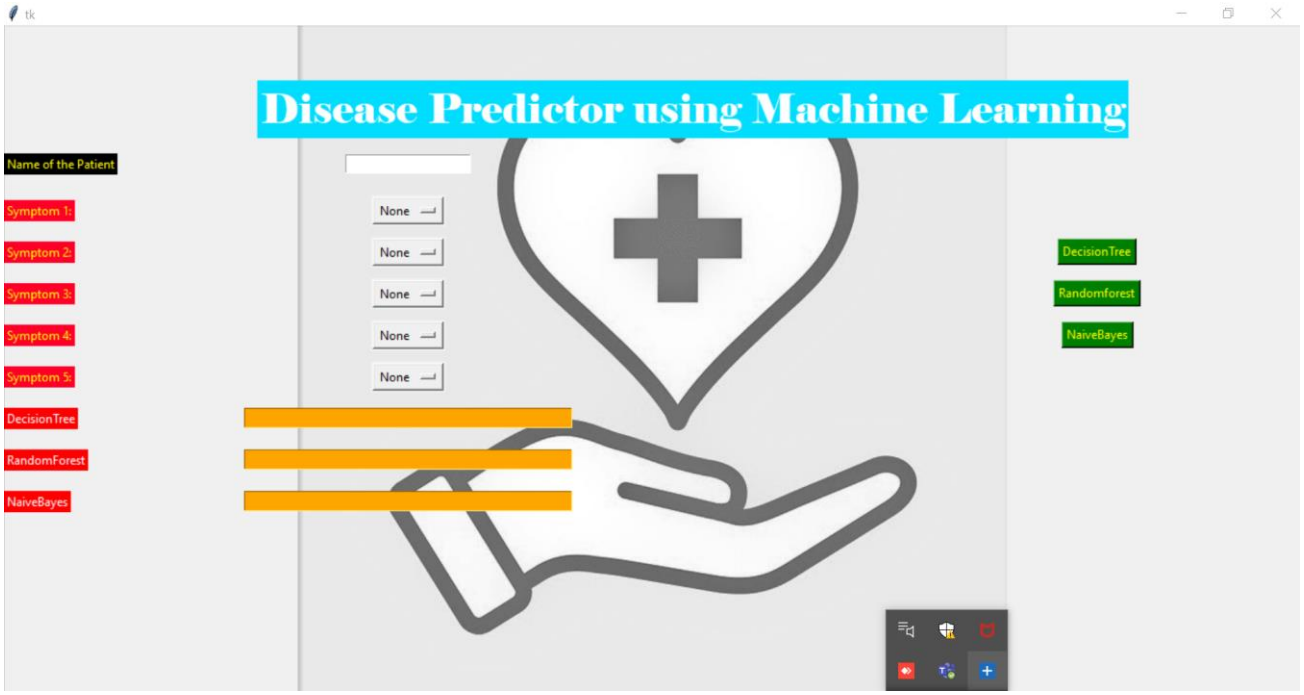
Recommendations

This project has not implemented recommendation of medications to the user. So, medication recommendation can be implemented in the project. History about the disease for a user can be kept as a log and recommendation can be implemented for medications.

‘

APPENDIX

Landing Page of Disease Predictor



Output showing the probability of the disease