A PROJECT REPORT

On

"Pharma Guide"

Submitted to
KIIT Deemed to be University

In Partial Fulfilment of
The Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE ENGINEERING

BY

Rahul Sahu 1705157
Rishika Sinha 1705159

UNDER THE GUIDANCE OF
Dr. Suresh Ch. Sathapathy

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
May 2020

Github Link: https://github.com/therealrahulsahu/MinorProject

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"Pharma Guide"

Submitted by

Rahul Sahu 1705157

Rishika Sinha 1705159

Is a record of genuine work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2019-2020, under our guidance.

Date: 27 May 2020

Dr. Suresh Ch. Sathapathy
(Project Guide)

Acknowledgment

We are profoundly grateful to Prof. Dr. Suresh C. Sathapathy for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

Rahul Sahu
Rishika Sinha

Abstract

Now-a-days, people face various diseases due to the environmental condition and their living habits. So the prediction of disease at earlier stage becomes important task. **"Pharma Guide"** 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 Decision Tree Classifier. Decision Tree Classifier makes a binary tree on the basis of common symptoms. Therefore, prediction accuracy depends on No. of symptoms confirmed by user.

Keywords:

Decision Tree Classifier, HTTP Request, Predictive modelling, flask, API (Application Program Interface), Web App

Contents

List of Figures

Introduction

1. Purpose

The purpose of this project is to build a virtual doctor to predict the disease base on the symptoms. Top suitable drugs based on review and ratings are displayed to the user. **The major purpose is to check possible disease or crosscheck doctor's diagnosis. This application can also use to diagnose some minor and major disease.**

2. Project Scope

The purpose of the project is to create a convenient and easy-to-use application for patients. The prediction model is based on a data-set which has disease with the symptoms. Above all, we hope to provide a comfortable user experience and to extend the scope of this project across the world.

3. Intend Audience

Our target audience is the people who are showing symptoms of illnesses that can be cured at home by taking either medicine or precautions. We let people know about their illness and whether they need to see a doctor. The patients have the convenience of checking their symptoms from anywhere like home, workplace etc.

Literature Survey

Lifestyle diseases are the diseases that are linked with the way people live their lives. Diseases that impact on our lifestyle are Heart disease, Stroke, Obesity and Type II Diabetes. Two of the lifestyle diseases are explained below:

A. Heart Disease

Heart is the most vital part of the human body as the life is dependent on efficient working of heart. If functioning of heart is not proper then it will influence the other body parts. Heart disease is the major cause of causalities in the different countries including India. Heart disease kills one person in every 34 seconds in the United States. A heart disease is caused due to narrowing or blocking of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol.

There are number of factors which increase the risk of Heart disease..

Family history
Smoking
Cholesterol
Poor Diet
High blood pressure
Obesity
Physical inactivity
Hypertension

1) Symptoms: The symptoms include tightness or pain in the chest, back or arms, neck, as well as fatigue, lightheadedness, abnormal heartbeat and anxiety. Women are more likely to have atypical symptoms than men.

a) Pain area: area between shoulders blades, arm, chest, jaw, left arm or upper abdomen.

b) Pain types: can be crushed, like a clenched first in the chest, radiating from the chest, sudden in the chest, or mild.

c) Pain circumstances: may occur during rest

d) Whole body: dizziness, fatigue, light-headedness, clammy skin, cold sweat, or sweating.

e) Gastrointestinal: heartburn, indigestion, nausea or vomiting.

f) Chest: discomfort, fullness or tightness.

g) Neck: discomfort or tightness.

h) Arm: discomfort or tightness.

2) Types of Heart diseases: Heart disease includes all types of disease affecting different components of the heart. Heart means ‘cardio’.

Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are: Coronary Artery Disease is the most common type of heart disease. In coronary artery disease, the arteries carry blood to the heart muscle which contain cholesterol and fat are lined with plague. Angina is a pain that occurs when your heart is not getting enough oxygen and

nutrients. It is the medical term for chest pain that occurs due to insufficient supply of blood to the heart. Myocarditis it is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is uncommon disease with few symptoms like join pain, leg swelling or fever that cannot be directly related to the heart.

3) Heart Failure: Heart failure happens when the heart isn't pumping enough blood to meet your body's needs.

4) Arrhythmia: Sometimes the heart's electrical system does not function normally. It may skip beats or sometimes the heart's electrical signal does not move in the proper sequence. These abnormal rhythms are called arrhythmia.

5) Cardiomyopathy: It is the weakening of the heart muscle or a change in the structure of the muscles due to inadequate heart pumping. Hypertension, alcohol consumption, viral infections, and genetic defects are common causes of cardiomyopathy.

6) Congenital heart disease: It is also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

B. Diabetes

Diabetes is a disease in which the body could not produce insulin or sometimes could not use the produced insulin properly. Diabetes leads to gathering of glucose particle in the blood instead of going into body cell. The gathering of glucose particle in the blood invites various kinds of instabilities in the body.

The various factors that help to determine the presence of diabetes are:

Number of times pregnant.
Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
Diastolic blood pressure (mm Hg)
Triceps skin fold thickness (mm)
2-Hour serum insulin (mu U/ml)
Body mass index (weight in kg/ (height in m)^2)
Diabetes pedigree function

1) Symptoms: The symptoms of diabetes are:

Polyuria (frequent urination)
Polydipsia (increased thirst)
Polyphagia (increased hunger)
Weight loss

Numerous works has been done related to lifestyle disease diagnosis using different data mining techniques. The data-set, algorithms, methods used by the authors and the observed results along with the future work is carried out in finding out efficient methods of medical diagnosis for various lifestyle diseases. Here is a brief discussion about two lifestyle diseases i.e. heart disease and type II diabetes and the work that has been already carried out in past few years.

A. Heart disease diagnosis using classification methods

Hlaudi Daniel Masethe predicts and diagnoses heart disease by using different data mining Algorithms such as J48, REPTREE, Naïve Bayes, Bayes Net and Simple CART. The author analyzes the performance of these algorithms through evaluation criteria such as Kappa Statistics, Mean Absolute Error, Root Mean Squared, Relative Absolute Error and Root Relative Squared Error. Accuracy of J48, REPTREE, Naïve Bayes, Bayes Net and CART are 99.0741%, 99.0741%, 97.222%, 98.1481% and 99.0741% respectively.

B. Heart Attack Prediction System using Clustering

Shantakumar B. Patil applied K-mean clustering algorithm on the pre-processed data. And the recurrent patterns applicable to heart disease are mined with the MAFIA algorithm from the data extraction. The neural network is trained with the selected important patterns for effective prediction of Heart Attack on the basis of computed significant weightage.

C. Heart Disease Diagnosis Using Fuzzy Logic Approach

P.K. Anooj has proposed a weighted fuzzy rule based CDSS for the diagnosis of heart disease. It automatically obtains the knowledge from the patient clinical data. The proposed CDSS for risk of heart patients consists of two phases. First is computerized approach for generation of weighted fuzzy rules and decision tree and the second is creating a fuzzy rule based decision support system.

D. Heart Disease Prediction Using Association Rule

V.Manikandan et al. extract the item set relations by using association rule. The data classification was based on MAFIA algorithms which resulted in better accuracy. The data was evaluated using entropy based cross validation and partition techniques and the results were compared. MAFIA (Maximal Frequent Item-set Algorithm) used a dataset with 19 attributes and the goal of the research work was to have highly accurate recall metrics with higher levels of precision.

E. Heart Disease Prediction System using Hybrid System

R. Chitra et.al. Present Hybrid Intelligent techniques for the prediction of heart disease. Some Heart disease classification system was reviewed in this study and concluded with justification importance of data mining in heart disease diagnosis and classification. The classification accuracy can be improved by reduction in features.

F. Type II Diabetes Prediction Using Hybrid Model

Jayaram et al. develop of a hybrid model for classifying Pima Indian Diabetic Database (PIDD). The model consisted of two stages. In the first stage, the K-means clustering was used to identified and eliminated incorrectly classified instances. In the second stage a fine tuned classification was done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify that cascaded K-means clustering and the rules generated by cascaded C4.5 tree with categorical data is easy to interpret as compared to rules generated with C4.5 alone with continuous data. The cascaded model with categorical data obtained him classification accuracy of 93.33%.

G. Type II Diabetes Prediction Using Classification

Han et al. used data mining techniques through Rapid Miner for the classification of diabetes data analysis and diabetes prediction model. A Decision tree and ID3 algorithm were used for prediction with 72% and 80% of accuracy respectively.

H. Type II Diabetes Prediction Using Rough Sets

Breault applied rough sets on the PIMA for the first time. He first pre-processed the data and discrete it by making intervals of data. He used the equal frequency binning criteria for intervals

and then he created redacts by using Johnson reducer algorithm and classified using the batch classifier with the standard/tuned voting method (RSES). The rules were constructed for each of the 10 randomizations of the PIDD training sets from above. The tests sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracy ranged from 69.6% to 85.5% with a mean of 73.8% and 95% CI of (71.3%, 76.3%)

I. Type II Diabetes Prediction Using Clustering

Vijayalakshmi et al. developed a clustering algorithm that is used for predicting diabetes based on graph b-coloring technique. They implement and perform experiments by comparing their approach with K-NN classification and K-means clustering. The results showed that the clustering based on graph coloring is much better than other clustering approaches in terms of accuracy and purity. The proposed technique presented a real representation of clusters by dominant objects that assures the inter cluster disparity in a partitioning and used to evaluate the quality of clusters.

Software Requirement And Specifications

1. **Functional Requirements**
   a) Prediction of Disease by list symptoms.
   b) Confirmation for other symptoms of predicted disease.
2. **Non-Functional Requirements**
   a) Application should be a web application.
   b) Application should run on any browser.
   c) Application should be accessible on any device.

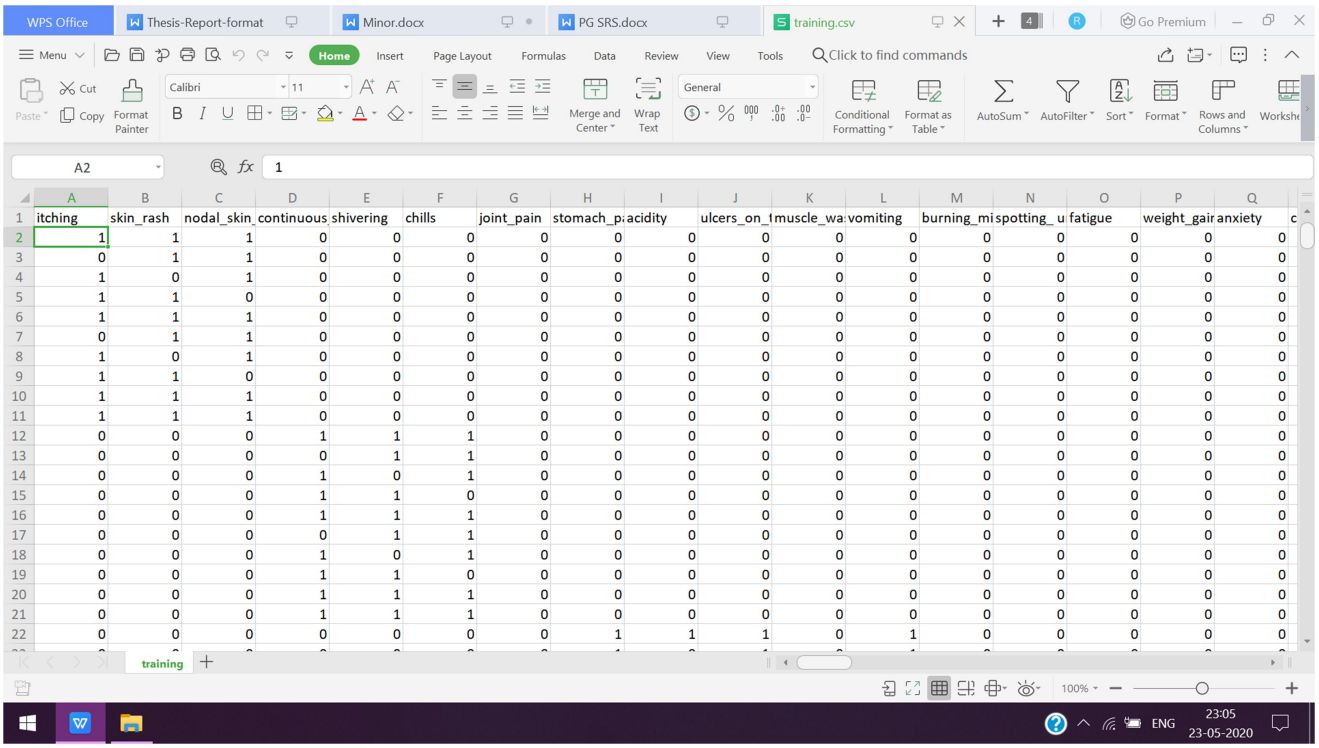
Requirement Analysis

1. **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 **flask** Framework, HTML, CSS, JavaScript and Python. The technology required by Disease Predictor is available and hence it is technically feasible.

2. **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 more processing power. In that case better processor might be needed.

3. **Operational Feasibility:** The project is operationally feasible as the user having basic knowledge about computer and Internet. Disease Predictor is based on client-server 3-tier architecture where client is users and server is the machine where data are stored.

Data-Set

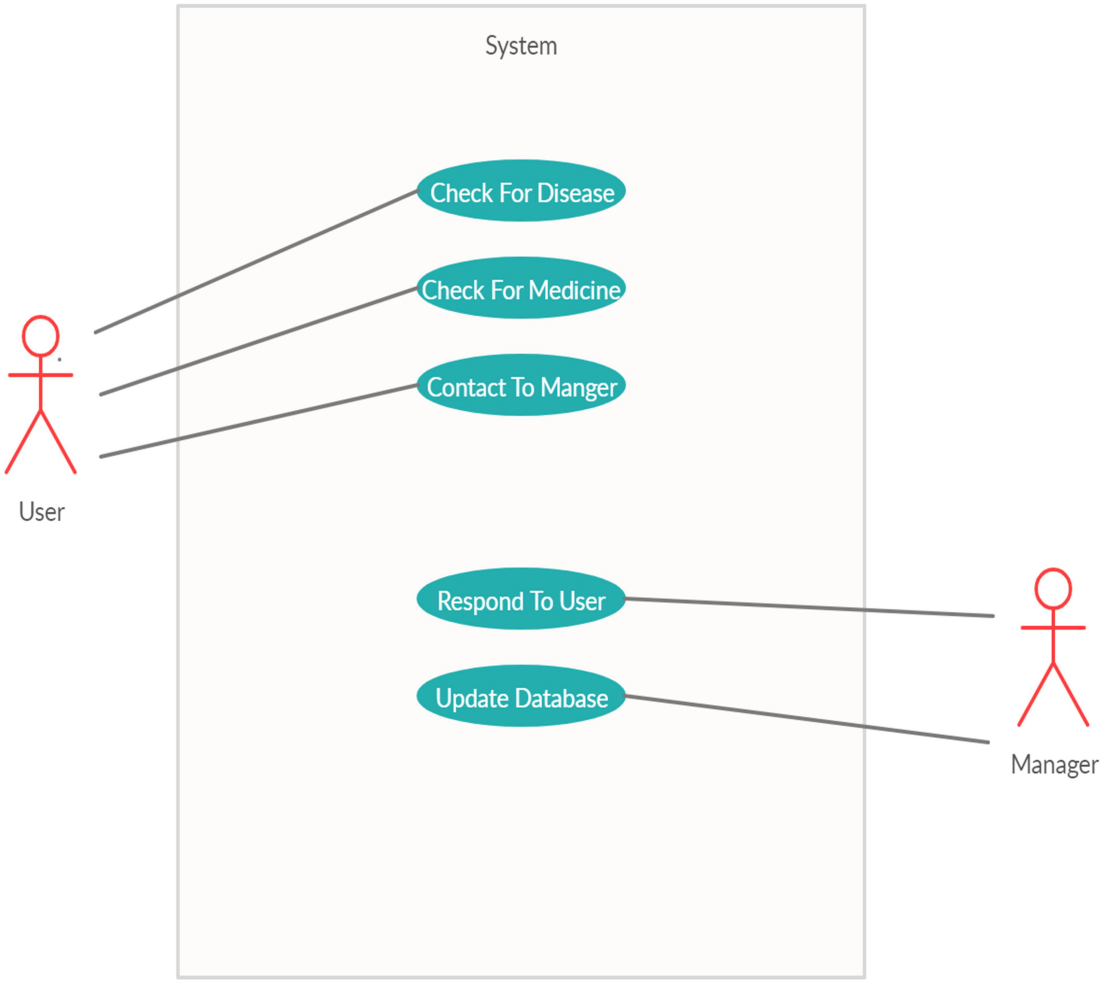


itching	skin_rash	nodal_skin	continuous	shivering	chills	joint_pain	stomach_p	acidity	ulcers_on_t	muscle_wa	vomiting	burning_mi	spotting_ u	fatigue	weight_gair	anxiety
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0

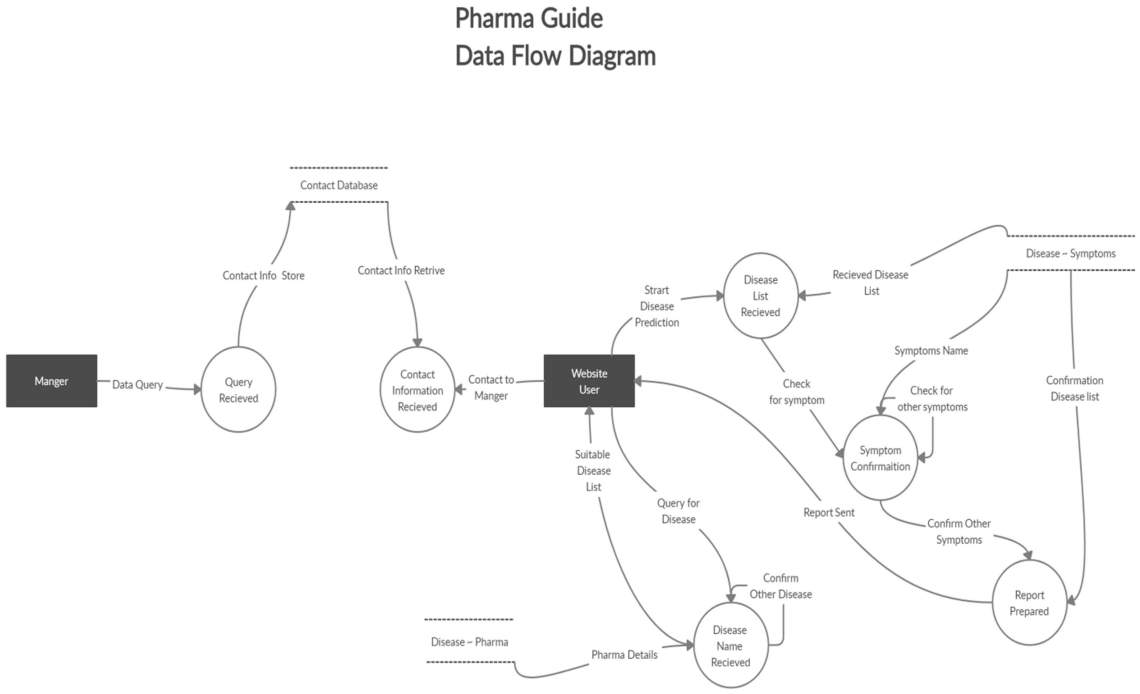
1. This data-set contains history of patient's symptoms and their diagnosed disease.
2. Patient may have multiple symptoms and last column (prognosis) contains likely disease.
3. This contains 4121 Rows and 133 columns.
4. First row contains names of symptoms (except last columns values).
5. Last column contains likely disease.
6. 4120 are data records of patients with observed symptoms and diagnosed disease.
7. Here '1' means symptom present and '0' means symptom not present.

System Design

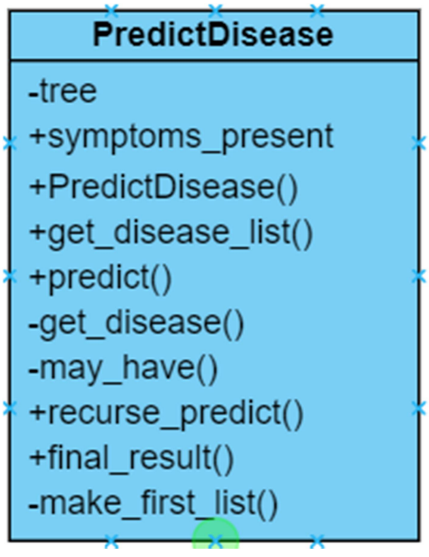
1. Use-Case Diagram



2. Data-Flow Diagram

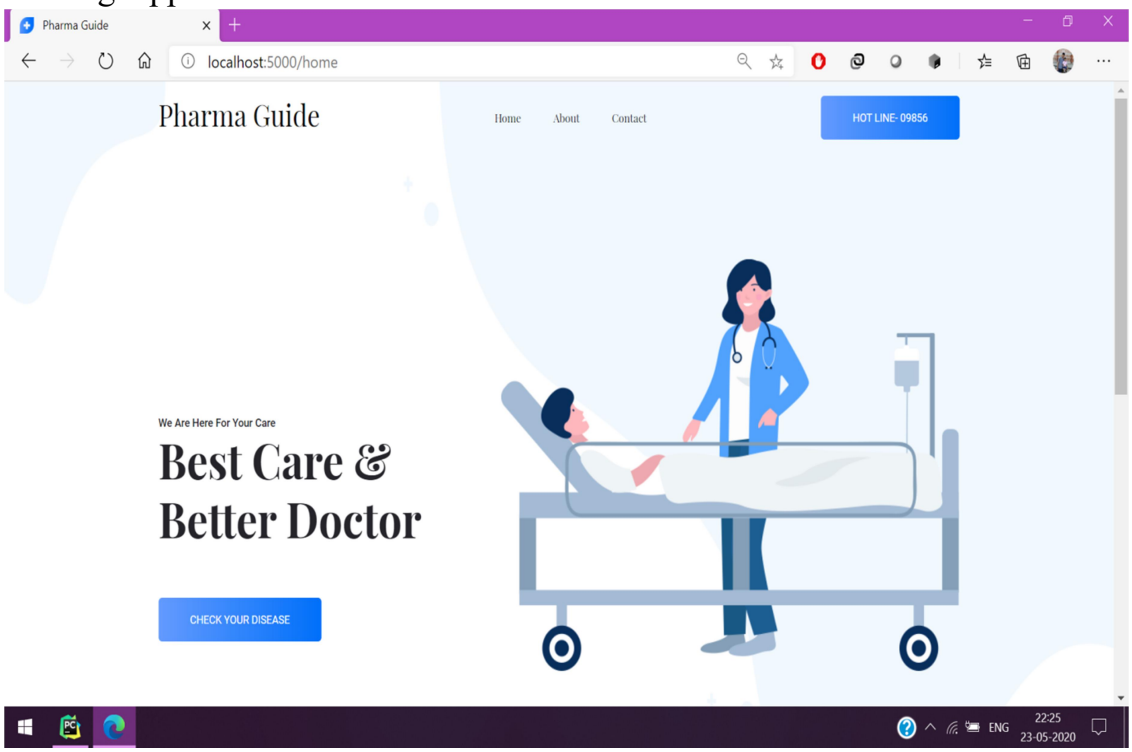


3. Class Diagram

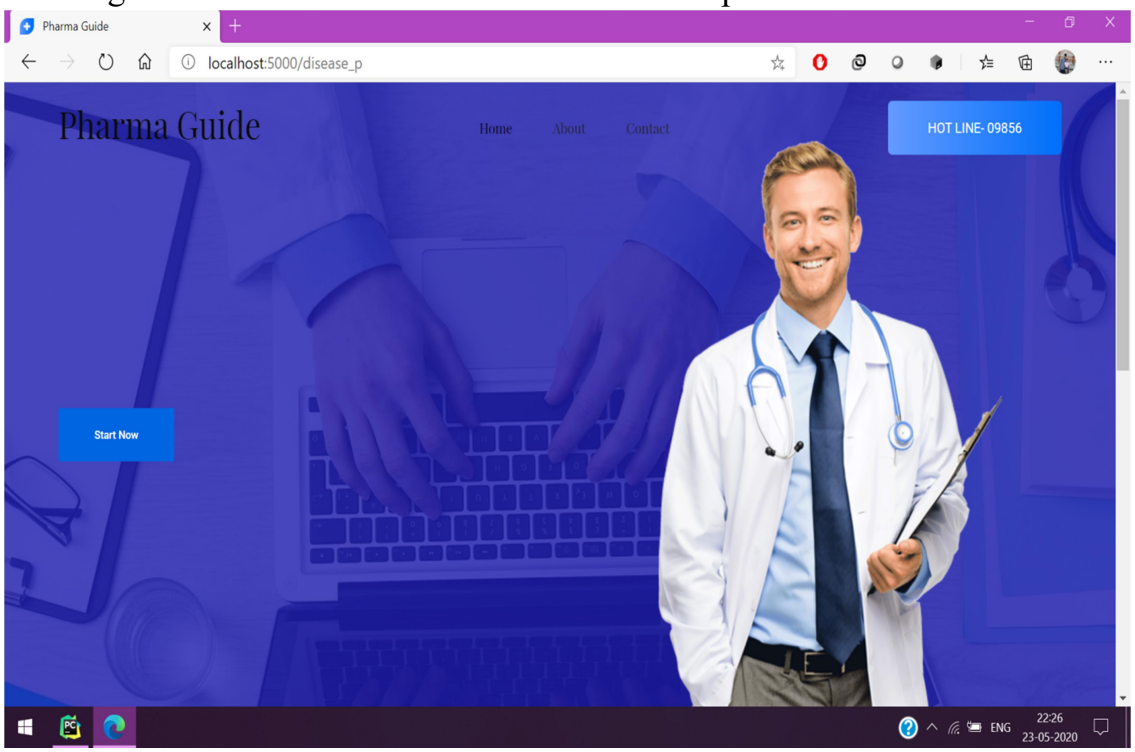


System Testing

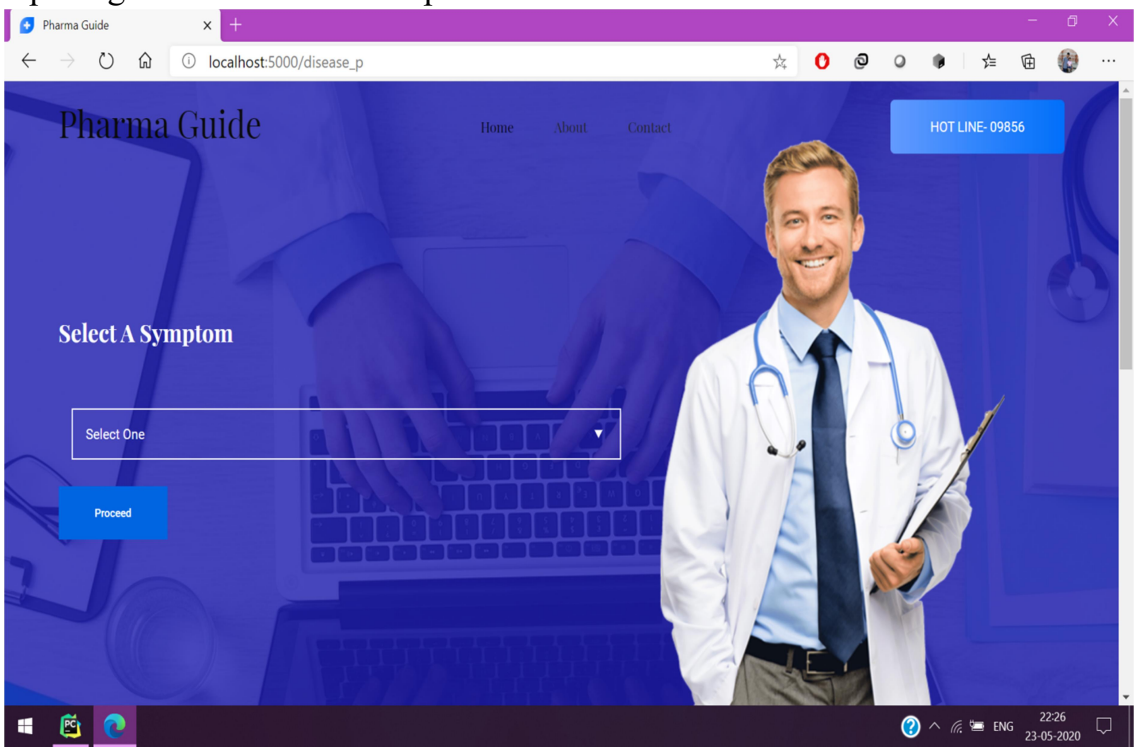
1. Starting Application



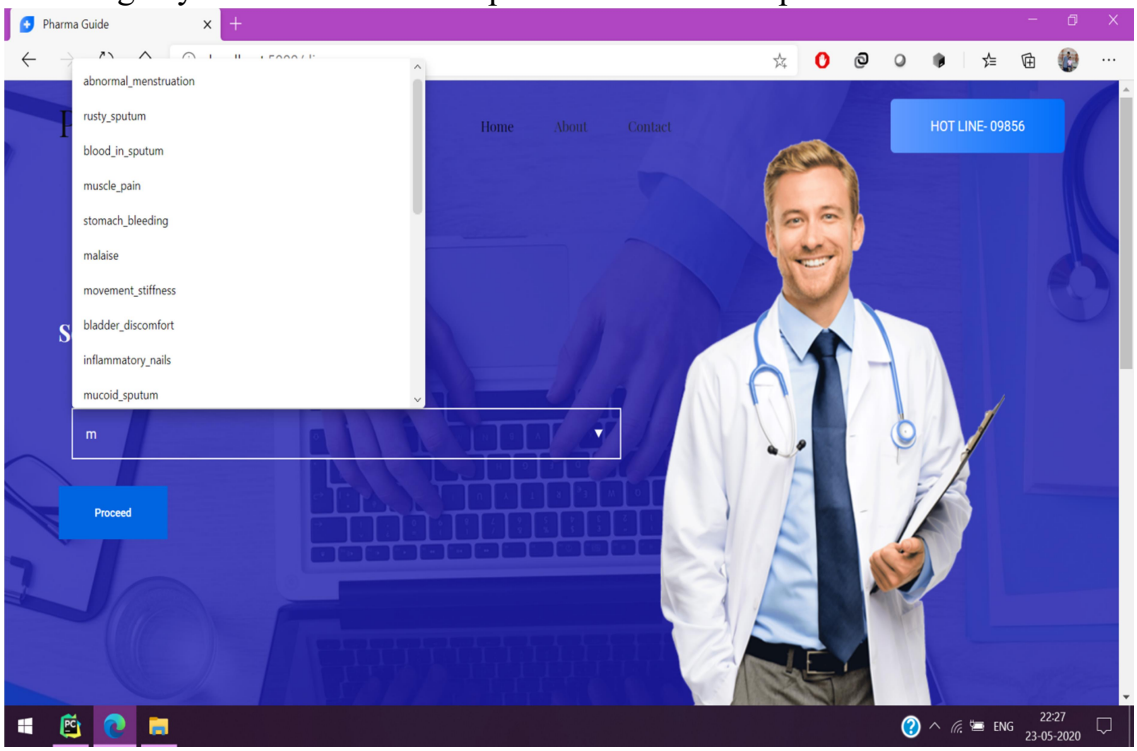
2. Clicking on “Check Your Disease” to start disease prediction.



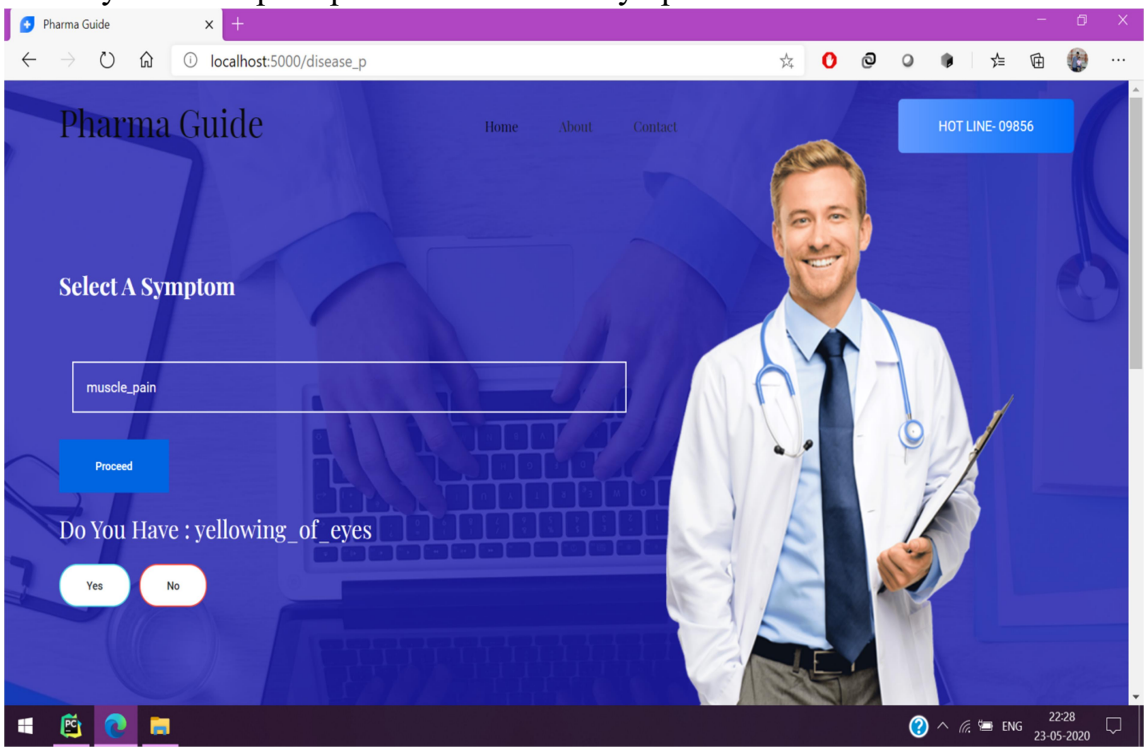
3. Opening disease selection drop down box.



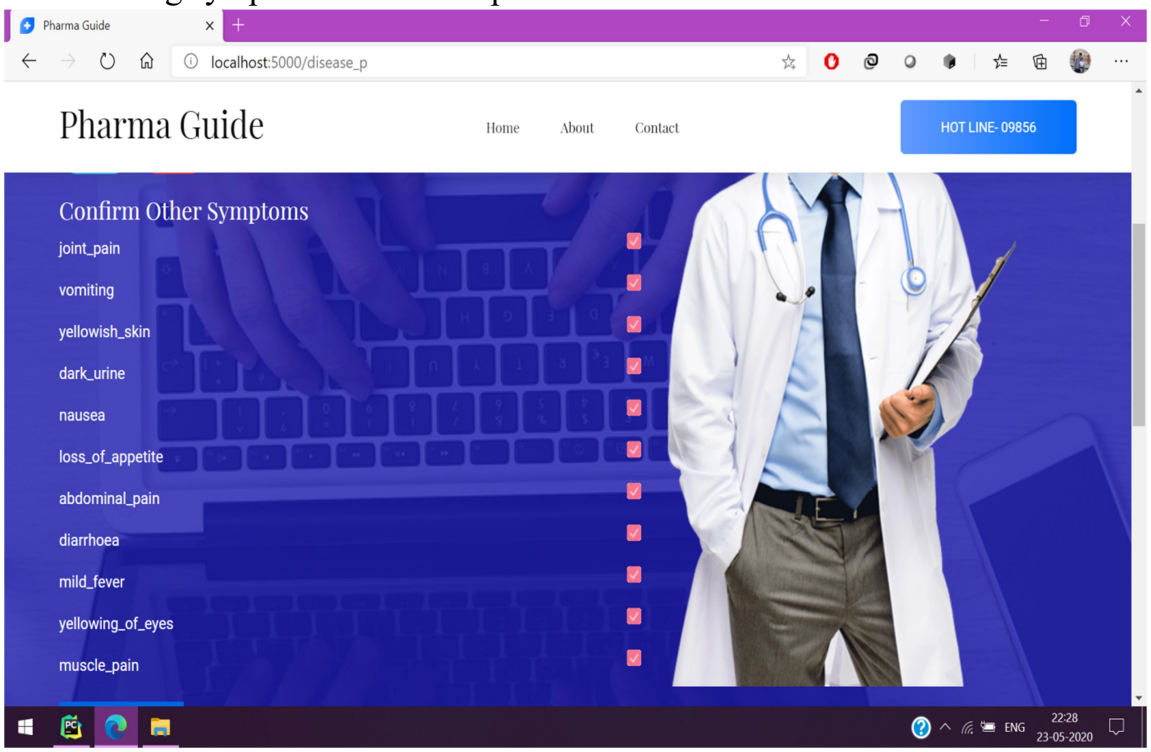
4. Selecting any disease ex: “muscle pain” and then click proceed.



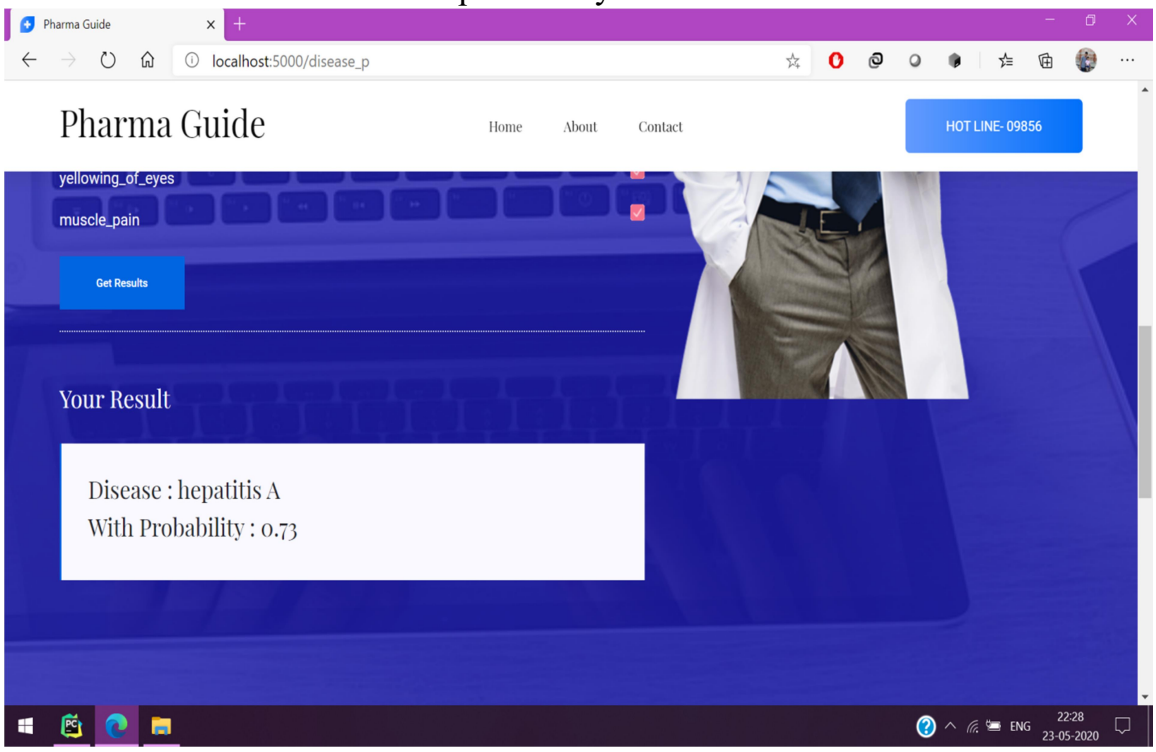
5. Now System will prompt to confirm other symptoms. Press Yes or No.



6. Now disease is already predicted and now confirming other symptoms by dis-selecting symptoms. And then press "Get Results".



7. Now results are fetched with its probability.



Project Planing

1. **Methodology**

Pharma Guide has been already implemented using different techniques like decision tree, Random Forest, SVR and Naive Bayes algorithm. From the analysis it was found that Decision Classifier is more accurate and most relevant than other techniques.

So, Pharma Guide uses Decision Tree Classifier for the prediction of different diseases by taking symptoms as independent variable and disease as dependent variable.

2. **Data Collection**

Data collection (data-set) 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. And data is based on already diagnosed patients by doctors. So, data is genuine and reliable.

3. **Algorithm Implemented**

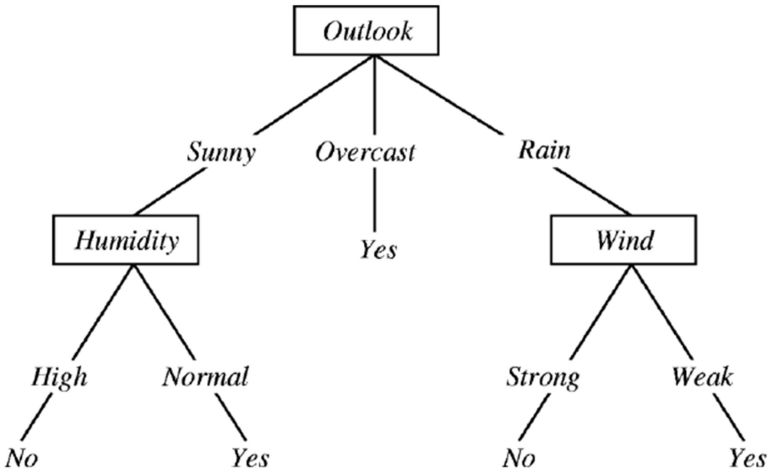
Decision Tree Classifier

Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. A decision tree is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree is composed of internal decision nodes decision node and terminal leaves. Each decision node m implements a test function fm(x) with discrete outcomes labeling the branches. Given an input, at each node, a test is applied and one of the branches is taken depending on the outcome. This process starts at the root and is repeated recursively until a leaf node is hit, at which point the value written in the leaf constitutes the output.

A decision tree is also a non-parametric model in the sense that we do not assume any parametric form for the class densities and the tree structure is not fixed a priori but the tree grows, branches and leaves are added, during learning depending on the complexity of the problem inherent in the data. Decision tree is a classifier in the form of a tree structure which consists of:

 Decision node: specifies a test on a single attribute.

- Leaf node: indicates the value of the target attribute.
- Edge: split of one attribute
- Path: a disjunction of test to make the final decision.



A Decision Tree to predict the weather

Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. Here we use GINI INDEX to partition a tree. Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. It works with categorical target variable "Success" or "Failure". It performs only Binary splits. High value of the value of Gini means high will bethe homogeneity. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a split

Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).

- Calculate Gini for split using weighted Gini score of each node of that split.

Methodology:

Split the data-set which involves iterating over each row, checking if the attribute value is below or above the split value and assigning it to the left or right group Given a data-set, we must check every value on each attribute as a candidate split, evaluate the cost of the split and find the best possible split we could make. Once the best split is found, we can use it as a node in our decision tree. Build a tree recursively until we get all the leaf nodes based on two criteria: Maximum tree depth and Minimum node record. Then we make predictions using

the tree by navigating the tree up to its leaf node. Then we evaluate the accuracy of the algorithm using training test set and cross-validation set.

Advantages of Decision trees:

a) Easy to interpret the decision rules.

b) Non-parametric so it is easy to incorporate a range of numeric or categorical data layers and there is no need to select uni-modal training data.

c) Robust with regard to outliers in training data.

Disadvantages of Decision trees:

a) Decision trees tend to over fit training data which can give poor results when applied to the full data set.

b) Not possible to predict beyond the minimum and maximum limits of the response variable in the training data.

Application of Decision trees:

a) It is used in filtering of spam emails.

b) Decision tree is used in field of medicine. Ex: To predict the type of people prone to specific type of Virus.

Implementation

Class "**PredictDisease**" with all its methods as mentioned in Class diagram. It was done in Python3.7 programming language.

And all required libraries are imported,

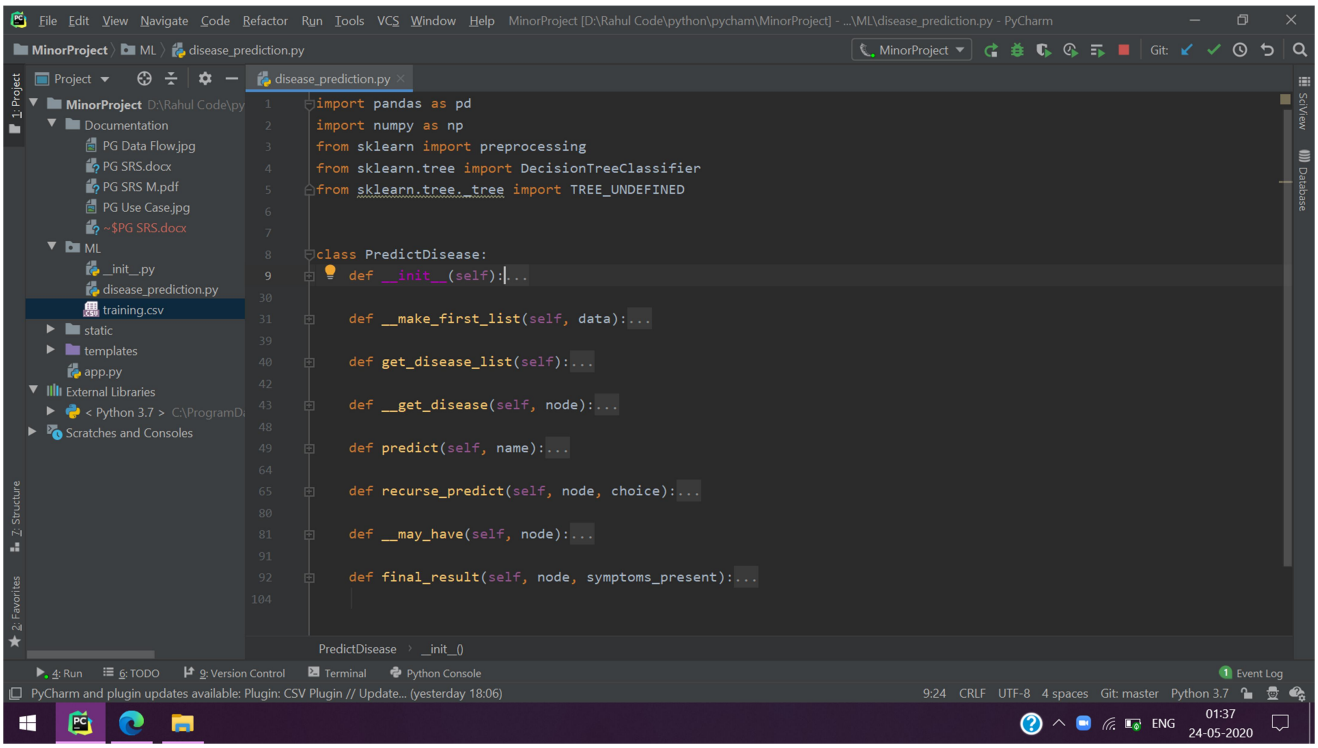
pandas: To handle and manage data-set.

numpy: For mathematical operations on data-array.

***sklearn-preprocessing*:** To get 'LabelEncoder' class for encoding string to numerical values.

sklearn.tree.DecisionTreeClassifier: Predefined class to implement decision tree algorithm.

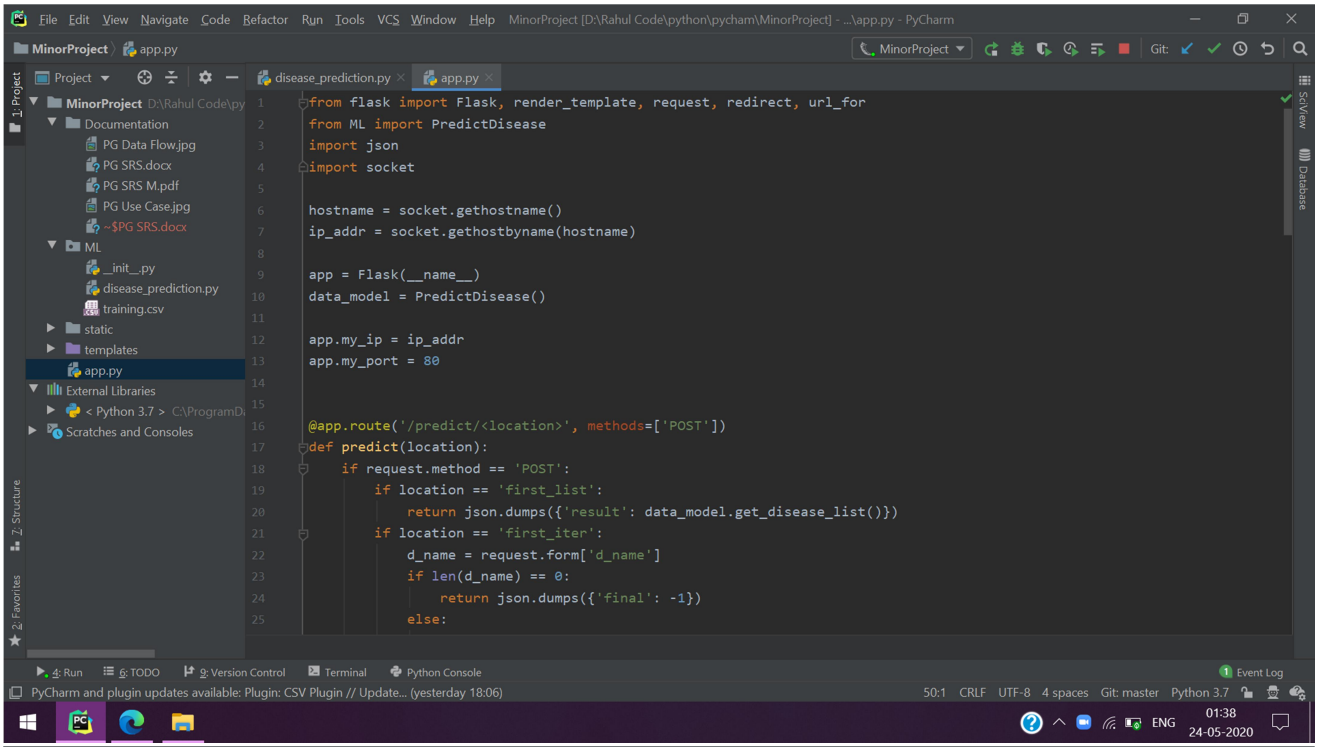
***sklearn.tree._tree*:** Macro to compare undefined nodes in tree.

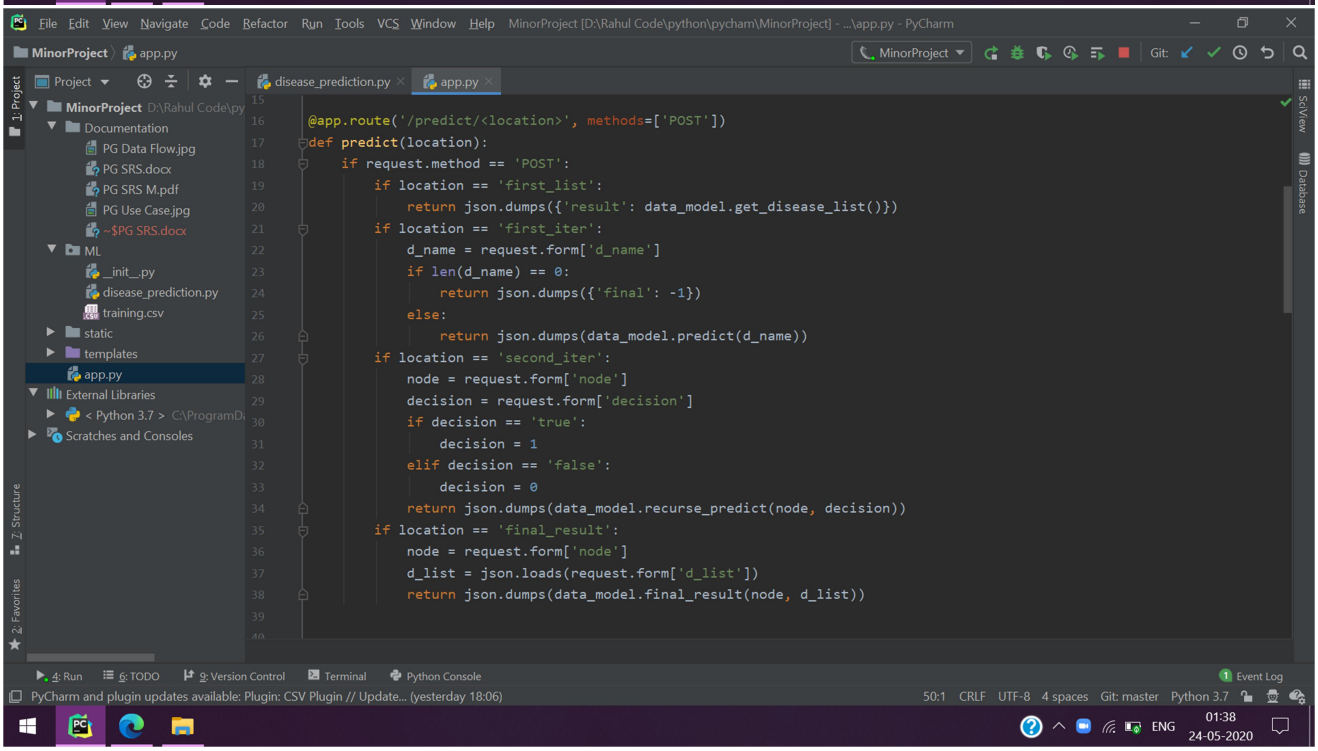


Here flask functions are implemented to accept URL with POST request and then result is returned according to URL.

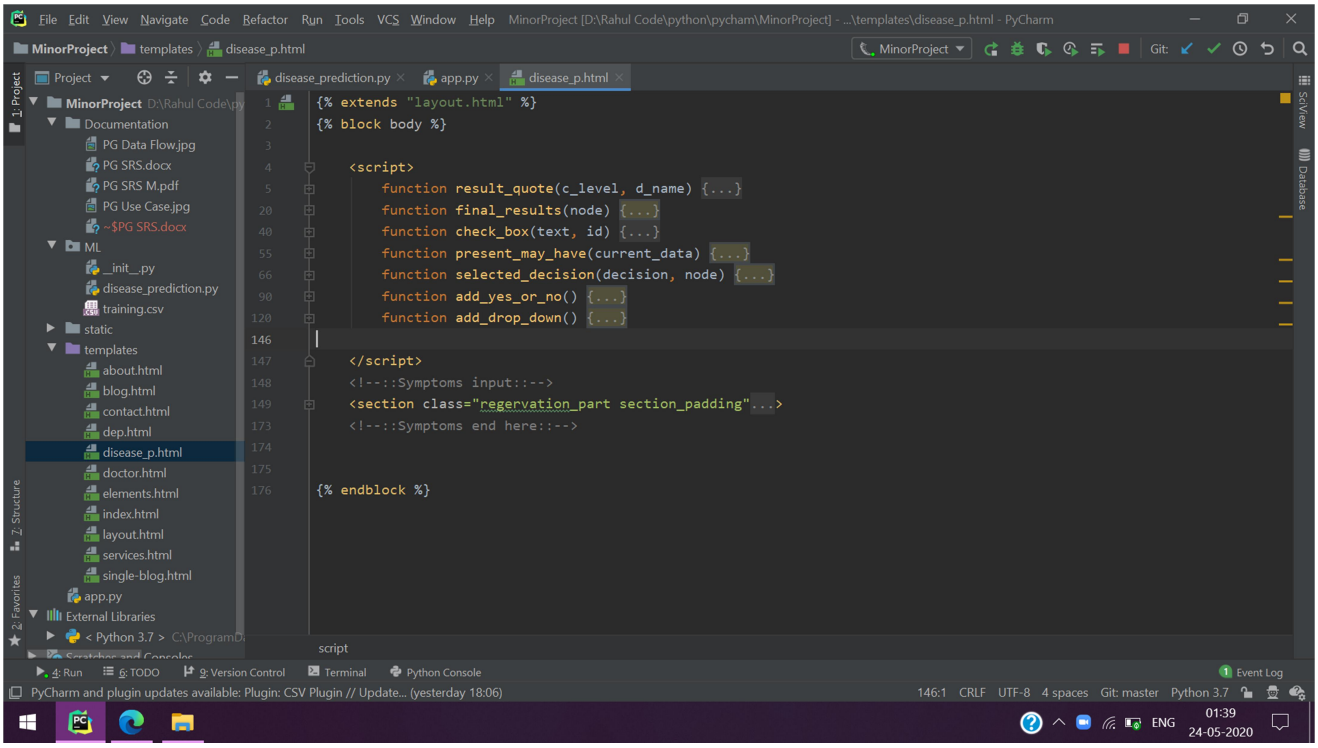
Here IP address set to local-host IPv4 and port to 80(default HTTP port).

Class “PredictDisease” is imported to use here. And other useful flask elements are imported for use.





Now in HTML file, A page was created to provide front-end to user. JavaScript is used to provide dynamic nature to the application. Overall 7 functions are used to achieve dynamic nature. And a section is made for required HTML tag.



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. Accuracy of prediction depends on symptoms entered by user. “Pharma Guide” was successfully implemented using flask framework. And this prediction is done by classifying data by using Decision Tree Classifier.

Recommendation

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.

Future Scope

In near future also focus on optimized selection of number of base classifier for proper selection of classifier in ensemble process. The diversity of medical diagnosis of disease data are increase day to day. Now in future dimension reduction process are also involved in ensemble classification technique.

As demand growing for this type of convenient and easily accessible application, this field possess a bright future for investment.

Individual Contribution Report

"Pharma Guide"

RAHUL SAHU
1705157

The "Pharma Guide" is a web application to provide a free service to its users. So they can check there disease in its early stage. This application takes symptoms as input and provides name of disease as output. This application can be used on any browser, so it can be used as a website.

Individual contribution and findings: I worked on frontend and backend. In frontend part HTML, CSS and JavaScript was used. Here JavaScript was used to provide dynamic nature to the data input in application. All user input requests were made by JavaScript, so that it only change specific block of HTML code. If form based input was used then it will load whole page after every input, which will lead to unnecessary data transfer. By this I learned more about HTML, CSS and JavaScript like making http request using JavaScript, manipulation of tag content by JavaScript, template rendering, etc.

In backend part flask framework of python programming language was used to handle http request**.** Data model which was made by fellow group member, I converted it into class (Object Oriented) to get encapsulated data model and then a module was made which consist that class. So it can be used easily.

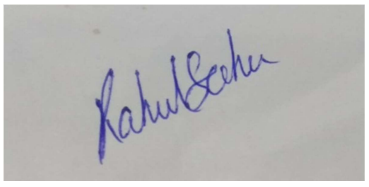
Here I learned request handling using flask framework and also learned to process them. I learned Jinja templating which is a feature of flask framework, where manipulation of template becomes easier. I also learned some more features of python programming language.

Individual Contribution in Project Report: Entire project report was done by me.

Individual Contribution in presentation and demonstration: Some of demonstration part was done by me like demonstrating application working and application code.

Signature of Supervisor

Signature of Student



Individual Contribution Report

"Pharma Guide"

RISHIKA SINHA

1705159

The "Pharma Guide" is a web application to provide a free service to its users. So they can check there disease in its early stage. This application takes symptoms as input and provides name of disease as output. This application can be used on any browser, so it can be used as a website.

Individual contribution and findings: Did Data Analysis and made built the prediction model of the project to predict disease based on the symptoms. Used Decision Tree Classifier in machine learning to implement the feature. The technical work includes designing the algorithm for prediction in python and statistically analyzing the data set. During this project I found out the various ways of checking the probabilities of occurrences of events when it depends on many parameters. One of which was Bayesian Network. I also studied various model in machine learning for different types of prediction. The entire project was a good learning experience in terms of Technical findings and Team Work. It is an agile based model. We'll continue the research to add more such features in the project.

Presentation and demonstration: Made the presentation and demonstrated the entire working of the project. The presentation includes the Novelty, Social Impact, Technology Stack and Scope of the project.

Signature of Supervisor:

Signature of the student: