Report On

DISEASE PREDICTION USING SUPERVISED LEARNING

Submitted in partial fulfillment of the requirements of the Course project in
Semester VII of Fourth Year Computer Engineering

by
Ayush Bujare (Roll No. 01)
Divyen Gharat (Roll No. 02)
Amruta Poojary (Roll No. 09)
Dhanashree Raut (Roll No. 19)

Mentor
Dr. Megha Trivedi

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(2023-24)

Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

CERTIFICATE

This is to certify that the Course Project entitled **"Disease prediction using supervised learning"** is a bonafide work of **1_Ayush Bujare, 2_Divyen Gharat, 9_Amruta Poojary, 19_Dhanashree Raut** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in Semester VII of Fourth Year **"Computer Engineering"** .

_____

Dr Megha Trivedi
Mentor

_____   _____

Dr Megha Trivedi   Dr. H.V. Vankudre

Head of Department   Principal

Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

Course Project Approval

This Mini Project entitled "Disease Prediction Using Supervised Learning" by **01_Ayush Bujare, 02_Divyen Gharat, 09_Amruta Poojary, 19_Dhanashree Raut** is approved for the degree of **Bachelor of Engineering** in Semester VII of Fourth Year **Computer Engineering .**

Examiners

1..
(Internal Examiner Name & Sign)

2……………………………………
(External Examiner name & Sign)

Date:

Place

Contents

Abstract

Disease prediction using Machine Learning ia a system which predicts the disease based on the information or the symptoms he/she enter into the system and provides the accurate results based on that information. If the patient is not much serious and the user just wants to know the type of disease, he/she has been through. It is a system which provides the user the tips and tricks to maintain the health system of the user and it provides a way to find out the disease using this prediction. Now-a-days health industry plays major role in curing the diseases of the patients so this is also some kind of help for the health industry to tell the user and also it is useful for the user in case he/she doesn't want to go to the hospital or any other clinic, so just by entering the symptoms all other useful information the user can get to know the disease he/she is suffering from and the health industry can also get benefit from this system by just asking the symptoms from the user and entering in the system and in just few seconds they can tell the exact and upto some extent the accurate diseases. This Disease Prediction Using Machine Learning is completely done with the help of Machine Learning and Python Programming language with Tkinter interface for it and also using the dataset that is available previously by the hospitals using that will predict the disease.

Acknowledgements

1 Introduction

Introduction

Disease Prediction Using Machine Learning is a system which predicts the disease based on the information provided by the user. It also predicts the disease of the patient based on the information or the symptoms he/she entered into the system. The user can type in the symptoms if he/she doesn't want to go to the hospital or clinics and they will come to know the disease they are suffering from by just sitting at home. This also benefits the health industry and makes it easier for them.

This DPUML is previously done by many other organizations but our intention is ti make it different and beneficial for the users who are using this system. This Disease Prediction is completely done with Machine Learning and Python Programming Language with Tkinter interface for it and also the dataset that is available previously by the hospitals using that we will predict the disease.

Now-a-days doctors are adoping many scientific technologies and methodology for both identification and diagnosing not only common diseases but also many fatal diseases. Doctors may sometime fail to take accurate decisions while diagnosing the disease of a patient, therefore disease prediction systems assist in such cases to get health according to research there are 40% people who ignore the general diseases which leads to harmful diseases later.

Problem Statement & Objectives

In Health industry sometimes there are various problems related to machines or devices which will give wrong or unaccepted results, so to avoid such results and to get correct and desired results we are building a program or project which will give accurate predictions based on the information provided by the user and also the dataset that is available. This will help the people in need. It also takes time and money to go to the hospital or clinics. To get to know the disease in early stages of the symptoms this disease prediction will help a lot of people ranging from teenagers to adults. All other necessary measures can be taken care of and the disease can be cured and known in early stages

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

2   Literature survey

Survey of Existing System/SRS

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

Limitation Existing system or Research gap

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 Disease Predictor 7 histories, suggest previously unconsidered concerns, and facilitating discussion about early testing and prevention. (A.Davis, V.Chawla, Blumm, Christakis, & Barbasi, 2008)

Mini Project Contribution

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)

3 Proposed System

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 cannot 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 Prediction system has data sets collected from different health related sites.

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.
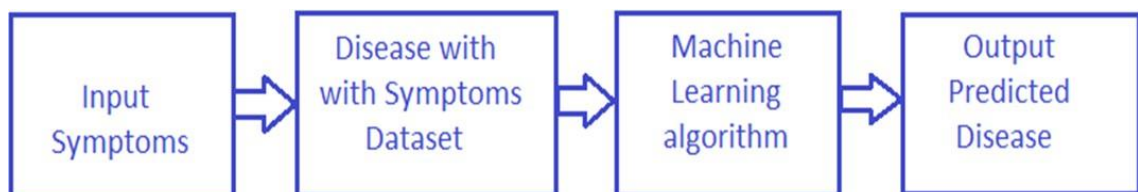
Block Diagram:

Fig a. Process diagram of our application

Algorithm and Process Design:

Module Description:

When the symptoms are given as input, the system will forecast the disease. The Naive Bayesian algorithm will be used to predict the sickness. This approach, according to the literature review, achieves the highest accuracy for a larger dataset. The collection includes illness labels as well as symptoms for each ailment. 70% of the dataset will be

utilised as training data, while 30% will be used as test data. The dataset would be used for training and testing, and the desired output would be obtained.

Naïve Bayes Algorithm:

This algorithm takes the user's input and forecasts the most likely ailment. The dataset and the machine learning technique are used to do this. The Naive Bayesian algorithm is used here, and it uses a probabilistic approach. We used Scikit to understand how to use the library and put it into practise. We utilised multinomial NB for this because there are multiple variants, i.e. multiple symptoms.

Bayesian Theorem:

- The Bayesian theorem is used to predict the class label (in our case, disease) for a given tuple.
- Let X be a tuple of symptoms, and H be a hypothesis, such as that the data tuple X (symptoms) belongs to the class C. (disease)
- We are looking for the chance that tuple X belongs to class C in classification issues, given that we know the attribute description of X.

Dataset

Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0392680_shortness of breath
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall
		UMLS:C0039070_syncope
		UMLS:C0042571_vertigo
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0030252_palpitation
		UMLS:C0027497_nausea
		UMLS:C0002962_angina pectoris
		UMLS:C0438716_pressure chest
UMLS:C0011847_diabetes	1421	UMLS:C0032617_polyuria
		UMLS:C0085602_polydypsia
		UMLS:C0392680_shortness of breath
		UMLS:C0008031_pain chest
		UMLS:C0004093_asthenia
		UMLS:C0027497_nausea
		UMLS:C0085619_orthopnea
		UMLS:C0034642_rale
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0241526_unresponsiveness
		UMLS:C0856054_mental status changes
		UMLS:C0042571_vertigo
		UMLS:C0042963_vomiting
		UMLS:C0553668_labored breathing

Table 1. Dataset of our project

The dataset is from a study conducted at Columbia University. It consists of 150 illnesses, with an average of 810 symptoms for each illness. 70% of the dataset used for training was created with all combinatorial inputs in mind. Symptoms present in the corresponding disease were marked as 1 and remained 0.

It consists of 5 drop-down options that have passed the list of symptoms. The user can select any five symptoms and click the Predict button to see the predicted illness in the text box.

GUI:

We used the Tkinter package for the user interface. Tkinter is Python's standard GUI library. Combining Python and Tkinter makes it quick and easy to create GUI applications. Tkinter provides a powerful object-oriented interface to the TkGUI toolkit.

Details of Hardware & Software

System reqirements:

- Operating System : Windows 7,10 or Higher Versions
- Platform : Jupiter Notebook
- Front End ; Python Tkinter
- Back End : Python and Files
- Programming Language : Python

Technologies used:

In this project we have used the core Python concepts.

We have also used the Machine Learning Algorithm. That is Naïve Bayes.

Vs code is used to run and implement the code.

Hardware used:

- System : Pentium 4, Intel Core i3,i5,i7 and 2 GHz Minimum
- RAM : 512Mb or above
- Hard Disk : 10 GB or above
- Input Device : Keyboard and Mouse
- Output Device : Monitor or PC

Code

```
from tkinter import *
from tkinter import messagebox
import numpy as np
```

```python
import pandas as pd

l1=['itching','skin_rash','nodal_skin_eruptions','continuous_sneezing','shivering','chills'
,'joint_pain',

'stomach_pain','acidity','ulcers_on_tongue','muscle_wasting','vomiting','burning_mictu
rition','spotting_ urination','fatigue',

'weight_gain','anxiety','cold_hands_and_feets','mood_swings','weight_loss','restlessne
ss','lethargy','patches_in_throat',

'irregular_sugar_level','cough','high_fever','sunken_eyes','breathlessness','sweating','de
hydration','indigestion',

'headache','yellowish_skin','dark_urine','nausea','loss_of_appetite','pain_behind_the_e
yes','back_pain','constipation',

'abdominal_pain','diarrhoea','mild_fever','yellow_urine','yellowing_of_eyes','acute_liv
er_failure','fluid_overload',

'swelling_of_stomach','swelled_lymph_nodes','malaise','blurred_and_distorted_vision'
,'phlegm','throat_irritation',

'redness_of_eyes','sinus_pressure','runny_nose','congestion','chest_pain','weakness_in
_limbs','fast_heart_rate',

'pain_during_bowel_movements','pain_in_anal_region','bloody_stool','irritation_in_an
us','neck_pain','dizziness','cramps',

'bruising','obesity','swollen_legs','swollen_blood_vessels','puffy_face_and_eyes','enlar
ged_thyroid','brittle_nails',

'swollen_extremeties','excessive_hunger','extra_marital_contacts','drying_and_tinglin
g_lips','slurred_speech','knee_pain','hip_joint_pain',
```

'muscle_weakness','stiff_neck','swelling_joints','movement_stiffness','spinning_movements','loss_of_balance','unsteadiness','weakness_of_one_body_side',
   'loss_of_smell','bladder_discomfort','foul_smell_of urine','continuous_feel_of_urine','passage_of_gases','internal_itching','toxic_look_(typhos)',

'depression','irritability','muscle_pain','altered_sensorium','red_spots_over_body','belly_pain','abnormal_menstruation','dischromic _patches',

'watering_from_eyes','increased_appetite','polyuria','family_history','mucoid_sputum','rusty_sputum','lack_of_concentration','visual_disturbances',

'receiving_blood_transfusion','receiving_unsterile_injections','coma','stomach_bleeding','distention_of_abdomen','history_of_alcohol_consumption',

'fluid_overload','blood_in_sputum','prominent_veins_on_calf','palpitations','painful_walking','pus_filled_pimples','blackheads','scurring','skin_peeling',

'silver_like_dusting','small_dents_in_nails','inflammatory_nails','blister','red_sore_around_nose','yellow_crust_ooze']

disease=['Fungal infection','Allergy','GERD','Chronic cholestasis','Drug Reaction',
    'Peptic ulcer diseae','AIDS','Diabetes','Gastroenteritis','Bronchial Asthma','Hypertension',
    ' Migraine','Cervical spondylosis',
    'Paralysis (brain hemorrhage)','Jaundice','Malaria','Chicken pox','Dengue','Typhoid','hepatitis A',
'Hepatitis B','Hepatitis C','Hepatitis D','Hepatitis E','Alcoholic hepatitis','Tuberculosis',
'Common Cold','Pneumonia','Dimorphic hemmorhoids(piles)',
'Heartattack','Varicoseveins','Hypothyroidism','Hyperthyroidism','Hypoglycemia','Osteoarthristis',
'Arthritis','(vertigo) Paroymsal Positional Vertigo','Acne','Urinary tract infection','Psoriasis',
'Impetigo']

```
l2=[]
for x in range(0,len(l1)):
    l2.append(0)
```

TESTING DATA

```
tr=pd.read_csv("Testing.csv")
tr.replace({'prognosis':{'Fungal          infection':0,'Allergy':1,'GERD':2,'Chronic
cholestasis':3,'Drug Reaction':4,
'Peptic    ulcer    diseae':5,'AIDS':6,'Diabetes    ':7,'Gastroenteritis':8,'Bronchial
Asthma':9,'Hypertension ':10,
'Migraine':11,'Cervical spondylosis':12,
'Paralysis        (brain        hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken
pox':16,'Dengue':17,'Typhoid':18,'hepatitis A':19,
'Hepatitis    B':20,'Hepatitis    C':21,'Hepatitis    D':22,'Hepatitis    E':23,'Alcoholic
hepatitis':24,'Tuberculosis':25,
'Common      Cold':26,'Pneumonia':27,'Dimorphic      hemmorhoids(piles)':28,'Heart
attack':29,'Varicose veins':30,'Hypothyroidism':31,
'Hyperthyroidism':32,'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,
'(vertigo)    Paroymsal    Positional    Vertigo':36,'Acne':37,'Urinary    tract
infection':38,'Psoriasis':39,
'Impetigo':40}},inplace=True)
```

```
X_test= tr[l1]
y_test = tr[["prognosis"]]
np.ravel(y_test)
```

TRAINING DATA

```
df=pd.read_csv("Training.csv")
```

```
df.replace({'prognosis':{'Fungal          infection':0,'Allergy':1,'GERD':2,'Chronic
cholestasis':3,'Drug Reaction':4,
'Peptic    ulcer    diseae':5,'AIDS':6,'Diabetes    ':7,'Gastroenteritis':8,'Bronchial
Asthma':9,'Hypertension ':10,
'Migraine':11,'Cervical spondylosis':12,
```

'Paralysis (brain hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken pox':16,'Dengue':17,'Typhoid':18,'hepatitis A':19,

'Hepatitis B':20,'Hepatitis C':21,'Hepatitis D':22,'Hepatitis E':23,'Alcoholic hepatitis':24,'Tuberculosis':25,

'Common Cold':26,'Pneumonia':27,'Dimorphic hemmorhoids(piles)':28,'Heart attack':29,'Varicose veins':30,'Hypothyroidism':31,

'Hyperthyroidism':32,'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,

'(vertigo) Paroymsal Positional Vertigo':36,'Acne':37,'Urinary tract infection':38,'Psoriasis':39,

'Impetigo':40}},inplace=True)

```
X= df[l1]

y = df[["prognosis"]]
np.ravel(y)

def message():
    if (Symptom1.get() == "None" and  Symptom2.get() == "None" and Symptom3.get() == "None" and Symptom4.get() == "None" and Symptom5.get() == "None"):
        messagebox.showinfo("OPPS!!", "ENTER  SYMPTOMS PLEASE")
    else :
        NaiveBayes()

def NaiveBayes():
    from sklearn.naive_bayes import MultinomialNB
    gnb = MultinomialNB()
    gnb=gnb.fit(X,np.ravel(y))
    from sklearn.metrics import accuracy_score
    y_pred = gnb.predict(X_test)
    print(accuracy_score(y_test, y_pred))
    print(accuracy_score(y_test, y_pred, normalize=False))

    psymptoms                                                     =
[Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Symptom5.get()]
```

```python
    for k in range(0,len(l1)):
        for z in psymptoms:
            if(z==l1[k]):
                l2[k]=1


    inputtest = [l2]
    predict = gnb.predict(inputtest)
    predicted=predict[0]


    h='no'
    for a in range(0,len(disease)):
        if(disease[predicted] == disease[a]):
            h='yes'
            break


    if (h=='yes'):
        t3.delete("1.0", END)
        t3.insert(END, disease[a])
    else:
        t3.delete("1.0", END)
        t3.insert(END, "No Disease")

root = Tk()
root.title(" Disease Prediction From Symptoms")
root.configure()


Symptom1 = StringVar()
Symptom1.set(None)
Symptom2 = StringVar()
Symptom2.set(None)
Symptom3 = StringVar()
Symptom3.set(None)
Symptom4 = StringVar()
Symptom4.set(None)
Symptom5 = StringVar()
```

```python
Symptom5.set(None)

w2 = Label(root, justify=LEFT, text=" Disease Prediction From Symptoms ")
w2.config(font=("Elephant", 30))
w2.grid(row=1, column=0, columnspan=2, padx=100)


NameLb1 = Label(root, text="")
NameLb1.config(font=("Elephant", 20))
NameLb1.grid(row=5, column=1, pady=10,  sticky=W)


S1Lb = Label(root, text="Symptom 1")
S1Lb.config(font=("Elephant", 15))
S1Lb.grid(row=7, column=1, pady=10 , sticky=W)


S2Lb = Label(root, text="Symptom 2")
S2Lb.config(font=("Elephant", 15))
S2Lb.grid(row=8, column=1, pady=10, sticky=W)


S3Lb = Label(root, text="Symptom 3")
S3Lb.config(font=("Elephant", 15))
S3Lb.grid(row=9, column=1, pady=10, sticky=W)


S4Lb = Label(root, text="Symptom 4")
S4Lb.config(font=("Elephant", 15))
S4Lb.grid(row=10, column=1, pady=10, sticky=W)


S5Lb = Label(root, text="Symptom 5")
S5Lb.config(font=("Elephant", 15))
S5Lb.grid(row=11, column=1, pady=10, sticky=W)


lr = Button(root, text="Predict",height=2, width=20, command=message)
lr.config(font=("Elephant", 15))
lr.grid(row=15, column=1,pady=20)


OPTIONS = sorted(l1)
```

```python
S1En = OptionMenu(root, Symptom1,*OPTIONS)
S1En.grid(row=7, column=2)


S2En = OptionMenu(root, Symptom2,*OPTIONS)
S2En.grid(row=8, column=2)


S3En = OptionMenu(root, Symptom3,*OPTIONS)
S3En.grid(row=9, column=2)


S4En = OptionMenu(root, Symptom4,*OPTIONS)
S4En.grid(row=10, column=2)


S5En = OptionMenu(root, Symptom5,*OPTIONS)
S5En.grid(row=11, column=2)


NameLb = Label(root, text="")
NameLb.config(font=("Elephant", 20))
NameLb.grid(row=13, column=1, pady=10,  sticky=W)


NameLb = Label(root, text="")
NameLb.config(font=("Elephant", 15))
NameLb.grid(row=18, column=1, pady=10,  sticky=W)


t3 = Text(root, height=2, width=30)
t3.config(font=("Elephant", 20))
t3.grid(row=20, column=1 , padx=10)


root.mainloop()
```

Experiment and Results for Validation and Verification:
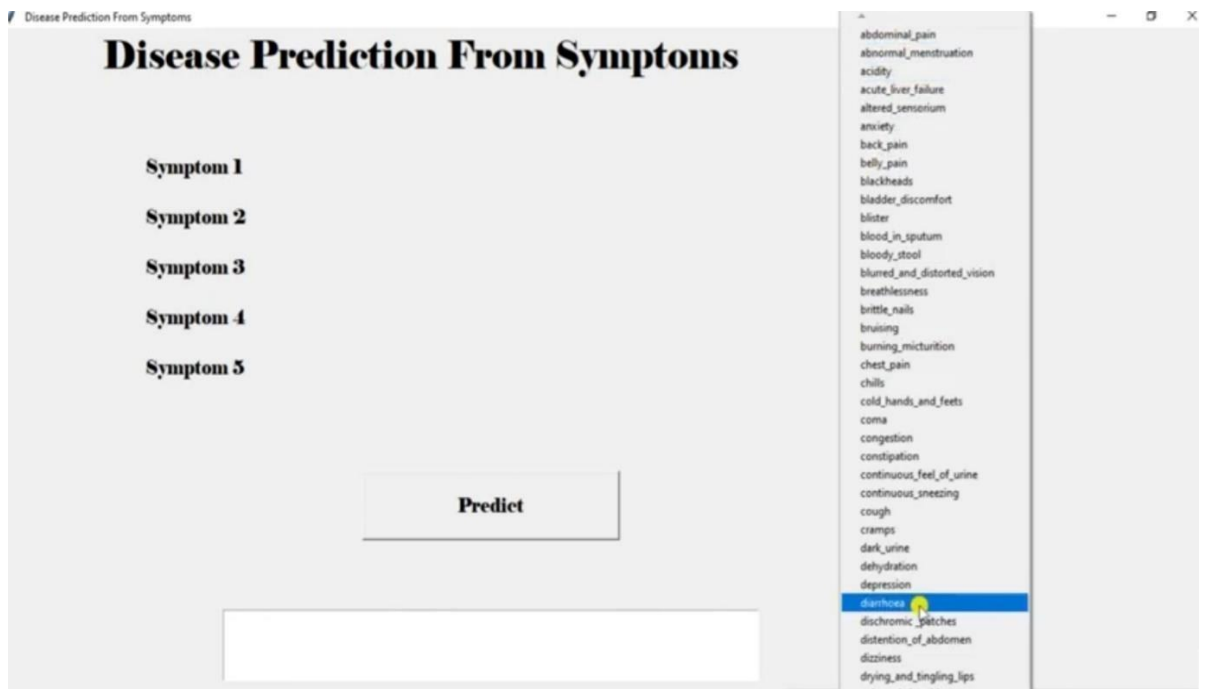

Fig. 1 GUI of the Output

Fig. 2 User entering symptoms

Fig. 3 Output generated by our algorithm

Analysis:

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.

It is easy to operate as the most of the users have 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.

Conclusion and Futurework:

The project is designed so that the system receives the user's symptoms as input and produces output. NS. Predict illness. Users can choose from a minimum of 1 to a maximum of 5 symptoms. If only one symptom is entered, the accuracy will be reduced. The more symptoms you have, the higher the accuracy.

- Facility for modifying user detail.
- More interactive user interface.
- Facilities for Backup creation.
- Can be done as Web page.
- Can be done as Mobile Application.
- More details and Latest Diseases.

References

[1] A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based On Medical History.

[2] Adam, S., & Parveen, A. (2012). Prediction System For Heart Disease Using Naive Bayes.

[3] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification With Naive Bayes Approach. Information Technology Journal .

[4] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical History.

[5] JyotiSoni, Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview Of Heart Disease Prediction.

[6] K.M. Al-Aidaroos, A. B. (n.d.). 2012. Medical Data Classification With Naive Bayes Approach .

[7] NisharBanu, MA; Gomathy, B;. (2013). Disease Predicting System Using Data Mining Techniques

1. www.youtube.com

2. https://stackoverflow.com/

Plagiarism