DISEASE PREDICTION USING NAIVE BAYES CLASSIFIER

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Roop Chandrika Mallela (37110648)
Reddy Lakshmi Bhavani (37110637)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119

MARCH – 2021

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Roop Chandrika Mallela (37110648)** and **Reddy Lakshmi Bhavani (37110637)** who carried out the project entitled “**Disease Prediction using Naive Bayes Classifier”** under my supervision from November 2020 to March 2021.

Internal Guide

Dr. B. ANKAYARKANNI M.E., Ph.D.

Head of the Department

Submitted for Viva voce Examination held on ____________________

Internal Examiner **External Examiner**

DECLARATION

I **Roop Chandrika Mallela (Reg No:37110648) and Reddy Lakshmi Bhavani (Reg No: 37110637)** hereby declare that the Project Report entitled "**Disease Prediction using Naïve Bayes Classifier**" done by us under the guidance of **Dr. B. Ankayarkanni M.E., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2017-2021.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D.**, **Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D. ,** Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Dr. B. Ankayarkanni M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Health is the most important in every human’s life. Weekly or monthly check up of one’s health is most important for the prevention and also to stay healthy. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus. In this situation, technology plays and important role. In this project, we have used Machine Learning. Machine Learning is the study of computer algorithms that improve automatically from the previous experience. It is widely used nowadays and it is the most efficient domain in health care. We will develop a GUI to get the symptoms from the user. The models used in this project are Naive Bayes and Decision Tree. The output is the disease, the accuracy of model, its definition and the treatment of the particular disease based on the symptoms given by the individual. As we all know the saying which tells that “Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease”. This project shows detailed explanation of how to find the diseases from symptoms, so that the individual can contact the respective doctor and stay healthy at an early stage.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
ML	Machine Learning
NB	Naive Bayes
DT	Decision Tree
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

CHAPTER 1

INTRODUCTION

Health is the most important in every human’s life. Weekly or monthly check up of one’s health is most important for the prevention and also to stay healthy. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus.

As we all know the saying which tells that “Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease”.

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people’s lives without even considering their own loves. There are also some villages which lack medical facilities.

We have designed a disease prediction system using ML algorithm (Naive Bayes), find the most accurate algorithm, and used it to find the disease and Tkinter for GUI. And also, we have created a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

This project helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor.

A disease predictor can also be called as a virtual doctor, which can predict the disease based on symptoms. Recently due to covid-19 no one are willing to go outside. This disease predictor system can be a most useful as it identifies the disease without even contacting the individual.

1.1 OVERVIEW

A disease is a condition that affects the individual functioning of body totally. Diseases if neglected will lead to the death of an individual. Diseases can be identified by the symptoms of the body of an individual. Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy.

Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus. As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease".

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

Accurate and on-time analysis of any health-related problem is important for the prevention and treatment of the illness. The traditional way of diagnosis may not be sufficient in the case of a serious ailment. In this situation, where everything has turned virtual, the doctors and nurses are putting up maximum efforts to save people's lives even if they have to danger their own.

There are also some remote villages which lack medical facilities. The dataset was processed in ML models Naive Bayes and Decision Tree. While processing the data, symptoms are given as input and the disease was received as an output.This project helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor.

1.2 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 Machine Learning Strategies

In machine learning, tasks square measure typically classified into broad classes. These classes square measure supported however learning is received or however feedback on the educational is given to the system developed. Two of the foremost wide adopted machine learning strategies square measure supervised learning that trains algorithms supported example input and output information that's tagged by humans, and unattended learning that provides the algorithmic program with no tagged information so as to permit it to search out structure at intervals its computer file.

1.3.1 ***Supervised Learning***

In supervised learning, the pc is given example inputs that square measure labelled with their desired outputs.The aim of this technique is for the algorithmic program to be ready to “learn” by comparison its actual output with the “taught” outputs to search out errors, and modify the model consequently. Supervised learning thus uses patterns to predict label values on extra unlabelled information. For example, with supervised learning, an algorithm may be fed data with images of sharks labelled as fish and images of oceans labelled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabelled shark images as fish and unlabelled ocean images as water.

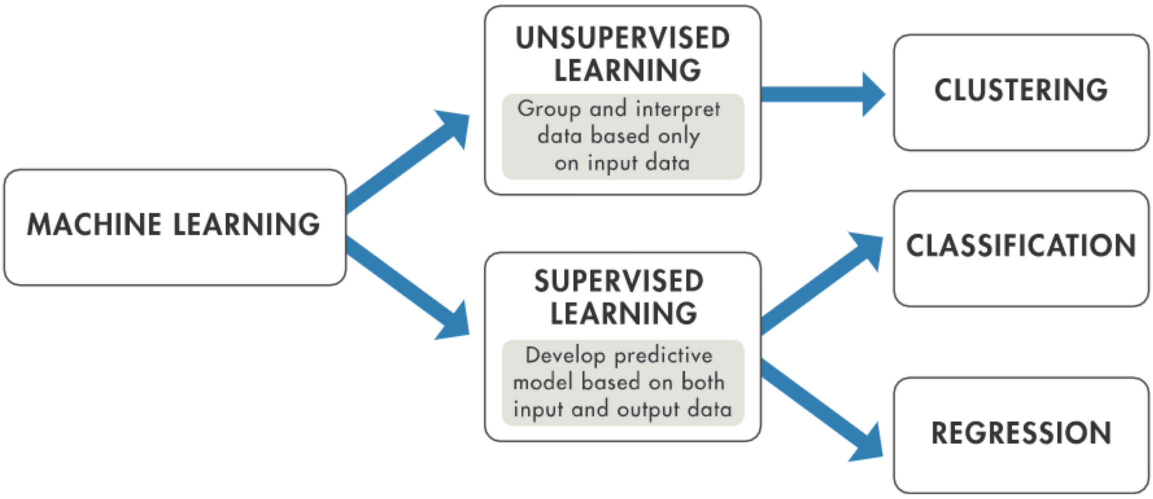
A common use case of supervised learning is to use historical information to predict statistically probably future events. It's going to use historical stock exchange info to anticipate approaching fluctuations, or be used to filter spam emails. In supervised learning, labeled photos of dogs are often used as input file to classify unlabeled photos of dogs.

1.3.2 ***Unattended Learning***

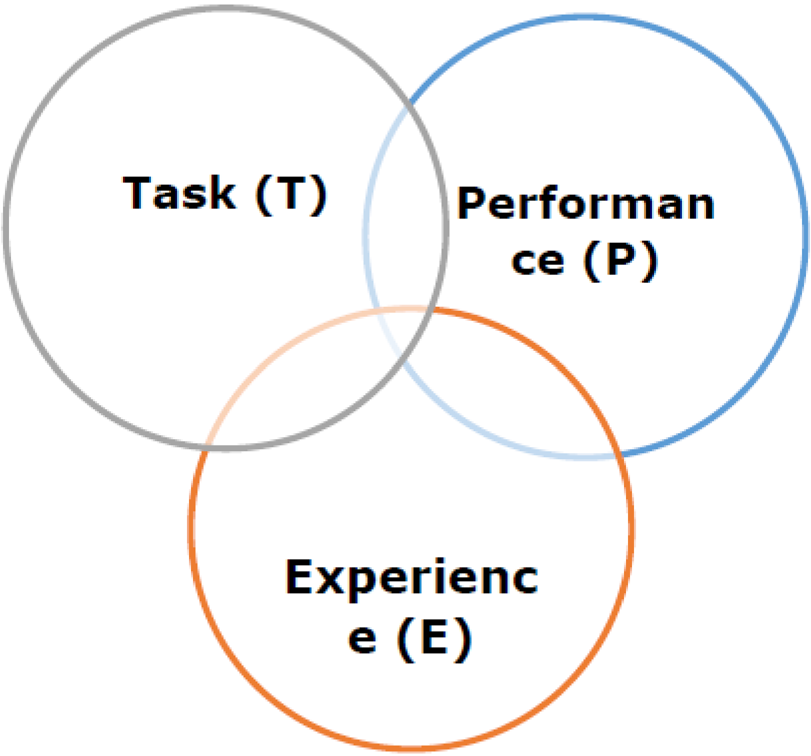
In unattended learning, information is unlabeled, that the learning rule is left to seek out commonalities among its input file. The goal of unattended learning is also as easy as discovering hidden patterns at intervals a dataset, however it should even have a goal of feature learning, that permits the procedure machine to mechanically discover the representations that square measure required to classify data.

Unsupervised learning is usually used for transactional information. You will have an oversized dataset of consumers and their purchases, however as a person's you'll probably not be able to add up of what similar attributes will be drawn from client profiles and their styles of purchases.

With this information fed into Associate in Nursing unattended learning rule, it should be determined that ladies of a definite age vary UN agency obtain unscented soaps square measure probably to be pregnant, and so a promoting campaign associated with physiological condition and baby will be merchandised.


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

1.1 MACHINE LEARNING CLASSIFICATION



1.2 MACHINE LEARNING TASK

CHAPTER 2

LITERATURE SURVEY

2.1 RELATED WORK

- Tarigoppula V.S Sriram et al. in [1] collects the voice dataset from UCI Machine Learning repository and train four algorithms on that dataset. The result is the prediction of Parkinson Disease by considering the most accurate algorithm.
- Shubham Bind et al. in [2] studies about all the available researches in literature to predict the Parkinson diseases.
- K. Gomathi, D. Shanmuga Priya in [3] used different data mining techniques to predict Heart disease, Breast Cancer, Diabetes. The models used are Decision Tree and Naive Bayes Classifier. Performance of both the models was compared and the best classifier is used to predict the above diseases.
- Isha Pandya et al. in [4] used two supervised machine learning algorithms Decision Tree, accuracy 91% and Naïve Bayes classifier, accuracy 87%. Here, they used the combination of both to get the best accuracy. Naïve Bayes Classifier accuracy should be improved.
- Akash C. Jamgade, Prof. S. D. Zade in [5] paper determined the most danger diseases which occur in a person in a locality and community. But, the data collection is difficult.
- Siddhika Arunachalam in [6] six classification algorithms are used after analyzing 14 attributes in the dataset. But, we may get confused which algorithm to use.
- Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE methodologies in [7] used are Support Vector Machines, Artificial Neural Networks, K-Means Algorithm, Decision Trees, Logistic Regression and predicted diseases are breast cancer, lung cancer, heart diseases, diabetes, thyroid or kidney diseases.

- H BENJAMIN FREDRICK DAVID in [8] predicted the occurrence of heart disease using ensemble learning algorithms. But, the Research work can be made to produce an impact in the accuracy of the Decision Tree and Bayesian Classification.
- Harshit Anand et al. in [9], domains of Machine learning and Data Science are used and models are built using numpy, pandas, sklearn, and so on and the model are deployed using Django.
- Goutam Chakraboty et al. in [10], takes into account six features from 23 features in the dataset and predict the risk of chronic kidney disease. It is used to reduce the impact of Chronic Kidney disease (CKD), where creatinine test is not available for all.
- Durga Praveen et al. in [11], applied 5 models namely KNN, SVM, Random Forest, Naïve Bayes and Adaboost and found that KNN, Adaboost has the highest accuracy of all the models. So, any of these two are used for prediction and prevention of the liver disease.
- Sejin Park et al. in [12], early prediction of disease using the previous real-time stroke symptoms. It is implemented at a low cost. Random Forest algorithm is used for validating clinical significance.
- Ahan Chatterjee et al. in [13], they have used machine learning models like Decision tree, SVM, Random Forest, and so on, find the best classifier using the accuracy and use it to predict cancer disease risk in the early stage and prevent it. Simulation model is also designed to manage the patient flow in OPDs.
- Sergio Grueso et al. in [14], they have used dataset taken from ADNI database and selected 47 out of 159 studies for analysis. Deep learning combined with multimodal and multidimensional data is used to achieve the best performance.
- Upendra Kumar in [15], have used Computer-Aided Pre-Screening Tool (CAPST) which improves the accuracy of diagnosis in medicine. By this, fast and accurate prediction risk of disease is found.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

From the above literature survey, We have inferred that all the systems existing predict only particular diseases namely lung disease, breast cancer, heart disease, diabetes by implementing various algorithms on the particular datasets.

After implementing various algorithms, the most accurate one is selected and it is used for prediction of disease. Sometimes, we may get confused of what algorithm to use. Also, all the systems find only the particular disease and not the disease based on the symptoms.

3.2 PROPOSED SYSTEM

We are proposing a system, which uses tkinter for GUI interface. It is a simple user Interface and also time efficient. Our aim with is to get the disease based on symptoms given by the user.

The domain we will use is the machine learning, in that we will be using Naïve Bayes Classifier, which will help us in getting the most accurate predictions easily and also the accuracy is given as output. To reduce time consuming, we will ask only less questions namely the name of the individual and the symptoms the individual is facing.

In this way, our system will be less time consuming and give accurate predictions. And also, we are creating a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

3.3 OBJECTIVE OF PROJECT

Developing a project based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method is the main objective of the project. We have designed a disease prediction system using ML algorithm (Naive Bayes), and also chatbot using decision tree which is used it to find the disease based on the symptoms. Based on the symptoms of an individual, the ML model gives the output, i.e., the disease that the individual might be suffering from. This project helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor.

3.4 SOFTWARE AND HARDWARE REQUIREMENTS

3.4.1 Software Requirements:

- ✓ Python
- ✓ Anaconda
- ✓ Jupyter Notebook

3.4.2 Hardware Requirements:

- ✓ Processor: Intel Core i5
- ✓ RAM: 8GB
- ✓ OS: Windows

3.4.3 Libraries:

- ✓ **Tkinter**- Tkinter is library of python used oftenly by everyone. It is a library which is useed to create GUI based applications easily. It contains so many widgets like radiobutton, textfiled and so on. We have used this for creating account registration screen, login or register screen, prediction interface which is a GUI based application
- ✓ **Sklearn**- Scikit Learn also known as sklearn is a open source library for python programming used for implementing machine learning algorithms. It features various classification, clustering, regression machine learning algorithms. In this it is used for importing machine learning models, get accuracy, get confusion matrix.
- ✓ **Pandas-** Library of python which can be used easily. It gives speed results and also easily understandable. It is a library which can be used without any cost. We have used it for data analysis and to read the dataset.
- ✓ **Matplotlib-** Library of python used for visualising the data using graphs, scatterplots and so on. Here, we have used it for data visualisation.

- ✓ **Numpy-** Library of python used for arrays computation. It has so many functions. We have used this module to change 2-dimensional array into contiguous flattened array by using ravel function.
- ✓ **Pandas Profiling**-This is library of python which can be used by anyone free of cost. It is used for data analysis. We have used this for getting the report of the dataset.

3.5. PROGRAMMING LANGUAGES

3.5.1 *PYTHON*

Python is that the best programing language fitted to Machine Learning. In step with studies and surveys, Python is that the fifth most significant language yet because the preferred language for machine learning and information science. It's owing to the subsequent strengths that Python has –

- ✓ **Easy to be told and perceive-** The syntax of Python is simpler; thence it's comparatively straightforward, even for beginners conjointly, to be told and perceive the language.
- ✓ **Multi-purpose language −** Python could be a multi-purpose programing language as a result of it supports structured programming, object-oriented programming yet as practical programming.
- ✓ **Support of open supply community −** As being open supply programing language, Python is supported by awfully giant developer community. Because of this, the bugs square measure simply mounted by the Python community. This characteristic makes Python terribly strong and adaptative.

3.5.2 DOMAIN

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches. In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. Thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

3.6. SYSTEM ARCHITECTURE

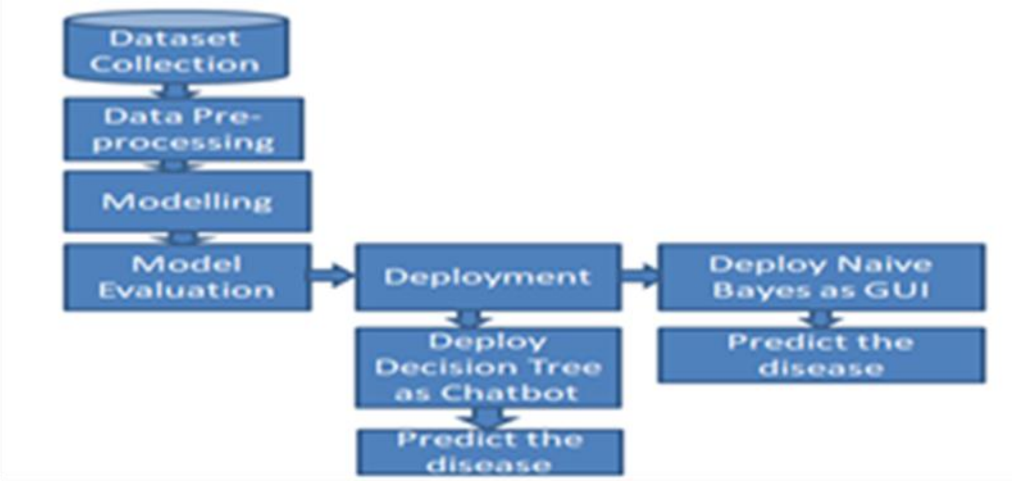


Fig 3.1 System Architecture

3.7 ALGORITHMS USED

3.7.1 NAIVE BAYES

It is a machine learning algorithm for classification problems which is based on Bayes theorem. The primary use of this is to do text classification. The Bayes theorem can be defined as:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

P(C|X) is the probability of hypothesis C for the given data X. This is called the posterior probability.

P(X|C) is the probability of data X given that the hypothesis C was true.

P(C) is the probability of hypothesis C being true. This is called the prior probability of C.

P(X) is the probability of the data and evidence of data and is called marginal probability.

Gaussian Naive Bayes

It follows the same procedure as the Naive Bayes. But for Naive Bayes we need a categorical dataset and for Gaussian Naive Bayes we need a dataset that has all the continuous features.

3.7.2 DECISION TREE

Decision Tree an algorithm whose input and output or known. The information is divided repeatedly using the particular parameter.

Decision tree consists of two main parts namely decision node and leaf nodes. The decision nodes specify the decision at which the parameter should be spilt.

The leaf nodes are the output bought by the decisions. A decision tree asks for either true or false to divide the data.

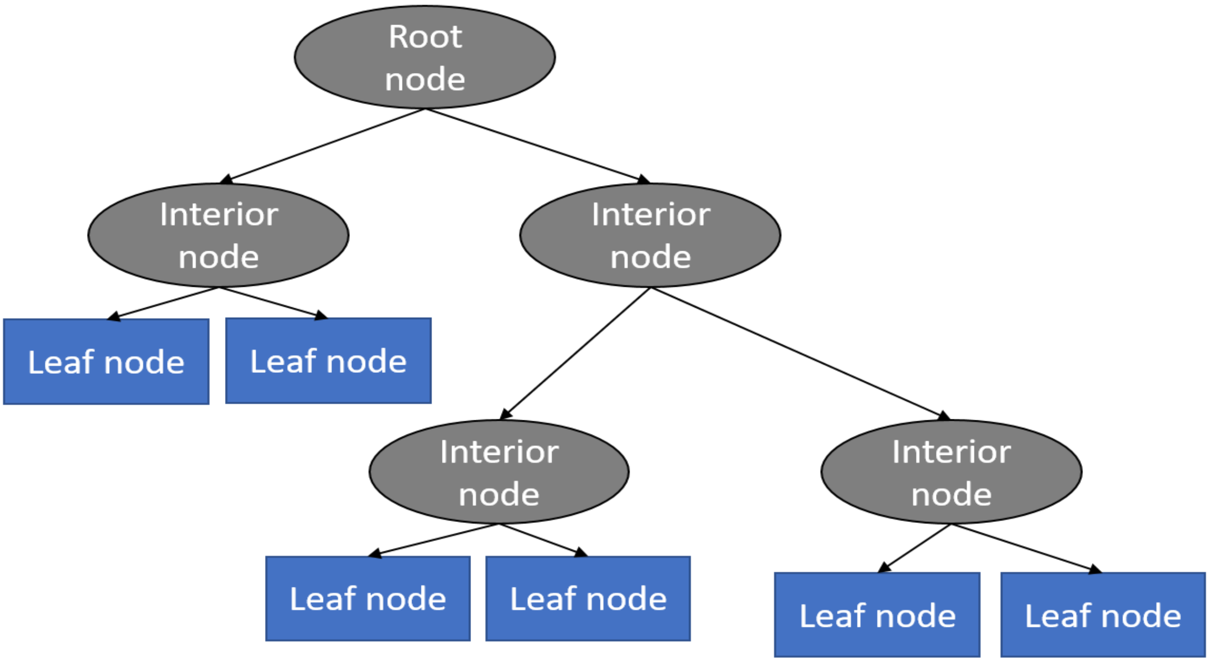


Fig 3.2. Decision Tree

3.8.MODULES

Project contains three parts:

3.8.1 DATASET COLLECTION.

3.8.2 TRAIN AND TEST THE MODEL.

3.8.3 DEPLOY THE MODELS.

- ❑ **Dataset Collection**- We had collected dataset from kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 4920 rows and 133 columns.
- ❑ **Train and Test the model**- We had used the Naïve Bayes Classifier as a model to train the dataset. After training, we had tested the model and found its accuracy.
- ❑ **Deploy the models**- Deployed Naïve bayes by creating interface to get the name, symptoms of an individual. By this, we will get the disease and accuracy of model as the output. We have also created a chatbot using Decision Tree which helps an individual to get the corresponding disease by checking whether he/she is being faced by the symptoms.

Following are the steps to do this project (use Jupyter Notebook):

1. Collect the dataset.
2. Import the necessary libraries.
3. Visualise the dataset.
4. Train the dataset using Naïve Bayes classifier and Decision Tree.
5. Test the model and find the accuracies of both.
6. Deploy the model- a) Naive Bayes as GUI Interface using Tkinter

   b) Chatbot using Decision Tree
7. Predict the disease based on the symptoms given by the user.

CHAPTER 4

RESULTS AND DISCUSSION

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where essential resources are unavailable and people are also not willing to go outside in fear of spreading virus, our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease.This also helps in reduction of the cost and give the correct and fast result.

4.1. WORKING

In this way machine learning when implemented in healthcare can help in satisfying the individual and also take care of their particular disease easily. Naive Bayes, which is the most easy model helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor. We have also created chatbot using decision tree by which an individual can easily find the disease by chatting. Accuracy is defined as the ratio of sum of TP and TN to the sum of TP, TN, FP, FN.

Algorithm	**Accuracy**
Decision Tree	97%
Naive Bayes	100%

Table 4.1 Accuracies

CHAPTER 5

CONCLUSION

5.1. CONCLUSION

The project presented the technique of predicting the disease based on the symptoms of an individual patient. Once the disease is predicted, we could easily manage the medicine resources required for the treatment. Doctors and medical professionals are always required in case of an emergency.

In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease. Our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease.This also helps in reduction of the cost and give the correct and fast result.

The project presented the technique of predicting the disease based on the symptoms of an individual patient. Almost all the ML models gave good accuracy values but the most accurate one is selected and the disease given by it is considered as the disease of an individual. Once the disease is predicted, we could easily manage the medicine resources required for the treatment.

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease. This project would help in lowering the cost required in dealing with the disease and would also make the recovery process easy.

REFERENCES

1. Isha Pandya et al, “Prediction of Heart Disease Using Machine Learning Algorithms”, 2018.
2. Akash C. Jamgade, Prof. S. D. Zade,”Disease Prediction using Machine Learning”, International Research Journal of Engineering and Technology Volume: 06 Issue: 05 May 2019.
3. Siddhika Arunachalam,” Cardiovascular Disease Prediction Model using Machine Learning Algorithms”, International Journal for Research in Applied Science & Engineering Technology ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.
4. Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE, “A REVIEW OF DATA MINING TECHNIQUES IN MEDICINE”, JOURNAL OF INFORMATION SYSTEMS & OPERATIONS MANAGEMENT, Vol. 14.1, May 2020.
5. H BENJAMIN FREDRICK DAVID, “IMPACT OF ENSEMBLE LEARNING ALGORITHMS TOWARDS ACCURATE HEART DISEASE PREDICTION”, DOI: 10.21917/ijsc.2020.0296.
6. Harshit Anand et al. “Hridaya Kalp: A Prototype for Second Generation Chronic Heart Disease Detection and Classification” ,31 July,2020.
7. Goutam Chakraboty et al. “Predicting the Risk of Chronic Kidney Disease using Machine Learning Algorithm”, 11(1), 28 December, 2020 202.
8. A. Durga Praveen et al. “Intelligent Liver Disease Prediction system using Machine Learning Models” ,vol 702. Springer, Singapore. 5 Jan, 2021.
9. Sejin Park et al. “Machine-Learning-Based Elderly Stroke Monitoring System Using Electroencephalography Vital Signals”, 2021, *11*(4), 1761.
10. Ahan Chatterjee et al. “A Machine Learning Approach to prevent cancer”, DOI: 10.4018/978-1-7998-2742-9.ch007, 2021.
11. Sergio Grueso et al. “Machine Learning methods for predicting conversion from mild cognitive impairment to Alzheimer’s disease”, 2021.
12. Upendra Kumar, “Applications of Machine Learning In Disease Pre-Screening”, 10.4018/978-1-7998-7705- 9.ch049, 2021.

APPENDICES

A) SOURCE CODE

CREATION OF CHATBOT

```
def tree_to_code(tree, feature_names):
   tree_ = tree.tree_
    feature_name = [feature_names[i] if i != _tree.TREE_UNDEFINED else
    "undefined!" for i in tree_.features]
    chk_dis=",".join(feature_names).split(",")
    symptoms_present = []
  while True:
            print("Enter the symptom you are experiencing  \t\t\t\t\t\t",end="->")
            disease_input = input("")
           conf,cnf_dis=check_pattern(chk_dis,disease_input)
      if conf==1:
         print("Searches related to input: ")
         for num,it in enumerate(cnf_dis):
            print(num,")",it)
         if num!=0:
            print(f"Select the one you meant (0 - {num}):  ", end="")
            conf_inp = int(input(""))
         else:
            conf_inp=0
         disease_input=cnf_dis[conf_inp]
         break
        else:
         print("Enter valid symptom.")
```

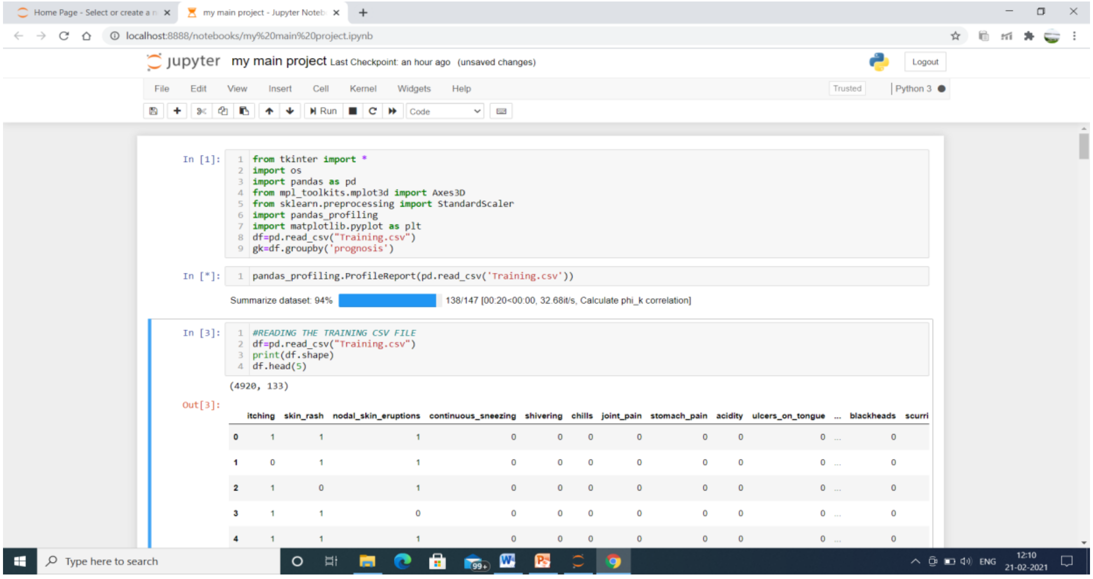
```
 while True:
    try:
        num_days=int(input("Okay. From how many days ? : "))
         break
    except:
        print("Enter number of days.")
def recurse(node, depth):
    indent = "  " * depth
    if tree_.feature[node] != _tree.TREE_UNDEFINED:
        name = feature_name[node]
        threshold = tree_.threshold[node]
         if name == disease_input:
            val = 1
        else:
            val = 0
        if  val <= threshold:
            recurse(tree_.children_left[node], depth + 1)
        else:
            symptoms_present.append(name)
            recurse(tree_.children_right[node], depth + 1)
    else:
        present_disease = print_disease(tree_.value[node])
        red_cols = reduced_data.columns

            symptoms_given=red_cols[reduced_data.loc[present_disease].values[0].nonzero()]
       print("Are you experiencing any ")
```

```
            symptoms_exp=[]
            for syms in list(symptoms_given):
                inp=""
                print(syms,"? : ",end='')
                while True:
                    inp=input("")
                    if(inp=="yes" or inp=="no"):
                        break
                    else:
                        print("provide proper answers i.e. (yes/no) : ",end="")
                if(inp=="yes"):
                    symptoms_exp.append(syms)
                         second_prediction=sec_predict(symptoms_exp)
                        calc_condition(symptoms_exp,num_days)
            if(present_disease[0]==second_prediction[0]):
                print("You may have ", present_disease[0])
                 print(description_list[present_disease[0]])
         else:
                print("You may have ", present_disease[0], "or ", second_prediction[0])
                print(description_list[present_disease[0]])
                print(description_list[second_prediction[0]])
            precution_list=precautionDictionary[present_disease[0])
            print("Take following measures : ")
            for  i,j in enumerate(precution_list):
                print(i+1,")",j)
recurse(0, 1)
```

TKINTER INTERFACE

```
def main_account_screen():
    global main_screen
    main_screen = Tk()
    main_screen.geometry("300x250")
    main_screen.title("Account Login")
     Label(text="Select Your Choice", bg="white", width="300", height="2", font=("Calibri", 13)).pack()
    Label(text="").pack()
    Button(text="Login", height="2", width="30", command = login).pack()
    Label(text="").pack()
    Button(text="Register", height="2", width="30", command=register).pack()
    main_screen.mainloop()
main_account_screen()
gk
```

B) SCREENSHOTS



Home Page - Select or create a n
my main project - Jupyter Noteb
localhost:8888/notebooks/my%20main%20project.ipynb
jupyter
my main project Last Checkpoint: an hour ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Type here to search
ENG
12:10
21-02-2021

```
In [1]:
from tkinter import *
import os
import pandas as pd
from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler
import pandas_profiling
import matplotlib.pyplot as plt
df=pd.read_csv("Training.csv")
gk=df.groupby('prognosis')
```

```
In [*]:
pandas_profiling.ProfileReport(pd.read_csv('Training.csv'))
```

```
Summarize dataset: 94%  138/147 [00:20<00:00, 32.68it/s, Calculate phi_k correlation]
```

```
In [3]:
#READING THE TRAINING CSV FILE
df=pd.read_csv("Training.csv")
print(df.shape)
df.head(5)
```

```
(4920, 133)
```

Out[3]:

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	scurri
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	

Fig B.1 Dataset

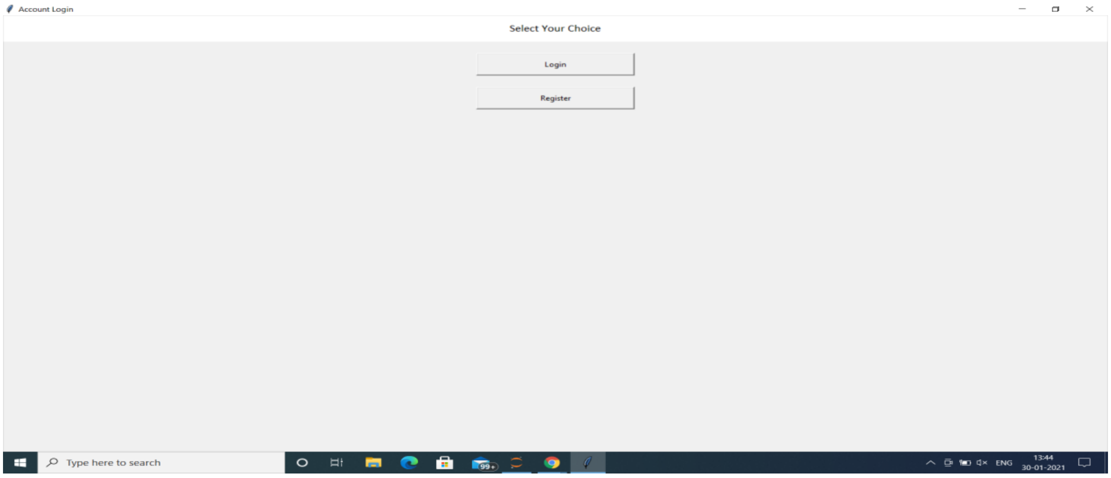
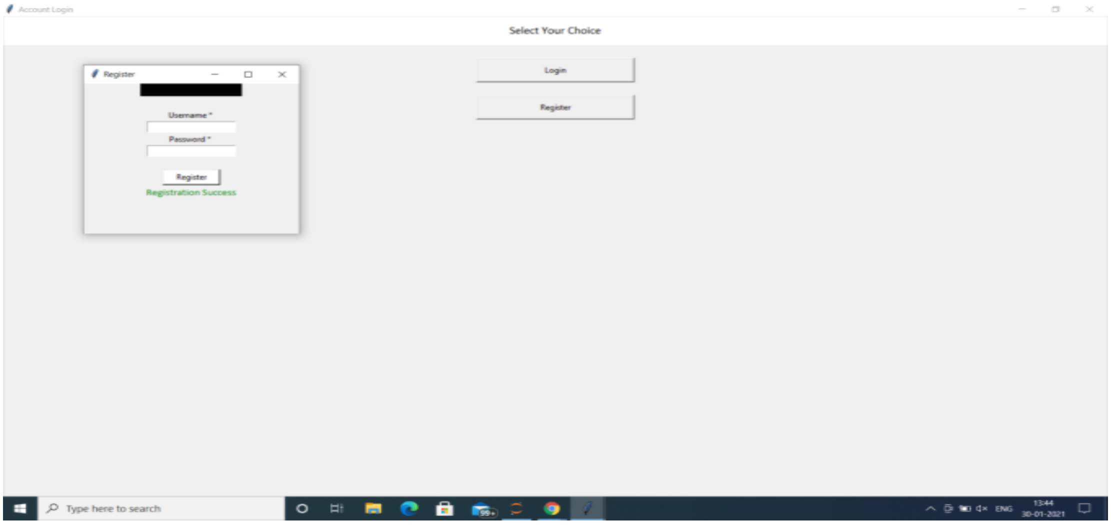


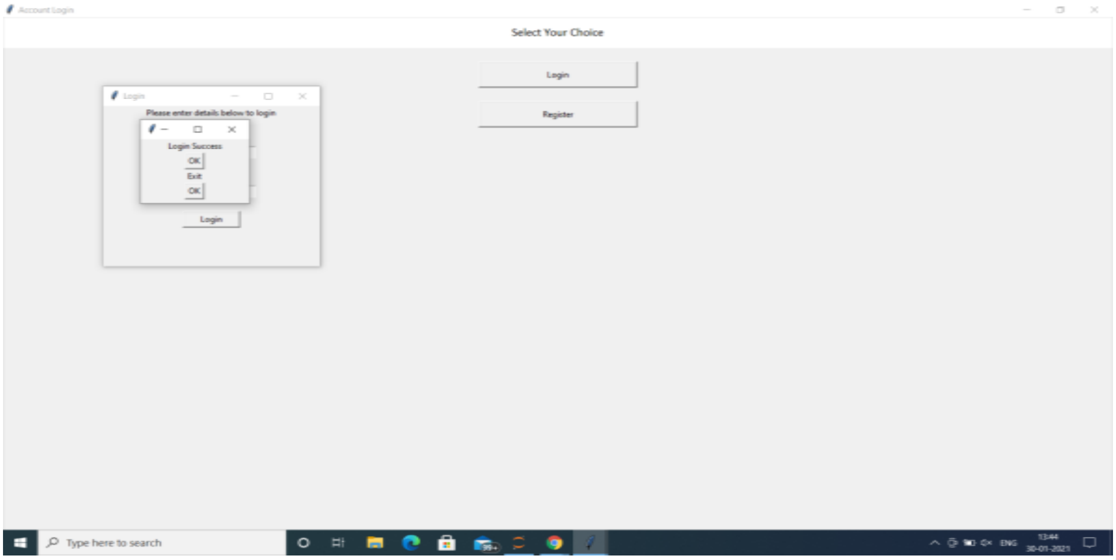
Fig B.2 Account Registration Screen



[redacted]

Registration Success

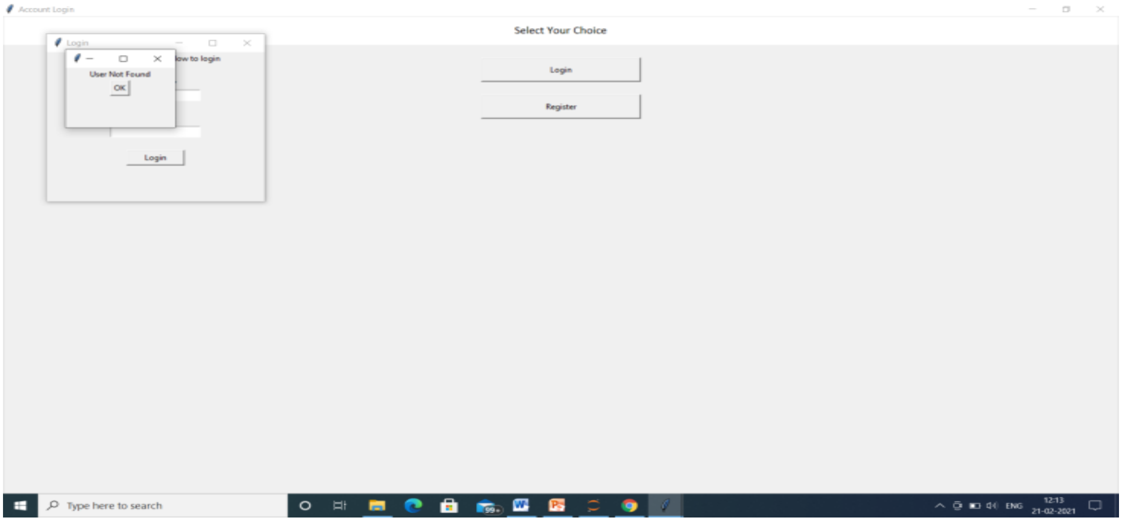
Fig B.3 Registration Success



Please enter details below to login

Login Success

Fig B.4 Login Success



User Not Found

OK

Fig B.5 Invalid Login

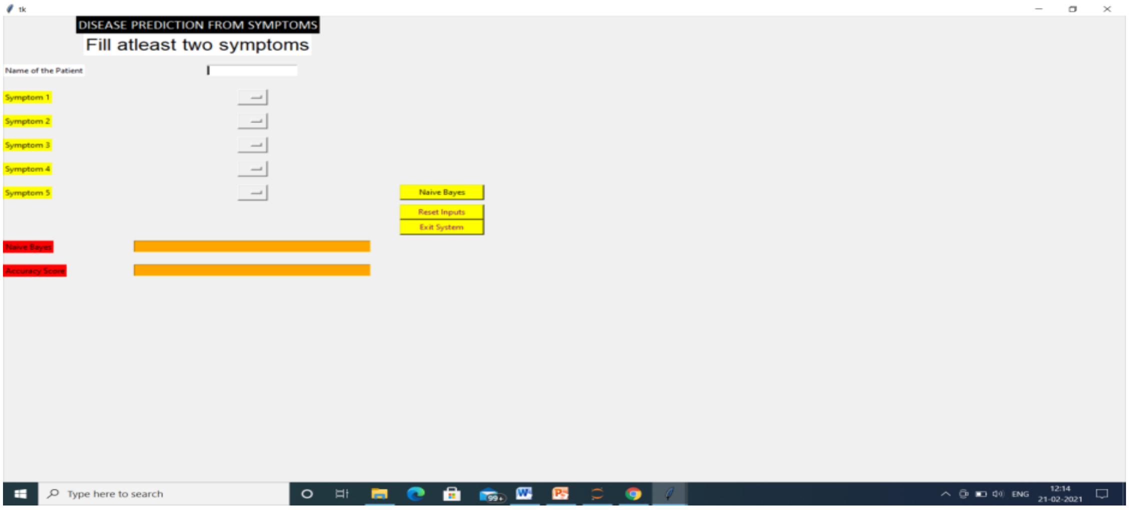
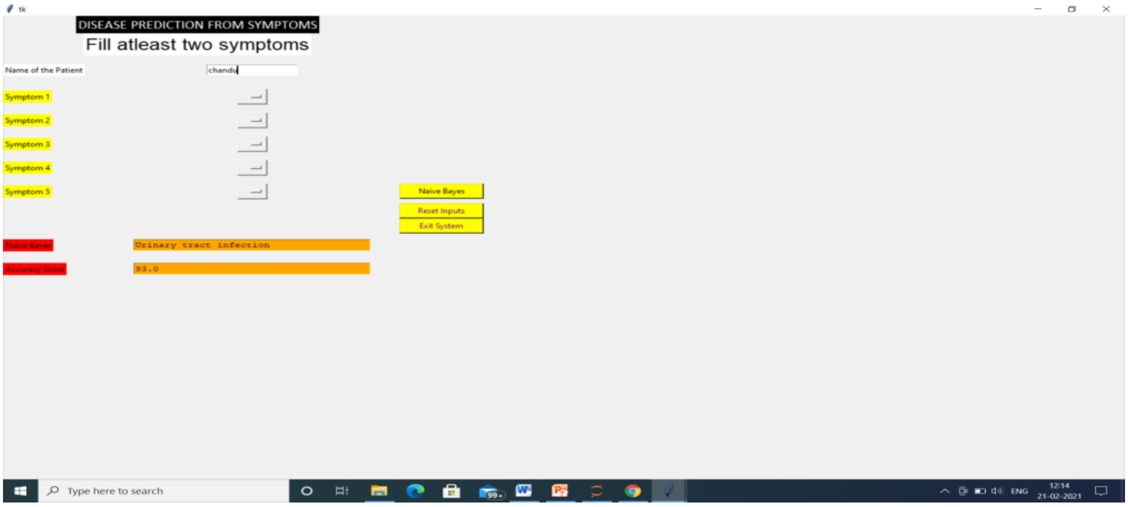


Fig B.6 GUI Interface

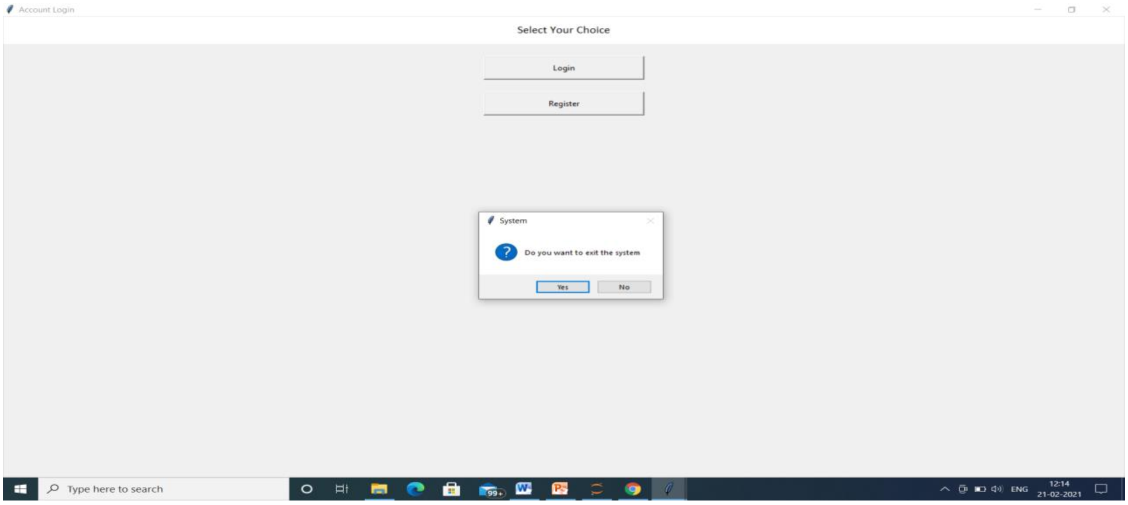


Name of the Patient chandu

Naive Bayes Urinary tract infection

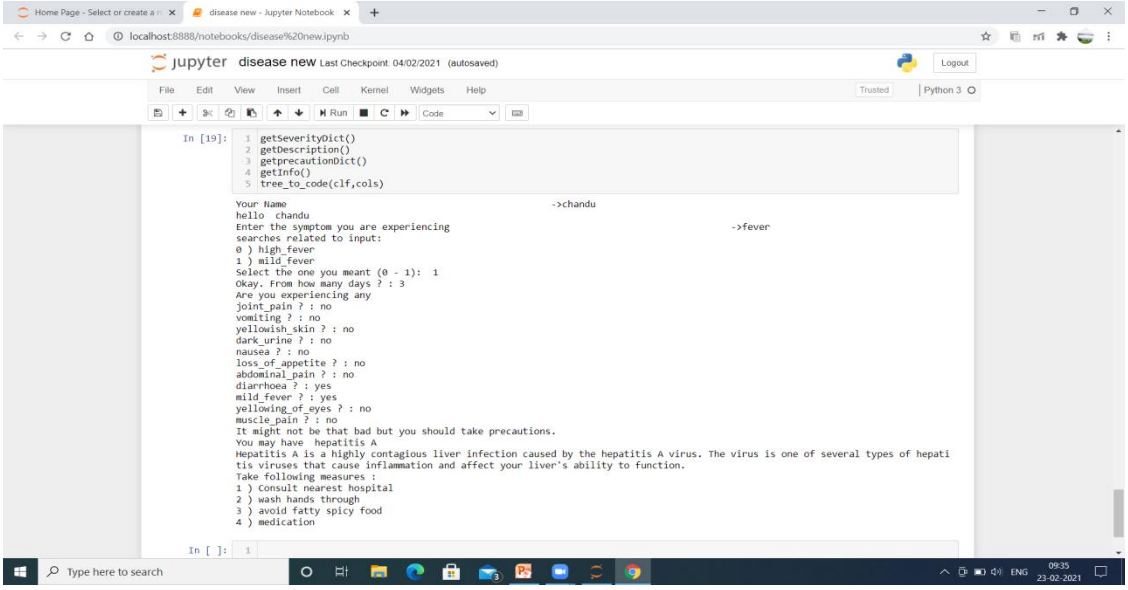
Accuracy Score 93.0

Fig B.7 Interface with inputs



Do you want to exit the system

Fig B.8 Exit System


Home Page - Select or create a n
disease new - Jupyter Notebook
localhost:8888/notebooks/disease%20new.ipynb
jupyter
disease new Last Checkpoint: 04/02/2021 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Type here to search
09:35
23-02-2021

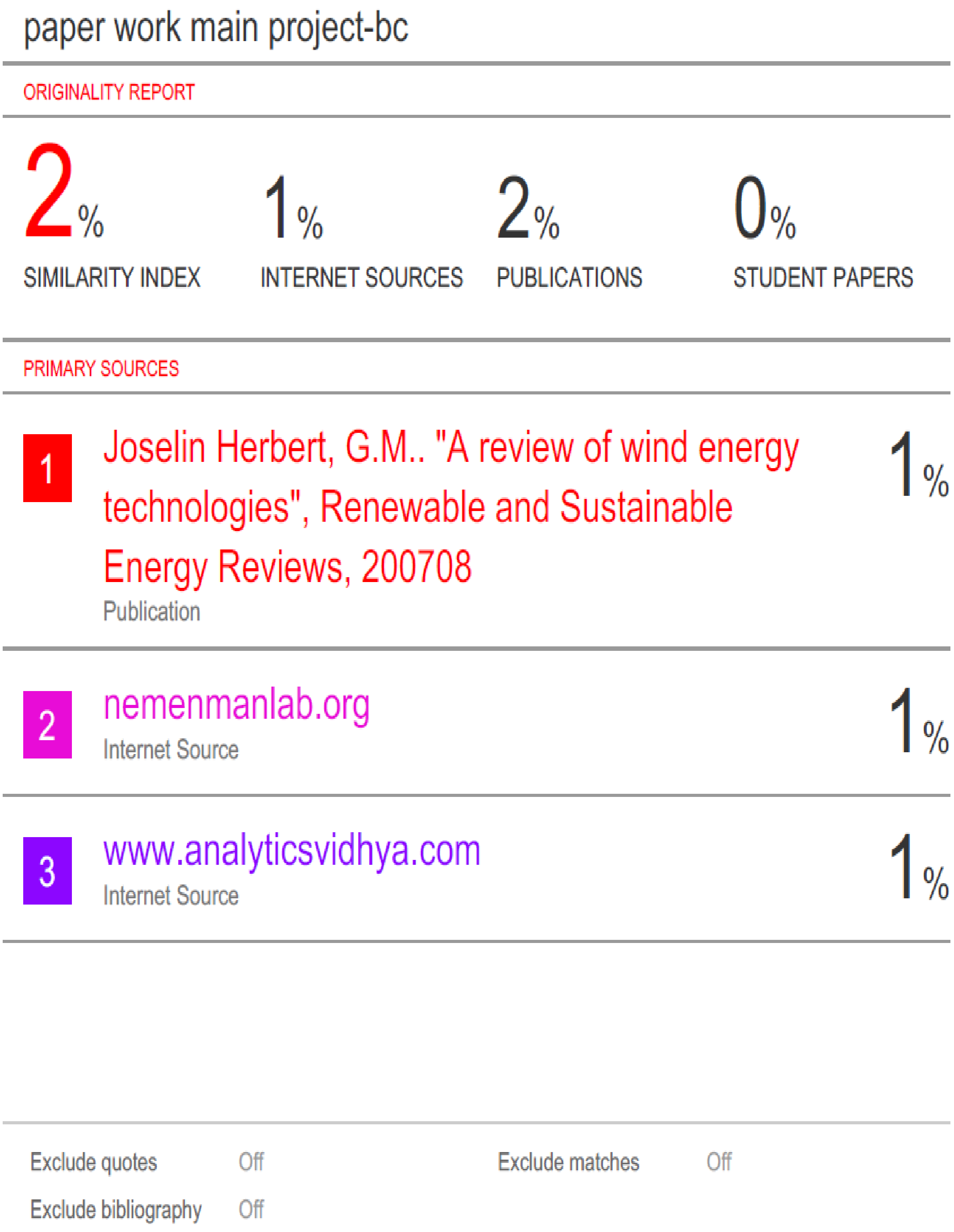
```
In [19]:
getSeverityDict()
getDescription()
getprecautionDict()
getInfo()
tree_to_code(clf,cols)
```

```
Your Name                                                     ->chandu
hello  chandu
Enter the symptom you are experiencing                                      ->fever
searches related to input:
0 ) high_fever
1 ) mild_fever
Select the one you meant (0 - 1):  1
Okay. From how many days ? : 3
Are you experiencing any
joint_pain ? : no
vomiting ? : no
yellowish_skin ? : no
dark_urine ? : no
nausea ? : no
loss_of_appetite ? : no
abdominal_pain ? : no
diarrhoea ? : yes
mild_fever ? : yes
yellowing_of_eyes ? : no
muscle_pain ? : no
It might not be that bad but you should take precautions.
You may have  hepatitis A
Hepatitis A is a highly contagious liver infection caused by the hepatitis A virus. The virus is one of several types of hepati
tis viruses that cause inflammation and affect your liver's ability to function.
Take following measures :
1 ) Consult nearest hospital
2 ) wash hands through
3 ) avoid fatty spicy food
4 ) medication
```

```
In [ ]:
```

Fig B.9 Output of Chatbot

C) PLAGIARISM REPORT



paper work main project-bc

ORIGINALITY REPORT

2%	1%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Joselin Herbert, G.M.. "A review of wind energy technologies", Renewable and Sustainable Energy Reviews, 200708 Publication	1%
2	nemenmanlab.org Internet Source	1%
3	www.analyticsvidhya.com Internet Source	1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

D) JOURNAL PAPER

DISEASE PREDICTION USING NAIVE BAYES CLASSIFIER

Roop Chandrika Mallela[1]
[1]Undergraduate Student, CSE Department, Sathyabama Institute of Science and Technology, Chennai. India.
[1]mrchandrika2000@gmail.com,

Reddy Lakshmi Bhavani[2]
[2] Undergraduate Student, CSE Department, Sathyabama Institute of Science and Technology, Chennai. India.
[2]reddybhavani4985@gmail.com,

Dr. B. Ankayarkanni[3]
[3]Assistant Professor, CSE Department, Sathyabama Institute of Science and Technology,Chennai. India.
[3]ankayarkanni.s@gmail.com

Abstract- Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one is willing to go to hospital for health check-up due to the fear of spreading virus. In this situation, technology plays and important role. The domain we used here is Machine Learning, it is the technique by which machines can learn from past experiences like a human being and make it efficient in future. ML is the domain which is widely used nowadays and it is the most efficient domain in health care. We will develop a GUI to get the symptoms from the user. The models used in this paper are Naive Bayes and Decision Tree. The output is the disease, the accuracy of model, its definition and the treatment of the particular disease based on the symptoms given by the individual. As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease". This paper shows detailed explanation of how to find the diseases from symptoms, so that the individual can contact the respective doctor and stay healthy at an early stage.

***Keywords:* Disease Prediction, Machine Learning, Tkinter, Chatbot, Symptoms.**

I. INTRODUCTION

A disease is a condition that affects the individual functioning of body totally. Diseases if neglected will lead to the death of an individual. Diseases can be identified by the symptoms of the body of an individual.

Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus.

As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease".

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

Some approaches tried to do predict the disease but they are restricted to particular disease like liver disease, heart disease, diabetes and so on. Machine Learning is the domain used in this paper. Machine Learning is a technique by which machines will be capable of learning and improving from past experiences using the algorithms.

The dataset which is used in this paper contains symptoms and the particular disease which was processed in Gaussian Naive Bayes and Decision Tree, which are the most easy models for disease prediction. While processing the data, symptoms are given as input and the disease was received as an output.

Machine learning technique makes easier to predict the disease and get treatment based on the disease. Various algorithms are implemented on the dataset and the accuracy of the same is also calculated. By this, we will consider only the algorithms which has highest accuracy and it is used for accurate disease prediction.

Developing a project based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method is the main objective of the project.

We have designed a disease prediction system using ML algorithm (Naive Bayes), find the most accurate algorithm, and used it to find the disease and Tkinter for GUI.

And also, we have created a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

This project helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor.

A disease predictor can also be called as a virtual doctor, which can predict the disease based on symptoms. Recently due to covid-19 no one are willing to go outside. This disease predictor system can be a most useful as it identifies the disease without even contacting the individual.

In this way machine learning when implemented in healthcare can help in satisfying the individual and also take care of their particular disease easily.

Naïve Bayes, which is the most easy model helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor. This system not only reduces expense of the individual but also gives the accurate prediction.

II.LITERATURE SURVEY

- Tarigoppula V.S Sriram et al. in [1] collects the voice dataset from UCI Machine Learning repository and train four algorithms on that dataset. The result is the prediction of Parkinson Disease by considering the most accurate algorithm.
- Shubham Bind et al. in [2] studies about all the available researches in literature to predict the Parkinson diseases.
- K. Gomathi, D. Shanmuga Priya in [3] used different data mining techniques to predict Heart disease, Breast Cancer, Diabetes. The models used are Decision Tree and Naive Bayes Classifier. Performance of both the models was compared and the best classifier is used to predict the above diseases.
- Isha Pandya et al. in [4] used two supervised machine learning algorithms Decision Tree, accuracy 91% and Naïve Bayes classifier, accuracy 87%. Here, they used the combination of both to get the best accuracy. Naïve Bayes Classifier accuracy should be improved.
- Akash C. Jamgade, Prof. S. D. Zade in [5] paper determined the most danger diseases which occur in a person in a locality and community. But, the data collection is difficult.
- Siddhika Arunachalam in [6] six classification algorithms are used after analyzing 14 attributes in the dataset. But, we may get confused which algorithm to use.
- Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE methodologies in [7] used are SVM, ANN, Logistic Regression so on. The predicted diseases are breast cancer, lung cancer, heart diseases, diabetes, thyroid or kidney diseases.
- H BENJAMIN FREDRICK DAVID in [8] predicted the occurrence of heart disease using ensemble learning algorithms. Accuracy of both Decision Tree and Naïve Bayes and analysed.
- Harshit Anand et al. in [9], domains of Machine learning and Data Science are used and models are built using numpy, pandas, sklearn, and so on and the model are deployed using Django.
- Goutam Chakraboty et al. in [10], takes into account six features from 23 features in the dataset and predict the risk of chronic kidney disease. It is used to reduce the impact of Chronic Kidney disease (CKD), where creatinine test is not available for all.
- A. Durga Praveen et al. in [11], applied 5 models namely KNN, SVM, Random Forest, Naïve Bayes and Adaboost and found that KNN, Adaboost has the highest accuracy of all the models. So, any of these two are used for prediction and prevention of the liver disease.
- Sejin Park et al. in [12], early prediction of disease using the previous real-time stroke symptoms. It is implemented at a low cost. Random Forest algorithm is used for validating clinical significance.
- Ahan Chatterjee et al. in [13], they have used machine learning models like Decision tree, SVM, Random Forest, and so on, find the best classifier using the accuracy and use it to predict cancer disease risk in the early stage and prevent it. Simulation model is also designed to manage the patient flow in OPDs.
- Sergio Grueso et al. in [14], they have used dataset taken from ADNI database and selected 47 out of 159 studies for analysis. Deep learning combined with multimodal and multidimensional data is used to achieve the best performance.
- Upendra Kumar in [15], have used a technique which is a computer screening tool. It checks for disease of an individual and gives the efficient and correct treatment in a fraction of seconds.

(A) Existing System

From the above literature survey, We have inferred that all the systems existing predict only particular diseases namely lung disease, breast cancer, heart disease, diabetes by implementing various algorithms on the particular datasets. After implementing various algorithms, the most accurate one is selected and it is used for prediction of disease. Sometimes, we may get confused of what algorithm to use. Also, all the systems find only the particular disease and not the disease based on the symptoms.

(B) Proposed System

We are proposing a system, which uses tkinter for GUI interface. It is a simple user Interface and also time efficient. Our aim with is to get the disease based on symptoms given by the user. The domain we will use is the machine learning, in that we will be using Naïve Bayes Classifier, which will help us in getting the most accurate predictions easily and also the accuracy is given as output. To reduce time consuming, we will ask only less questions namely the name of the individual and the symptoms the individual is facing. In this way, our system will be less time consuming and give accurate predictions. And also, we are creating a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

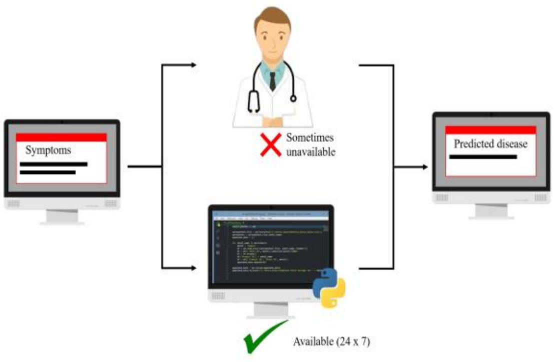


Fig 1 Proposed system for disease prediction. The doctor may not be available always when needed. By using this system we can predict the disease based on our symptoms anytime.

III ARCHITECTURE OF THE SYSTEM

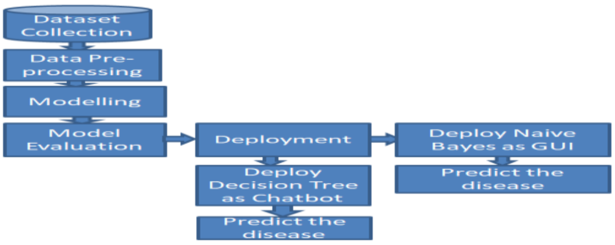


Fig 2 Architecture of the system

IV METHODOLOGY

The modules used in this project are-

- ✓ **Tkinter**- Tkinter is library of python used oftenly by everyone. It is a library which is useed to create GUI based applications easily. It contains so many widgets like radiobutton, textfiled and so on. We have used this for creating account registration screen, login or register screen, prediction interface which is a GUI based application

- ✓ **Sklearn**- Scikit Learn also known as sklearn is a open source library for python programming used for implementing machine learning algorithms. It features various classification, clustering, regression machine learning algorithms. In this it is used for importing machine learning models, get accuracy, get confusion matrix.

- ✓ **Pandas-** Library of python which can be used easily. It gives speed results and also easily understandable. It is a library which can be used without any cost. We have used it for data analysis and to read the dataset.

- ✓ **Matplotlib-** Library of python used for visualising the data using graphs, scatterplots and so on. Here, we have used it for data visualisation.

- ✓ **Numpy-** Library of python used for arrays computation. It has so many functions. We have used this module to change 2-dimensional array into contiguous flattened array by using ravel function.

- ✓ **Pandas Profiling**-This is library of python which can be used by anyone free of cost. It is used for data analysis. We have used this for getting the report of the dataset.

Naive Bayes Classifier

Naïve Bayes, a supervised machine learning algorithm used for classifying problems which uses Bayes theorem. Naïve Bayes is mainly used for doing the text classification. The Bayes theorem can be defined by:

$$P\,(y|x) = \frac{P\,(x|y)\,.\,P\,(y)}{P\,(x)}$$

P (y|x) is the chance of assumption y for given data x This is also called chance or possibilty of updation.

P(x|y) is the chances of data x given the assumption y was true.

P(y) is the possibility of assumption y being true. This is called the initial possibility or chance of y .

P(x) is the possibility and evidence of data and is called marginal possibility or chance.

Gaussian Naive Bayes

It follows the same procedure as the Naive Bayes. But for Naive Bayes we need a categorical dataset and for Gaussian Naive Bayes we need a dataset that has all the continuous features.

Decision Tree

Decision Tree an algorithm whose input and output or known. The information is divided repeatedly using the particular parameter. Decision tree consists of two main parts namely decision node and leaf nodes. The decision nodes specify the decision at which the parameter should be spilt. The leaf nodes are the output bought by the decisions. A decision tree asks for either true or false to divide the data.

Project contains three parts:

1. DATASET COLLECTION.
2. TRAIN AND TEST THE MODEL.
3. DEPLOY THE MODEL USING TKINTER.

- ✓ **Dataset Collection-** We had collected dataset from kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 4920 rows and 133 columns.
- ✓ **Train and Test the model-** We had used the Naïve Bayes Classifier as a model to train the dataset. After training, we had tested the model and found its accuracy.
- ✓ **Deploy the model using Tkinter-** Deployed Naïve bayes by creating interface to get the name, symptoms of an individual. By this, we will get the disease and accuracy of model as the output. We have also created a chatbot using Decision Tree which helps an individual to get the corresponding disease by checking whether he/she is being faced by the symptoms.

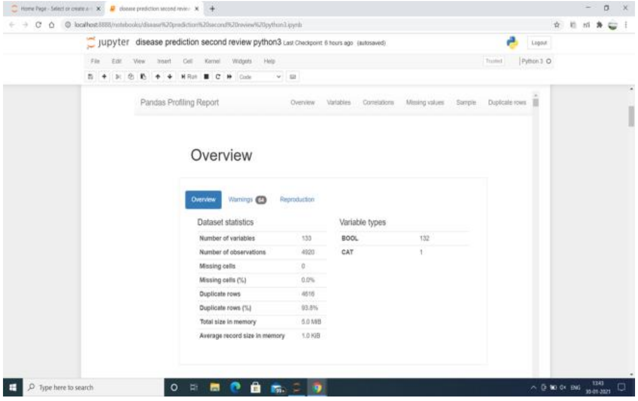
V RESULTS

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where essential resources are unavailable and people are also not willing to go outside in fear of spreading virus.

In this situation, our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease. This also helps in reduction of the cost and give the correct and fast result.

Algorithm	Accuracy
Decision Tree	97%
Naive Bayes	100%

Accuracies Table


jupyter disease prediction second review python3 Last Checkpoint 6 hours ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Pandas Profiling Report
Overview Variables Correlations Missing values Sample Duplicate rows
Overview
Overview Warnings Reproduction
Type here to search

Dataset statistics

Number of variables	133
Number of observations	4920
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	4616
Duplicate rows (%)	93.8%
Total size in memory	5.0 MiB
Average record size in memory	1.0 KiB

Variable types

BOOL	132
CAT	1

Fig Report of dataset

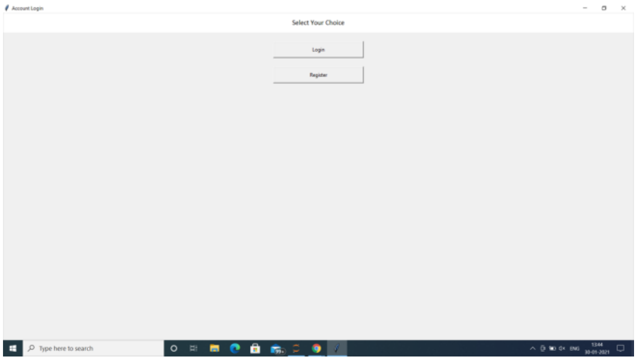


Fig 4 Account Registration Screen

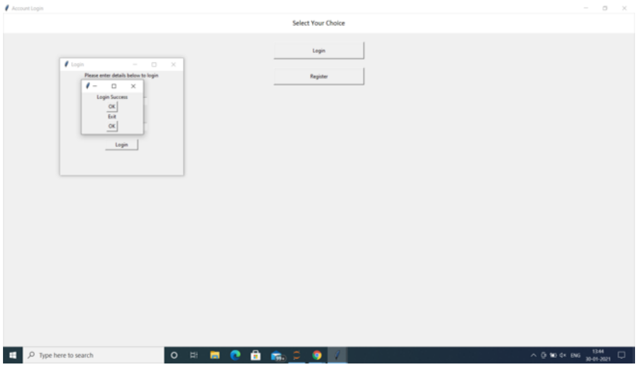
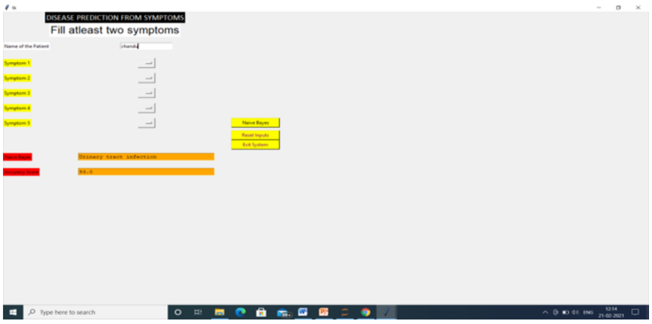


Fig 5 Login Successful Screen

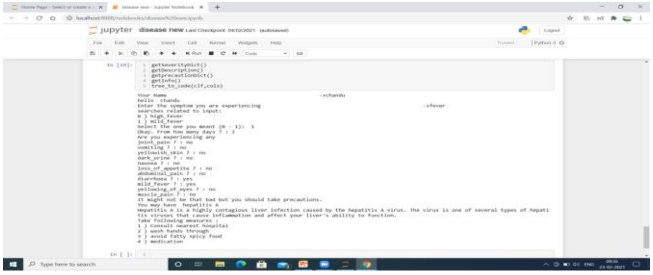


Name of the Patient: chandu

Naive Bayes: Urinary tract infection

Accuracy Score: 93.0

Fig 6 GUI Interface


jupyter disease new Last Checkpoint: 04/02/2021 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Logout
Trusted
Python 3

```
In [10]:  getSeverityDict()
          getDescription()
          getprecautionDict()
          getInfo()
          tree_to_code(clf,cols)

Your Name                                         ->chandu
hello  chandu
Enter the symptom you are experiencing                          ->fever
searches related to input:
0 ) high_fever
1 ) mild_fever
Select the one you meant (0 - 1):  1
Okay. From how many days ? : 3
Are you experiencing any
joint_pain ? : no
vomiting ? : no
yellowish_skin ? : no
dark_urine ? : no
nausea ? : no
loss_of_appetite ? : no
abdominal_pain ? : no
diarrhoea ? : yes
mild_fever ? : yes
yellowing_of_eyes ? : no
muscle_pain ? : no
It might not be that bad but you should take precautions.
You may have  hepatitis A
Hepatitis A is a highly contagious liver infection caused by the hepatitis A virus. The virus is one of several types of hepati
tis viruses that cause inflammation and affect your liver's ability to function.
Take following measures :
1 ) Consult nearest hospital
2 ) wash hands through
3 ) avoid fatty spicy food
4 ) medication
```

Fig 7 Chatbot

VI CONCLUSION

The project presented the technique of predicting the disease based on the symptoms of an individual patient. Once the disease is predicted, we could easily manage the medicine resources required for the treatment. Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease. Our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease. This also helps in reduction of the cost and give the correct and fast result.

REFERENCES

[1] Tarigoppula V.S Sriram et al, "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013

[2]Shubham Bind et al, "A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction" International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1648-1655.

[3]K. Gomathi, D. Shanmuga Priya, "Multi Disease Prediction using Data Mining Techniques", 2016.

[4]Isha Pandya et al, "Prediction of Heart Disease Using Machine Learning Algorithms", 2018.

[5] Akash C. Jamgade, Prof. S. D. Zade,"Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology Volume: 06 Issue: 05 May 2019.

[6] Siddhika Arunachalam," Cardiovascular Disease Prediction Model using Machine Learning Algorithms", International Journal for Research in Applied Science & Engineering Technology ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.

[7] Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE, "A REVIEW OF DATA MINING TECHNIQUES IN MEDICINE", JOURNAL OF INFORMATION SYSTEMS & OPERATIONS MANAGEMENT, Vol. 14.1, May 2020.

[8] H BENJAMIN FREDRICK DAVID, "IMPACT OF ENSEMBLE LEARNING ALGORITHMS TOWARDS ACCURATE HEART DISEASE PREDICTION", DOI: 10.21917/ijsc.2020.0296.

[9] Harshit Anand et al. "Hridaya Kalp: A Prototype for Second Generation Chronic Heart Disease Detection and Classification" ,31 July,2020.

[10] Goutam Chakraboty et al. "Predicting the Risk of Chronic Kidney Disease using Machine Learning Algorithm", 11(1), 28 December, 2020, 202.

[11] A. Durga Praveen et al. "Intelligent Liver Disease Prediction system using Machine Learning Models" , vol 702. Springer, Singapore. 5 Jan, 2021.

[12] Sejin Park et al. “Machine-Learning-Based Elderly Stroke Monitoring System Using Electroencephalography Vital Signals”, 2021, *11*(4), 1761.

[13] Ahan Chatterjee et al. “A Machine Learning Approach to prevent cancer”, DOI: 10.4018/978-1-7998-2742-9.ch007, 2021.

[14] Sergio Grueso et al. “Machine Learning methods for predicting conversion from mild cognitive impairment to Alzheimer’s disease”, 2021.

[15] Upendra Kumar, “Applications of Machine Learning In Disease Pre-Screening”, 10.4018/978-1-7998-7705-9.ch049, 2