

DISEASE PREDICTION

A PROJECT REPORT

Submitted by

Group 58

Jigyasa Shukla (21BAI10160)
Purva Samar (21BAI10356)
Rikku Prasad (21BAI10488)
Vaibhav Anand (21BAI10136)
Swarnankita Saha(21BAI10436)

*in partial fulfillment for the award of the degree
of*

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING



SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

**KOTRIKALAN, SEHORE
MADHYA PRADESH - 466114**

OCT 2022

BONAFIDE CERTIFICATE

Certified that this project report titled “**DISEASE PREDICTION**” is the bonafide work of “**Jigyasa Shukla (21BAI10160), Purva Samar (21BAI10356), Rikku Prasad (21BAI10488), Vaibhav Anand (21BAI10136), Swarnankita Saha (21BAI10436)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR

<<Dr. Suthir Sriram>>,
<<Program Chair-Division of AIML>>
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE

<<Dr. Anil Kumar Yadav>>,
<< Assistant Professor Grade 2>>
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on – 29/07/22

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr. Suthir Sriram, Head of the Department, School of Computing Science & Engineering for much of his valuable support encouragement in carrying out this work.

I would like to thank my internal guide Dr. Anil Kumar Yadav for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computing Science & Engineering+, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	List of Abbreviations	7
	List of Figures and Graphs	8
	List of Tables	9
	Abstract	10
1	<p style="text-align: center;">CHAPTER-1:</p> <p style="text-align: center;">PROJECT DESCRIPTION AND OUTLINE</p> <ol style="list-style-type: none"> 1. Introduction 2. Motivation for the work 3. [About Introduction to the project including techniques] 4. Problem Statement 5. Objective of the work 6. Organization of the project 7. Summary 	<p>11</p> <p>11</p> <p>11</p> <p>11</p> <p>12</p> <p>12</p> <p>12</p> <p>13</p> <p>13</p>
2	<p style="text-align: center;">CHAPTER-2:</p> <p style="text-align: center;">RELATED WORK INVESTIGATION</p> <ol style="list-style-type: none"> 1. Introduction 2. Machine Learning 3. Existing Approaches/Methods <ol style="list-style-type: none"> 1. Approaches/Methods -1 2. Approaches/Methods -2 3. Approaches/Methods -3 4. Summary 	<p>14</p> <p>14</p> <p>14</p> <p>14</p> <p>15</p> <p>15</p> <p>15</p>
3	<p style="text-align: center;">CHAPTER-3:</p> <p style="text-align: center;">REQUIREMENT ARTIFACTS</p> <ol style="list-style-type: none"> 1. Introduction 2. Hardware and Software requirements 	<p>16</p> <p>16</p>

	3. Specific Project requirements 1. Data requirement 2. Functions requirement 4. Summary	16 16-17 18
4	CHAPTER-4: DESIGN METHODOLOGY AND ITS NOVELTY 1. Methodology and goal 2. Functional modules design and analysis 3. Software Architectural designs 4. User Interface designs 5. Summary	19 19 19 20 21 23 24
5	CHAPTER-5: TECHNICAL IMPLEMENTATION & ANALYSIS 1. Technical coding and code solutions 2. Summary	25 26 27
6	CHAPTER-6: PROJECT OUTCOME AND APPLICABILITY 1. Outline 2. Significant project outcomes 3. Project applicability on Real-world applications 4. Inference	28 28 28 29
7	CHAPTER-7: CONCLUSIONS AND RECOMMENDATION 1. Outline 2. Limitation/Constraints of the System	29 30 31 31

	3. Future Enhancements 4. Inference	
	References	32

LIST OF ABBREVIATIONS

1. RAM-RANDOM ACCESS MEMORY
2. MB-MEGABYTE
3. GB-GIGABYTE
4. KNN ALGORITHM-K NEAREST NEIGHBOUR
5. CNN-CONVOLUTIONAL NEURAL NETWORK
6. MDRP-MULTIMODAL DISEASE RISK PREDICTION
7. SVM- SUPPORT VECTOR CLASSIFIER

LIST OF FIGURES AND GRAPHS

TABLE NUMBER	TITLE	PAGE NUMBER
1	ARCHITECTURE DIAGRAM	21
2	PROCESS FLOW	22

LIST OF TABLES

TABLE NUMBER	TITLE	PAGE NUMBER
1	HARDWARE REQUIREMENT	16
2	SOFTWARE REQUIREMNT	16

ABSTRACT

In modern day it has become a necessity to detect diseases as early as possible as it helps supply better patient care. Technologies like artificial intelligence and machine learning have made it easier for doctors to pinpoint diseases and make correct decisions about a patient's health and treatment based on vast amounts of data available. When our body starts showing symptoms that something is wrong within our body, we must travel to the doctor to get an exact medical diagnosis. Sometimes it's a minor problem and sometimes it can be life threatening illness if we do not investigate it and take care of it in the early stages. Therefore, to save the time needed for complete diagnosis we are proposing this disease prediction system that can predict the possible diseases a patient might have based on symptoms provided. In this project we will be using machine learning algorithms like Random Forest Classifier, Naïve Bayes Theorem, Support Vector Classifier & K-Fold Cross Validation Algorithm. The results we have gotten from this proposed system have an accuracy up to 84%. It has tremendous potential to predict the possible diseases more precisely. Our main aim is to make the work of doctors easier and help the first-year doctors to make correct decisions.

CHAPTER 1

PROJECT DESCRIPTION AND OUTLINE

1.1 Introduction

All technology users today benefit from machine learning. Machine learning is a subfield of artificial intelligence. In general, the goal of machine learning is to understand the structure of data and fit that data into models that humans can understand and use. Hypothesis uses ML algorithms to allow doctors to use supervised learning as a powerful tool to diagnose diseases more effectively.

With the number of cases and illnesses rising each year, many countries' healthcare systems have become overloaded and expensive over time. With enough data, disease prediction becomes very easy and cheap. Observing symptoms to predict disease is now an integral part of treatment. The intention is to derive efficient, accurate and satisfactory machine learning algorithms for disease prediction.

A machine learning algorithm consists of two phases: 1) training and 2) testing. Apply full machine learning concepts to keep your patients healthy. This project uses supervised machine learning concepts to predict disease. A key feature is machine learning, using algorithms such as decision trees, random forests, naïve Bayes, and ANNs to help predict early disease and improve patient care.

1.2 Motivation for the Project

The main motivation for our project and the need for this prediction system is less availability of health centers. The distance between patients and health centers during the time of need is a great motivator. The main factor that drove us forward to complete this project was to reduce the load on doctors in this postcovid era when need of doctors has increased as people require consultation for every minute inconvenience. We are motivated to propose a system that does not waste the valuable time of patients as well as doctors.

1.3 About Introduction to the project Including Techniques

In our proposed system we will be using techniques and algorithms from machine learning like Support Vector Classifier, Naïve Bayes Classifier, Random Forest Classifier & K – Fold Cross validation.

K- Fold Cross-validation is a statistical method used to estimate the skill of machine learning models.

Naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation

Random forest algorithm is one such algorithm used for machine learning. It is used to train the data based on the previously fed data and predict the possible outcome for the future.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points

1.4 Problem Statement

To propose a system to predict a list of probable diseases. It also aims to produce more accurate results and reduce workload of doctors.

The conventional approach requires patients a visit to doctor and making appointments to get a diagnosis. In most cases it leads to multiple tests in order to get accurate disease indications. Sometimes patients have to wait for test results like blood reports. This process takes up precious time of patients as well as doctors.

Through this project we aim to propose a disease prediction system which depends on the input from the user to bring forth probable diagnosis.

When we are not feeling well the first thing we do is to check our temperature to get an estimate or baseline idea of our fever so we can consult our doctor if the temperature is high enough similarly a medical disease prediction application can be used to get a baseline idea of disease and can indicate us whether we should take immediate doctor consultation or not, or at least start some home- remedies for the same to find temporary relief.

1.5 Objective of the work

Our main aim is to provide a quick medical diagnosis to the patients. The current situation is that if a patient exhibits any symptoms, he or she must visit a doctor or a hospital to have the ailment diagnosed.

But our primary goal is to lessen the time and energy patients spend only on diagnosis the illness. Many individuals are dying merely because of their disease's tardy diagnosis. So, our primary goal is to decrease these facilities.

1.6 Organization of the project

The purpose of making this project called "Disease Prediction Using Machine Learning" is to predict a nearly accurate disease of the patient using all their general information's and also the symptoms.

Using this information, there we will compare with our previous datasets of the patients and predicts the disease of the patient he/she is been through, If this Prediction is done at the early stages of the disease with the help of this project and all other necessary measure the disease can be cured and in general this prediction system can also be very useful in health industry.

1.7 Summary

This illness prediction system's primary objective is to make disease predictions based on symptoms. This method captures the user's symptoms, which he or she may have as an input, producing a final product as a diagnosis of illness. Infectious Disease Predictor Grails framework-based implementations that were effective. This system provides an atmosphere that is comfortable and simple to utilize. The user can access the system via a web application since it can access this system at any time and from any location. In conclusion, for disease risk modelling the accuracy depends upon hospital data.

CHAPTER 2

RELATED WORK INVESTIGATION

2.1 Introduction

The intent is to deduce a satisfactory Machine Learning algorithm which is efficient and accurate for the prediction of disease. In this paper, the supervised Machine Learning concept is used for predicting the diseases. The main feature will be Machine Learning in which we will be using algorithms such as Decision Tree, Random Forest, Naïve Bayes and KNN which will help in early prediction of diseases accurately and better patient care .

2.2 Machine Learning

All technology users today benefit from machine learning. Machine learning is a subfield of artificial intelligence. In general, the goal of machine learning is to understand the structure of data and fit that data into models that humans can understand and use. Hypothesis uses ML algorithms to allow doctors to use supervised learning as a powerful tool to diagnose diseases more effectively. With the number of cases and illnesses rising each year, many countries' healthcare systems have become overloaded and expensive over time. With enough data, disease prediction becomes very easy and cheap. Observing symptoms to predict disease is now an integral part of treatment. The intention is to derive efficient, accurate and satisfactory machine learning algorithms for disease prediction. A machine learning algorithm consists of two phases: 1) training and 2) testing. Apply full machine learning concepts to keep your patients healthy. This project uses supervised machine learning concepts to predict disease. A key feature is machine learning, using algorithms such as decision trees, random forests, naive Bayes, and ANNs to help predict early disease and improve patient care.

2.3 Existing Approaches/Methods

2.3.1 Approaches/Methods -1

There is numerous work that has been done related to disease prediction system using different Machine Learning algorithms and achieved different results for different methods in medical field. The paper [1] “Disease Prediction Using Data Mining Techniques” used KNN, Naïve Bayes and SVM algorithms and collated with respect to the accuracy using heart disease dataset and achieved the highest accuracy of 86.6% using Naïve Bayes.

PROS-

It showed that Naïve Bayes algorithm provided the highest accuracy.

CONS-

KNN algorithm does not work well with large datasets. The cost of calculating the distance between the new point and each existing point is huge, which degrades performance.

2.3.2 Approaches/Methods -2

The paper [2] "Application of Machine Learning Predictive Models in the Chronic Disease " focused on SVM and LR algorithms and evaluate the study models associated with diagnosis of chronic disease. These models are highly applicable in classification and diagnosis of CD.

PROS- SVM algorithm is better than some other algorithms, like K-Nearest Neighbours, is because it chooses the best line to classify your data points.

CONS-

Linear Regression deals with continuous values whereas classification problems mandate discrete values hence it makes it difficult to use LR for classification.

2.2.3 Approaches/Methods -3

The paper [3] "Disease Prediction Using Machine Learning over Big Data" has proposed a CNN-MDRP algorithm which combines structured and unstructured data and proved that CNN-MDRP is more accurate than previous prediction algorithm.

PROS-

Automatically detects important features without human supervision.

CONS-

A Convolutional neural network is significantly slower due to an operation such as maxpool. If the CNN has several layers, then the training process takes a lot of time if the computer doesn't consist of a good GPU.

2.4 Summary

We are proposing such a system that will flaunt a simple, cost effective, elegant User Interface and also be time efficient. This system is used to predict diseases according to symptoms. In this proposed system we are going to take down five symptoms from the users and evaluate them by applying algorithms such as Decision Tree, Random Forest, Naïve bayes and KNN which will help in getting accurate prediction.

CHAPTER 3

REQUIREMENT ARTIFACTS

3.1 Hardware and Software Requirement

PROCESSOR-11th Gen Intel(R) Core (TM) i7 (H)

GRAPHICS PROCESSING UNIT(GPU)-Intel(R) Iris(R) Xe Graphics, 1024 MB (H)

MEMORY-128GB (H)

OPERATING SYSTEM-Microsoft Windows 11 Home Single Language 64-bit (S)

PROGRAMMING LANGUAGE-PYTHON (S)

DEEP LEARNING LIBRARY-KAGGLE DATASET (S)

3.2 Specific Project Requirements

3.2.1 Data requirement

We will be using dataset from Kaggle for this problem. This dataset will include two files one for training and the other for testing. This dataset could be an information database of disease symptom associations generated by discharge summaries of various hospitals and doctors.

3.2.2 Functions requirement

We will be requiring the following algorithms in our project-

Random Forest Classifier-Random Forest is an ensemble learning-based supervised machine learning classification algorithm that internally uses multiple decision trees to make the classification.

Gaussian Naive Bayes Classifier: It is a probabilistic machine learning algorithm that internally uses Bayes Theorem to classify the data points. Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts based on the probability of an object.

CONS-

Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

PROS-

It is used for Credit Scoring.

It is used in medical data classification.

It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.

It is used in Text classification such as Spam filtering and Sentiment analysis.

Fold cross-validation is one of the cross-validation techniques in which the whole dataset is split into k number of subsets, also known as folds, then training of the model is performed on the k-1 subsets and the remaining one subset is used to evaluate the model performance.

CV is easy to understand, easy to implement, and it tends to have a lower bias than other methods used to count the model's efficiency scores. All this makes cross-validation a powerful tool for selecting the best model for the specific task.

There are a lot of different techniques that may be used to cross-validate a model. Still, all of them have a similar algorithm:

1. Divide the dataset into two parts: one for training, other for testing
2. Train the model on the training set
3. Validate the model on the test set

4. Repeat 1-3 steps a couple of times. This number depends on the CV method that you are using

PROS-

Reduces Overfitting
Hyperparameter Tuning

CONS-

Increases Training Time
Needs Expensive Computation:

Support Vector Classifier is a discriminative classifier i.e., when given a labeled training data, the algorithm tries to find an optimal hyperplane that accurately separates the samples into different categories in hyperspace. Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

PROS-

It works really well with a clear margin of separation. It is effective in high dimensional spaces. ...

CONS-

It doesn't perform well when we have large data set because the required training time is higher.

3.3 Summary

These algorithms need to study in detail so that the program can be executed properly. Detailed study of python is required to make the code that will be able to achieve the output. The knowledge of knowing how to use the algorithms like Naïve Bayes Classifier, Random Forest Classifier etc. within the python program and then using it to achieve the output is mandatory. Deep understanding about what is machine learning how it works and how it is helping us in the modern-day environment is also required.

CHAPTER 4

DESIGN METHODOLOGY AND ITS NOVELTY

4.1 Methodology and goal

Our concept is based on multiple illness prediction based on symptoms supplied by the patient. The first step is to identify the issue statement. The dataset is then prepared for work. Following that, we conceptualize our data using scatter plots, distribution graphs, and so on, to identify anomalies, missing values, and so on in our data and make our dataset suitable for prediction. Finally, the key feature will be Machine Learning, which will use algorithms such as Decision Tree, Random Forest, Naive Bayes, and KNN to predict accurate illness for early prediction and improved patient treatment. We utilized Python as a platform to run our Machine Learning algorithms for this model. We also created an appealing user interface.

4.2 Functional modules design and analysis

GATHERING THE DATA-

Data preparation is the primary step for any machine learning

project. **CLEANING THE DATA-**

Cleaning is the most important step in a machine learning project.

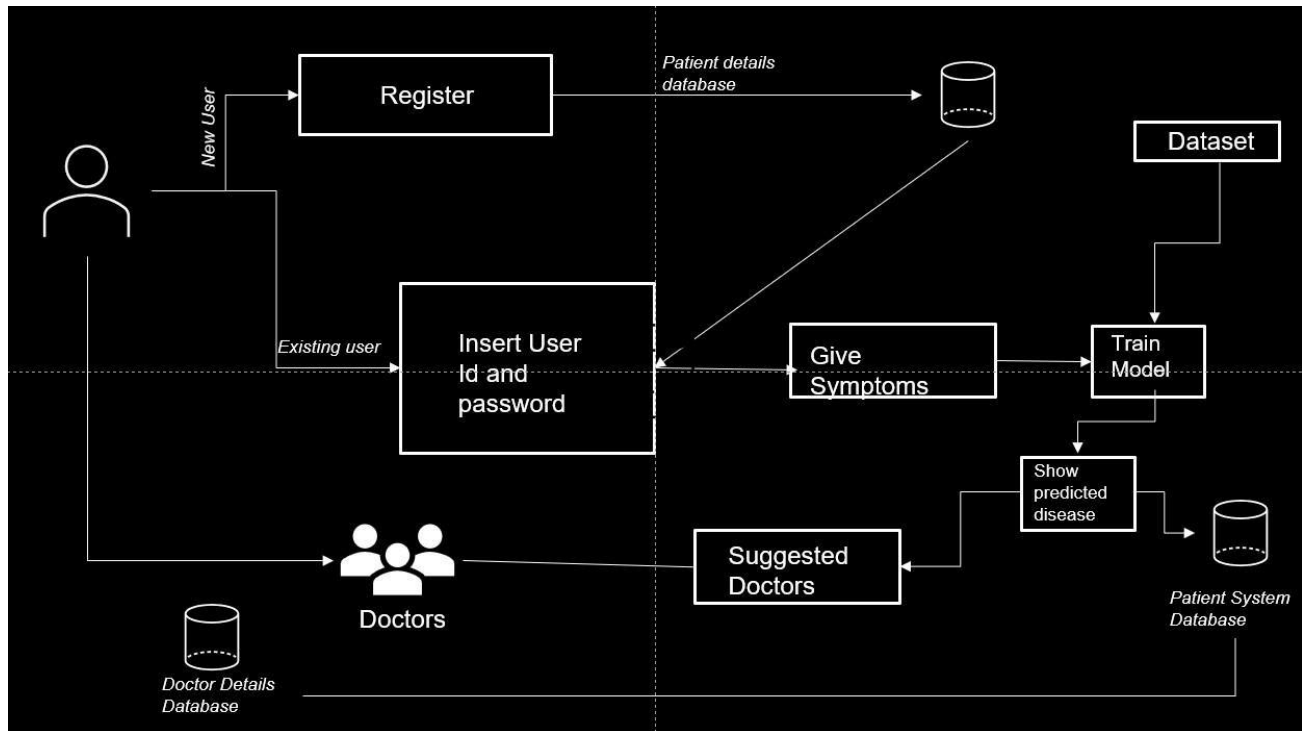
MODEL BUILDING-

After gathering and cleaning the data, the data is ready and can be used to train a machine learning model.

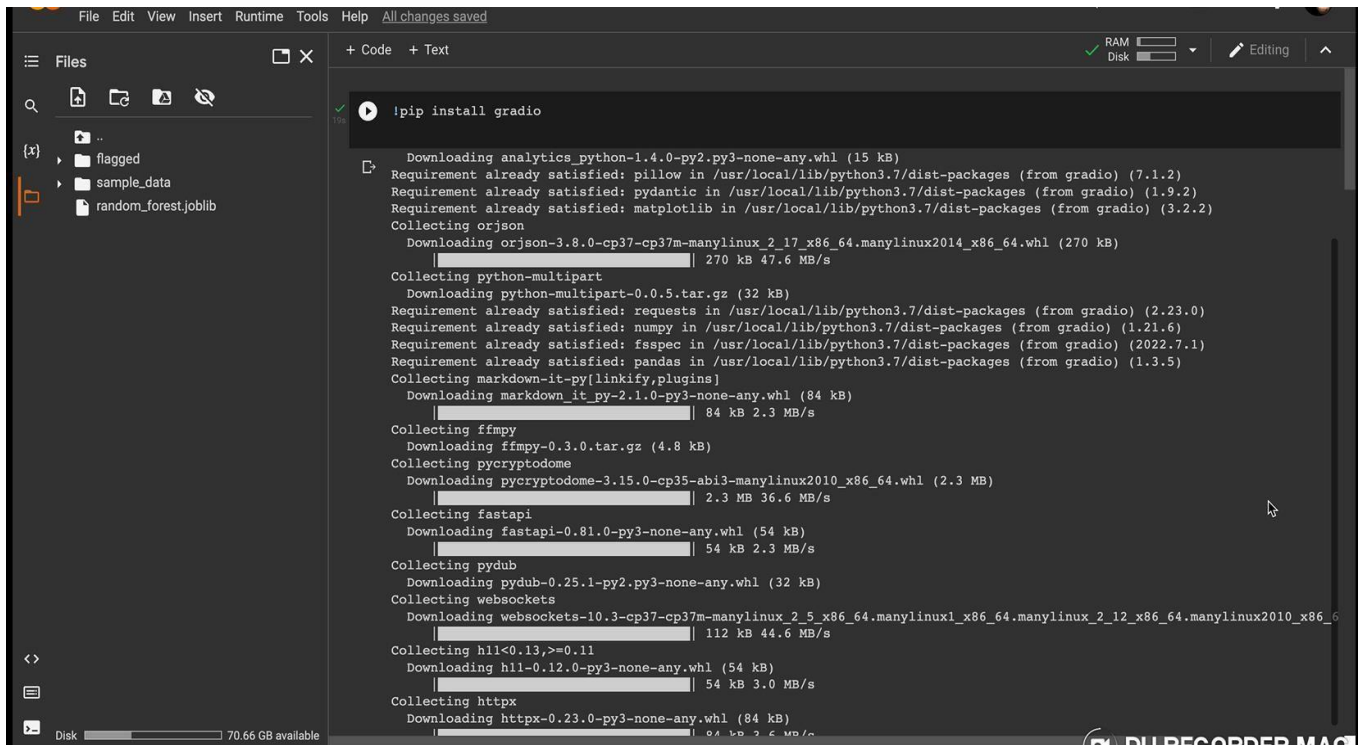
INFERENCE-

After training the three models we will be predicting the disease for the input symptoms by combining the predictions of all three models. This makes our overall prediction more robust and accurate.

4.3 Software architecture designs



4.4 User Interface Designs

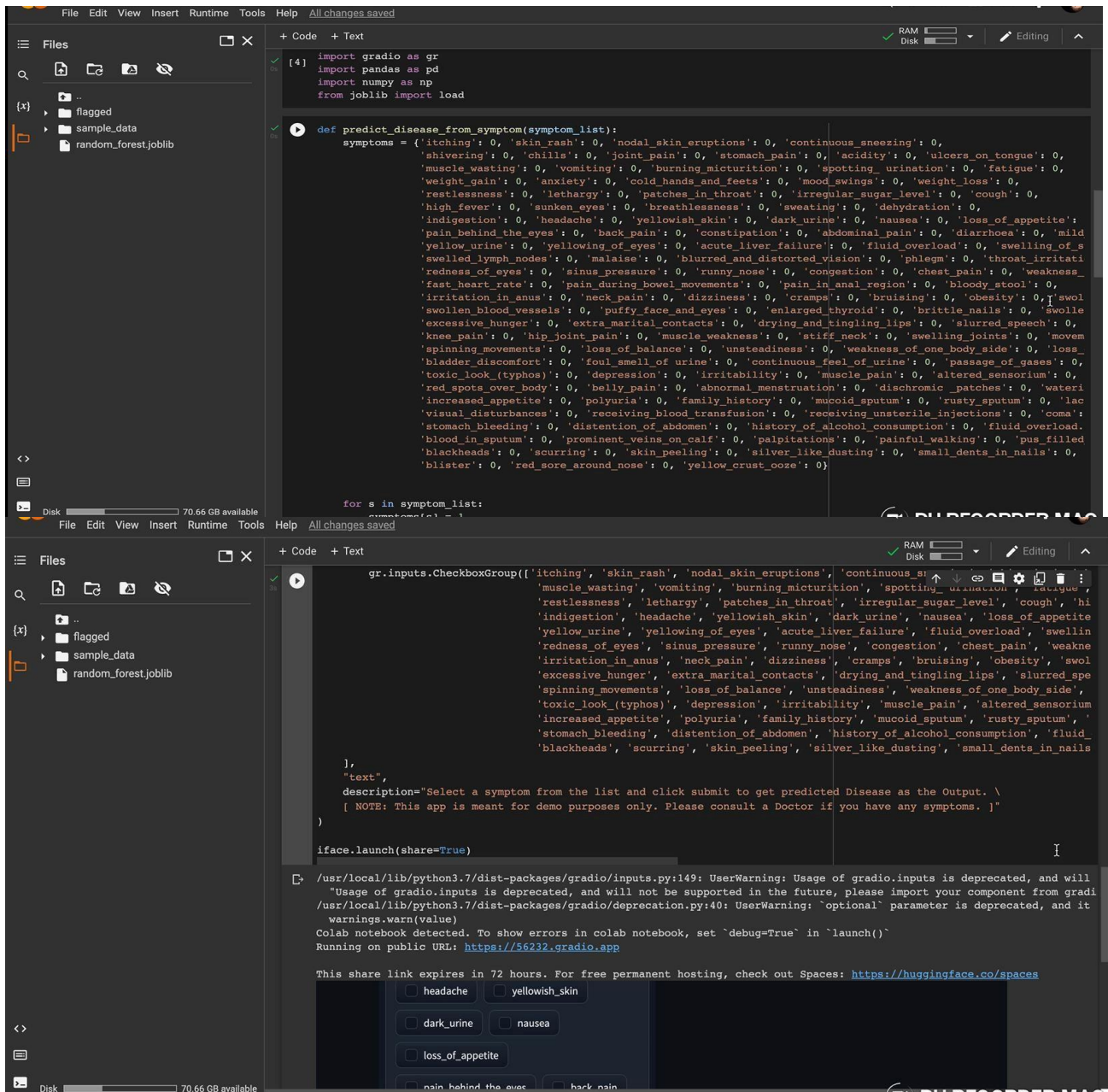


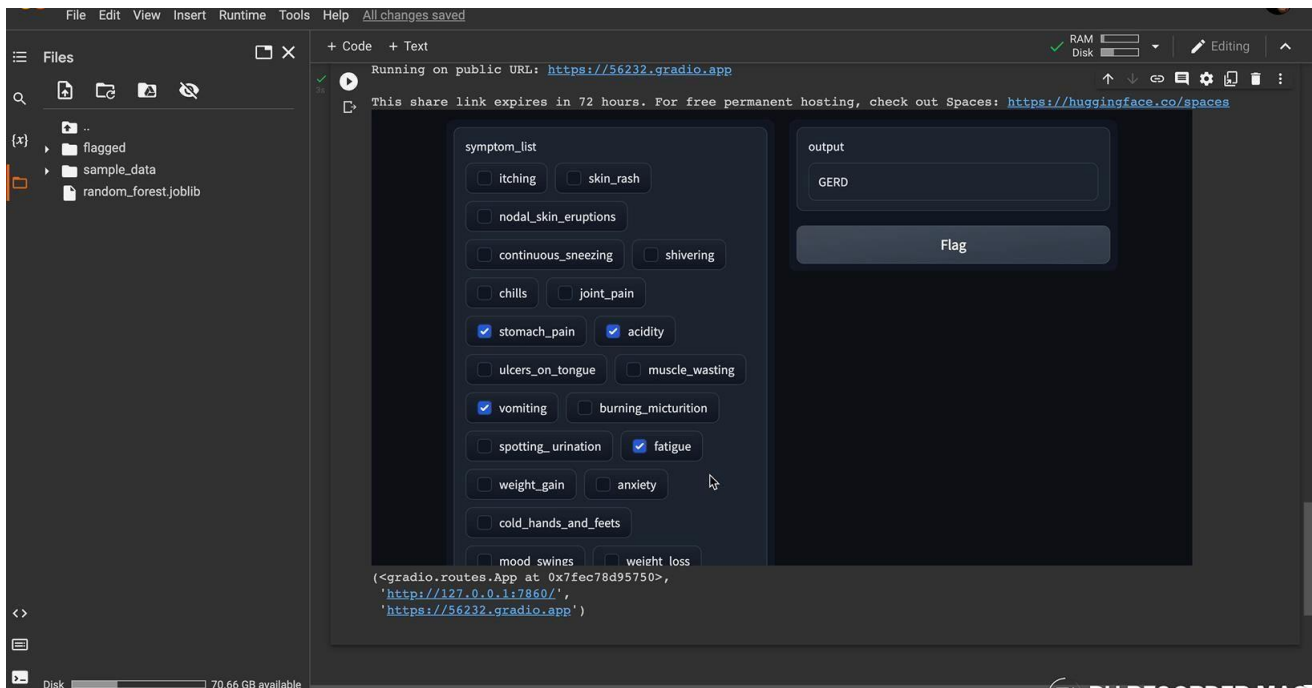
```
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
RAM
Disk
Editing

Files
{x}
..
  flagged
  sample_data
  random_forest.joblib

!pip install gradio

Downloading analytics_python-1.4.0-py2.py3-none-any.whl (15 kB)
Requirement already satisfied: pillow in /usr/local/lib/python3.7/dist-packages (from gradio) (7.1.2)
Requirement already satisfied: pydantic in /usr/local/lib/python3.7/dist-packages (from gradio) (1.9.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from gradio) (3.2.2)
Collecting orjson
  Downloading orjson-3.8.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (270 kB)
    |-----| 270 kB 47.6 MB/s
Collecting python-multipart
  Downloading python-multipart-0.0.5.tar.gz (32 kB)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from gradio) (2.23.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from gradio) (1.21.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.7/dist-packages (from gradio) (2022.7.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from gradio) (1.3.5)
Collecting markdown-it-py[linkify,plugins]
  Downloading markdown_it_py-2.1.0-py3-none-any.whl (84 kB)
    |-----| 84 kB 2.3 MB/s
Collecting ffmpeg
  Downloading ffmpeg-0.3.0.tar.gz (4.8 kB)
Collecting pycryptodome
  Downloading pycryptodome-3.15.0-cp35-abi3-manylinux2010_x86_64.whl (2.3 MB)
    |-----| 2.3 MB 36.6 MB/s
Collecting fastapi
  Downloading fastapi-0.81.0-py3-none-any.whl (54 kB)
    |-----| 54 kB 2.3 MB/s
Collecting pydub
  Downloading pydub-0.25.1-py2.py3-none-any.whl (32 kB)
Collecting websockets
  Downloading websockets-10.3-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2_12_x86_64.manylinux2010_x86_64.whl (112 kB)
    |-----| 112 kB 44.6 MB/s
Collecting h11<0.13,>=0.11
  Downloading h11-0.12.0-py3-none-any.whl (54 kB)
    |-----| 54 kB 3.0 MB/s
Collecting httpx
  Downloading httpx-0.23.0-py3-none-any.whl (84 kB)
    |-----| 84 kB 2.4 MB/s
```





4.5 Summary

In this way, with the help of Machine Learning, there can be early diagnosis of the patient.

Boon to common people and sometimes, doctors.

Saves time required for diagnosing the problem.

CHAPTER 5

TECHNICAL IMPLEMENTATION & ANALYSIS

5.1 Technical coding and code solutions

```
[ ] import gradio as gr
import pandas as pd
import numpy as np
from joblib import load
```

```
def predict_disease_from_symptom(symptom_list):
    symptoms = {'itching': 0, 'skin_rash': 0, 'nodal_skin_eruptions': 0, 'continuous_sneezing': 0,
                'shivering': 0, 'chills': 0, 'joint_pain': 0, 'stomach_pain': 0, 'acidity': 0, 'ulcers_on_tongue': 0,
                'muscle_wasting': 0, 'vomiting': 0, 'burning_micturition': 0, 'spotting_urination': 0, 'fatigue': 0,
                'weight_gain': 0, 'anxiety': 0, 'cold_hands_and_feets': 0, 'mood_swings': 0, 'weight_loss': 0,
                'restlessness': 0, 'lethargy': 0, 'patches_in_throat': 0, 'irregular_sugar_level': 0, 'cough': 0,
                'high_fever': 0, 'sunken_eyes': 0, 'breathlessness': 0, 'sweating': 0, 'dehydration': 0,
                'indigestion': 0, 'headache': 0, 'yellowish_skin': 0, 'dark_urine': 0, 'nausea': 0, 'loss_of_appetite': 0,
                'pain_behind_the_eyes': 0, 'back_pain': 0, 'constipation': 0, 'abdominal_pain': 0, 'diarrhoea': 0, 'mild_fever': 0,
                'yellow_urine': 0, 'yellowing_of_eyes': 0, 'acute_liver_failure': 0, 'fluid_overload': 0, 'swelling_of_stomach': 0,
                'swelled_lymph_nodes': 0, 'malaise': 0, 'blurred_and_distorted_vision': 0, 'phlegm': 0, 'throat_irritation': 0,
                'redness_of_eyes': 0, 'sinus_pressure': 0, 'runny_nose': 0, 'congestion': 0, 'chest_pain': 0, 'weakness_in_limbs': 0,
                'fast_heart_rate': 0, 'pain_during_bowel_movements': 0, 'pain_in_anal_region': 0, 'bloody_stool': 0,
                'irritation_in_anus': 0, 'neck_pain': 0, 'dizziness': 0, 'cramps': 0, 'bruising': 0, 'obesity': 0, 'swollen_legs': 0,
                'swollen_blood_vessels': 0, 'puffy_face_and_eyes': 0, 'enlarged_thyroid': 0, 'brittle_nails': 0, 'swollen_extremeties': 0,
                'excessive_hunger': 0, 'extra_marital_contacts': 0, 'drying_and_tingling_lips': 0, 'slurred_speech': 0,
                'knee_pain': 0, 'hip_joint_pain': 0, 'muscle_weakness': 0, 'stiff_neck': 0, 'swelling_joints': 0, 'movement_stiffness': 0,
                'spinning_movements': 0, 'loss_of_balance': 0, 'unsteadiness': 0, 'weakness_of_one_body_side': 0, 'loss_of_smell': 0,
                'bladder_discomfort': 0, 'foul_smell_of_urine': 0, 'continuous_feel_of_urine': 0, 'passage_of_gases': 0, 'internal_itching': 0,
                'toxic_look(typhos)': 0, 'depression': 0, 'irritability': 0, 'muscle_pain': 0, 'altered_sensorium': 0,
                'red_spots_over_body': 0, 'belly_pain': 0, 'abnormal_menstruation': 0, 'dischromic_patches': 0, 'watering_from_eyes': 0,
                'increased_appetite': 0, 'polyuria': 0, 'family_history': 0, 'mucoid_sputum': 0, 'rusty_sputum': 0, 'lack_of_concentration': 0,
                'visual_disturbances': 0, 'receiving_blood_transfusion': 0, 'receiving_unsterile_injections': 0, 'coma': 0,
                'stomach_bleeding': 0, 'distention_of_abdomen': 0, 'history_of_alcohol_consumption': 0, 'fluid_overload.1': 0,
                'blood_in_sputum': 0, 'prominent_veins_on_calf': 0, 'palpitations': 0, 'painful_walking': 0, 'pus_filled_pimples': 0,
                'blackheads': 0, 'scurring': 0, 'skin_peeling': 0, 'silver_like_dusting': 0, 'small_dents_in_nails': 0, 'inflammatory_nails': 0,
                'blister': 0, 'red_sore_around_nose': 0, 'yellow_crust_ooze': 0}
```

```

for s in symptom_list:
    symptoms[s] = 1

df_test = pd.DataFrame(columns=list(symptoms.keys()))
df_test.loc[0] = np.array(list(symptoms.values()))

clf = load(str("./random_forest.joblib"))
result = clf.predict(df_test)

del df_test

return f"{result[0]}"

```

```

iface = gr.Interface(
    predict_disease_from_symptom,
    [
        gr.inputs.CheckboxGroup([
            'itching', 'skin_rash', 'nodal_skin_eruptions', 'continuous_sneezing', 'shivering', 'chills', 'joint_pain', 'stomach_pain',
            'muscle_wasting', 'vomiting', 'burning_micturition', 'spotting_urination', 'fatigue', 'weight_gain', 'anxiety', 'cold_han',
            'restlessness', 'lethargy', 'patches_in_throat', 'irregular_sugar_level', 'cough', 'high_fever', 'sunken_eyes', 'breathles',
            'indigestion', 'headache', 'yellowish_skin', 'dark_urine', 'nausea', 'loss_of_appetite', 'pain_behind_the_eyes', 'back_pai',
            'yellow_urine', 'yellowing_of_eyes', 'acute_liver_failure', 'fluid_overload', 'swelling_of_stomach', 'swelled_lymph_nodes',
            'redness_of_eyes', 'sinus_pressure', 'runny_nose', 'congestion', 'chest_pain', 'weakness_in_limbs', 'fast_heart_rate', 'pa',
            'irritation_in_anus', 'neck_pain', 'dizziness', 'cramps', 'bruising', 'obesity', 'swollen_legs', 'swollen_blood_vessels',
            'excessive_hunger', 'extra_marital_contacts', 'drying_and_tingling_lips', 'slurred_speech', 'knee_pain', 'hip_joint_pain',
            'spinning_movements', 'loss_of_balance', 'unsteadiness', 'weakness_of_one_body_side', 'loss_of_smell', 'bladder_discomfort',
            'toxic_look_typhos', 'depression', 'irritability', 'muscle_pain', 'altered_sensorium', 'red_spots_over_body', 'belly_pai',
            'increased_appetite', 'polyuria', 'family_history', 'mucoid_sputum', 'rusty_sputum', 'lack_of_concentration', 'visual_dist',
            'stomach_bleeding', 'distention_of_abdomen', 'history_of_alcohol_consumption', 'fluid_overload.1', 'blood_in_sputum', 'pro',
            'blackheads', 'scurring', 'skin_peeling', 'silver_like_dusting', 'small_dents_in_nails', 'inflammatory_nails', 'blister',
        ]),
        "text",
        description="Select a symptom from the list and click submit to get predicted Disease as the Output. \
        [ NOTE: This app is meant for demo purposes only. Please consult a Doctor if you have any symptoms. ]"
    )

iface.launch(share=True)

```

5.2 Summary

The main aim of this paper is to predict the disease in accordance with symptoms put down by the patients with proper implementation of Machine Learning algorithm. In this paper we have used four Machine Learning algorithm for prediction and achieved the mean accuracy of more than 84% which shows remarkable rectification and high accuracy than previous work and also makes this system more reliable than the existing one for this job and hence provides better satisfaction to the user in comparison with the other one. It also stores the data entered by the user and the name of the disease the patient is suffering from in the Database which can be used as past record and will help in future for future treatment and thus contributing in easier health management .We have also created a GUI for better interaction with the system by users which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease easily with different parameters and models. In the end we can say that our system has no threshold of the users because everyone can use this system.

CHAPTER 6

PROJECT OUTCOME AND APPLICABILITY

6.1 Outline

There is a demand to make such a system that will help end users to predict diseases on the basis of symptoms given in it without visiting hospitals. By doing so, it will decrease the rush at OPD's of hospitals and bring down the workload on medical staff. Not only this, this system will reduce the costly treatment and panic moment at the end stages so that proper medication can be provided at the right time and we can lower down the death rate as well. This system also consists of a feature of Database which stores the data entered by the end users and the name of the disease the patient is suffering from that can be used as a past record and will help in further treatment in future. The analysis accuracy is increased by using Machine Learning algorithms. Altogether this system will help in easier health management.

6.2 Significant project outcomes

This project aims to predict the disease based on 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. In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data

6.3 Project applicability on Real-world applications

It can be useful for the students who are studying medicine for predictive assessment of any illness. It can give an idea for emergency usage of any generic medicine.

It can be useful for suggesting the best possible doctors for the best possible results.

It can be useful for maintaining medical records of the patients to reduce workload of doctors.

6.4 Inference

With the proposed system, higher accuracy can be achieved. We not only use structured data, but also the text data of the patient based on the proposed k-mean algorithm. To find that out, we combine both data, and the accuracy rate can be reached up to 84%. None of the existing system and

work is focused on using both the data types in the field of medical big data analytics. We propose a Random Forest Classifier algorithm for both structured and unstructured data. The disease risk model is obtained by combining both structured and unstructured features.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATION

7.1 Outline

We proposed general disease prediction system based on machine learning algorithm. We utilized Radom forest algorithm to classify patient data because today medical data growing very vastly and that needs to process existed data for predicting exact disease based on symptoms. We got accurate general disease risk prediction as output, by giving the input as patients record which help us to understand the level of disease risk prediction. Because of this system may leads in low time consumption and minimal cost possible for disease prediction and risk prediction.

7.2 Limitations/Constraints of the System

We need to make an app/website to provide better access to everyone.

We need to make sure the accuracy comes up to a hundred percent.

We need to find ways to make the system suggest medicines for the disease too.

7.3 Future Enhancement

The results of this study confirm the application of machine learning algorithms in prediction and early detection of diseases. To our best understanding, the model built according to the proposed method exhibits better accuracy than the existing ones. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 84.8% with a regular speed which is quicker than that of the unimodal disease risk prediction algorithm and produces report. The Further work will mainly focus on the Medical Assistance and proper Medication to the patients as soon as possible so as to build the best infrastructure and quick easiest way in the medical sectors.

7.4 Inference

The feature extracted data is further evaluated to predict the disease by using the classifiers such as Random forest and Naïve Bayes Classifiers, Decision tree, K-Nearest Neighbour.

On comparing the four machine learning algorithms, K Nearest Neighbour shows 95.6%, Naïve Bayes Classifiers shows 94.5%, Random Forest model has the highest accuracy of 95.7% than the Decision Tree algorithm shows 92.4%.

The results were evaluated with accuracy, sensitivity, specificity, positive predictive value and negative predictive value.

REFERENCES

1. Disease and symptoms Dataset from Kaggle
2. Nivethitha.A, Narendran.G, Pramoth Krishnan.T, 'Smart Disease Prediction Using Machine Learning'.
3. www.github.com
4. Priyanka J.Panchal, Shaezah A. Mhaskar, Tejal S.Ziman, 'Disease Prediction using Machine Learning'
5. Research Gate
6. IJRASET
7. Disease Prediction Project in Python GUI using ML
8. Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach" IEEE Xplore Part Number:CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp.1211-1215, 2019.
9. Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.
10. "Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil," Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.