MIT ADT UNIVERSITY

SCHOOL OF ENGINEERING

MINI PROJECT ON

" Disease PREDICTION SYSTEM"

Submitted by :

Achal Rajesh Mate (2203541)

Pushkar Ashok Narkhede(2203528)

Ritu Kumari (22193195)

Under the guidance of :

Prof. Dr. Rajeshree Nayak

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MIT SCHOOL of Engineering

Loni Kalbhor ,Pune.

DEPARTMENT OF COMPUTER ENGINEERING

MIT ADT SCHOOL OF ENGINEERING,PUNE

CERTIFICATE

This is to certify that the Mini- Project report entitled

"Disease PREDICTION SYSTEM"

Submitted by:

Achal Rajesh Mate
Pushkar Narkhede
Ritu Kumari

is a record of bonafide work carried out by them, under my guidance, in partial fulfillment of the requirement for the **Second Year of Engineering(CSE)** at **MIT School of Engineering, Pune** under MIT Art, Design & Technology University.

Date:

Dr. RajneeshKaur Sachdeo

Prof. Dr. Rajeshree Nayak

Department of CSE,

M.I.T. School of Engineering

Loni Kalbhor, Pune.

Dean Engineering,

Head ,Department of CSE

MIT School of Engineering

Loni-Kalbhor, Pune.

ACKNOWLEDGEMENT

It is our proud privilege and duty to acknowledge the kind of help and guidance received from several people in the preparation of this report. It would not have been possible to prepare this report in this form without their valuable help, cooperation and guidance.

First and foremost, we wish to record our sincere gratitude to the management of this college and to our Respected Principal for his constant support and encouragement in the preparation of this report and for the availability of library and laboratory facilities needed to prepare this report. Our sincere thanks to Prof. Dr. Rajneeshkaur Sachdeo Bedi Head of Computer Engineering Department for his valuable suggestions and guidance throughput the preparation of this report.

We express our sincere gratitude to our guide **Prof. Dr. Rajshree Nayak** for guiding us in investigations of this project and in carrying out experimental work. Our numerous discussions were extremely helpful.

Last but not the least we wish to thank my parents for financing my studies and helping us throughout my life achieving perfection and excellence. Their personal help in making this report and seminar worth presentation is gratefully acknowledged.

Achal Rajesh Mate (2203541)

Pushkar Ashok Narkhede (220328)

Ritu Kumari (22193195)

Abstract

Now a days health has greater effect in our life. Healthy life is an important factor in onces life . Number of diseases are increasing day by day if they are not predicted on time one can cause death

In this project we have develop a System which predict the disease present in the system by taking user input as symptoms

Prediction system is a python based Web project which uses structure dataset for developing the Machine Learning Models. System Ask User to enter the symptoms and predict where the person is surviving from any disease or not.

Contents

1.INTRODUCTION

In this project we have develop a System which predict the diseases. Mainly three Disease are present in system . Covid 19, Breast Cancer, Heart Cause Disease.

This is web Base System which ask user to enter the symptom which they want to predict that they are surviving from disease or not.

By using various machine learning classification algorithms have train and build the model for disease prediction. Structure Dataset is used for Machine Learning Model development. User or patient first need to login into the system . After the successfully login of patient they need to select the disease which they want to predict . After selecting the disease patient need to enter the symptoms which has been ask by the system . The Model will predict where the particular person is surving from disease or not .

2.PROJECT PLAN

2.1 PROBLEM :

Now a days number of disease are increasing day by day which can cause death if they are cure or predicted on time. Therefore there is need or system which can predict the disease in early stage so patient can come to know if it surviving from any disease or not

2.3 SOLUTION :

Develop a System which can predict the disease base on symptoms

2.3 PROJECT OUTCOME :

After Completing this project you should be able to predict the disease which are present in the system .

2.4 TECHNOLOGIES :

Tools--

1. Jupyter-Lab or Jupyter Notebook
2. Pycharm
3. SQL Workbench

Technology--

1. Html 5
2. Css
3. Javascript for authentication
4. Python Programming (for web development and integration of ML model)
5. flask as a web framework
6. Machine Learning. (Classification algorithm (Supervised ML)
7. Mysql Database.

3.PROJECT REQUIREMENTS

Requirement Analysis is the first phase of software development process. This phase focuses to understand the problem. Requirement Analysis is on identifying what is need from these systems, not how the system will achieve its goals. In this phase often at least two parties are involved in Software Development-a client and a developer. The developer has to develop the system to satisfy the clients' needs. The developer and client arrange a meeting and discuss his/her own views. The developer asks the clients for his/her needs. After a meeting the developer understands what the requirements of the client are. Before starting of the development process, the developer analyse , test the requirements which are given by the clients. According to those requirements the developer starts development process. Hence the developer needs a user's problem.

In the software requirement we are dealing with the requirements of the proposed system, that's the capabilities of that system, which is yet to be developed, should have. The software requirement specification (SRS) is a document that completely describes what the proposed software should do without describing how the software will do it. So the basic goal of Requirement Phase is to produce the SRS, which describes the complete external behaviour of the proposed software.

2.1 PERFORMANCE REQUIREMENTS

- PORTABILITY

The system portability should be taken care of without any interventions. Portability means the capability of the software to be transferred from one environment to another. Thus our system provides the portability that it can run on any machine with backward compatibility.

- EFFICIENCY

The system should be capable of providing the required performance related to the amount of resources of the organization.

- RELIABILITY

  Our system is capable enough to maintain the level of performance.

- ACCURACY

  The system also calculates the accuracy of the images classified. To give 99% accurate system is our motive.

2.2 HARDWARE REQUIREMENTS

- MIN HARD-DISK – 1 TB
- LAPTOP OR PC
- WIFI\ROUTER

2.3 SOFTWARE REQUIREMENTS

- OS
- PYTHON
- IDE (PYCHARM\VS CODE)

2.4 PACKAGES

- TENSORFLOW
- NUMPY
- MATPLOTLIB

4.PROJECT DESIGN

4.1FEASIBILITY STUDY

A **feasibility study** is an evaluation of a proposal designed to determine the difficulty in carrying out a designated task. Generally, a feasibility study precedes technical development and project implementation.

1.TECHNOLOGY AND SYSTEM FEASIBILITY :

The assessment is based on an outline design of system requirements in terms of Input, Processes, Output, Fields, Programs, and Procedures. This can be quantified in terms of volumes of data, trends, frequency of updating, etc. in order to estimate whether the new system will perform adequately or not. Technological feasibility is carried out basically to determine whether the company has the capability in terms of software, hardware, personnel and expertise to handle the completion of the project. HCL fulfilled all the above requirements for the efficient working of web application.

2.ECONOMIC FEASIBILITY :

Economic analysis is the most frequently used method for evaluating the effectiveness of a new system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system. An entrepreneur must accurately weigh the cost versus benefits before taking an action

3.COST BASED STUDY :

It is important to identify cost and benefit factors. Cost and benefits can be categorized into the following categories. Basically it is an analysis of the costs to be incurred in

the system and benefits derivable out of the system. In a broad sense the costs can be divided into two types 1. Development costs 2. Operating costs.

4.TIME BASED STUDY :

Contrast to the traditional system management it can generate any information of data just by single click and it saves user time .No extra time is being provided to deliver application.

5.OPERATIONAL FEASIBILITY :

It is a measure of how well a proposed system solves the problems, and takes advantages of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

SYSTEM DEVELOPMENT

Details Description of each disease

1. **Breast Cancer Prediction**

   Assign to = Achal Mate

Dataset Description –

- Number of Instances: 569
- Number of Attributes: 30 numeric, predictive attributes and the class

Attribute Information:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.
For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

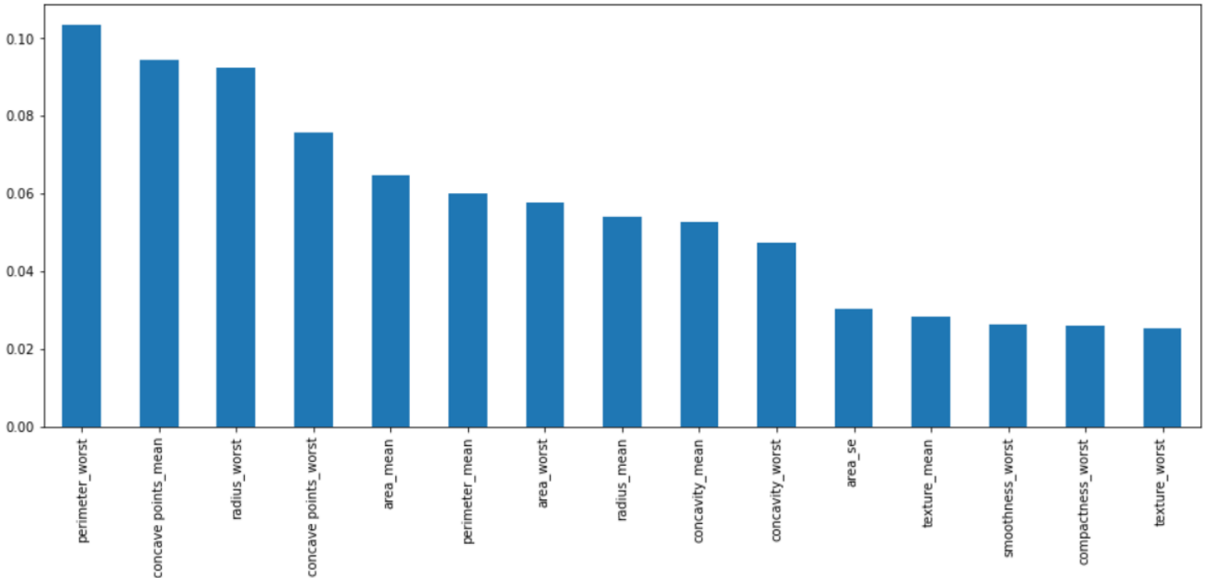
- Missing Values: 569
- Class Distribution: 212 - Malignant, 357 – Benign

Depending on the types of cells in a tumor, it can be:
Benign - The tumor doesn’t contain cancerous cells.
Malignant - The tumor contains cancerous cells.

Observation:

fractal_dimension_mean,texture_se, smoothness_se, symmetry_se, and fractal_dimension_se are least correlated with the target variable.
All other features have a significant correlation with the target variable.(i.e diagnosis)

Feature Importance -



top 15 features which contributes to target variable

1. Need to verify which feature is significant and which is not from this top 15
2. for this we will be using Hypothesis testing and other feature selection technique

Hypothesis testing -

- Null Hypothesis ==> all the features are not significent
- Alternate Hypothesis ==> all the feature are significent
- alpha values ==>0.05

by Hypothesis testing
if(p_value < alpha) then **reject Null Hypothesis**
else **Fail to reject Null Hypothesis**
or
if(p_value > alpha) then **Fail to reject Null Hypothesis**
else **reject Null Hypothesis**

calculating VIF score we get top 9 feature

1. area_mean
2. texture_mean
3. fractal_dimension_mean
4. fractal_dimension_se
5. texture_se
6. smoothness_se
7. concavity_se
8. symmetry_se
9. area_se

But Our Dataset Contain 10 Independent features which further they are categories into three categories

1. Mean
2. Worst
3. Square(Se)

As Above selected feature does not include any worst categories' feature so to get more significant we drop all worst categories features from the main dataset

Top 15 RFE features

[('radius_mean', True, 1),
('texture_mean', True, 1),
('perimeter_mean', True, 1),
('area_mean', True, 1),
('smoothness_mean', True, 1),
('compactness_mean', True, 1),
('concavity_mean', True, 1),
('concave points_mean', True, 1),
('symmetry_mean', True, 1),
('fractal_dimension_mean', True, 1),
('radius_se', True, 1),
('texture_se', True, 1),
('perimeter_se', True, 1),
('area_se', True, 1),
('smoothness_se', False, 5),
('compactness_se', False, 4),
('concavity_se', False, 2),
('concave points_se', False, 6),
('symmetry_se', False, 3),
('fractal_dimension_se', True, 1)]

Observation

Among this features

- smoothness_mean
- concave points_mean
- radius_se
- area_mean
- perimeter_mean
- fractal_dimension_mean
- perimeter_se
- radius_mean
- ompactness_mean

are the features with higher **p value** than alpha value
So they are drop to get the significant feature
Now, VIF score is also in the range .

Selected Features

1. texture_mean
2. concavity_mean
3. symmetry_mean
4. texture_se
5. area_se
6. fractal_dimension_se

Model Buliding

Model and Training Accuracy

Algorithm	Accuracy
ExtraTreesClassifier	0.982456
Random Forest Classifier	0.973684
XGBClassifier	0.973684
Gradient Boosting Classifier	0.956140
svc rbf Kernal	0.947368
K Neighbors Classifier	0.947368
navieBayes_gaussian	0.938596
Ada Boost Classifier	0.938596
svc linear Kernal	0.929825
logistic_name	0.921053
Decision Tree Classifier	0.912281

After Perfoming Parameter Tuning

Algorithm	Accuracy
ExtraTreesClassifier	0.982456
xgb_classifier_test	0.982456
RandomForestClassifier	0.973684
Logistic Regression	0.964912
SVC	0.964912
KNeighborsClassifier	0.956140
AdaBoostClassifier	0.947368
GaussianNB	0.938596
Gradient Boosting Classifier	0.921053
DecisionTreeClassifier	0.903509

2 . Covid 19 Prediction

Assign To= Pushkar Narkhede

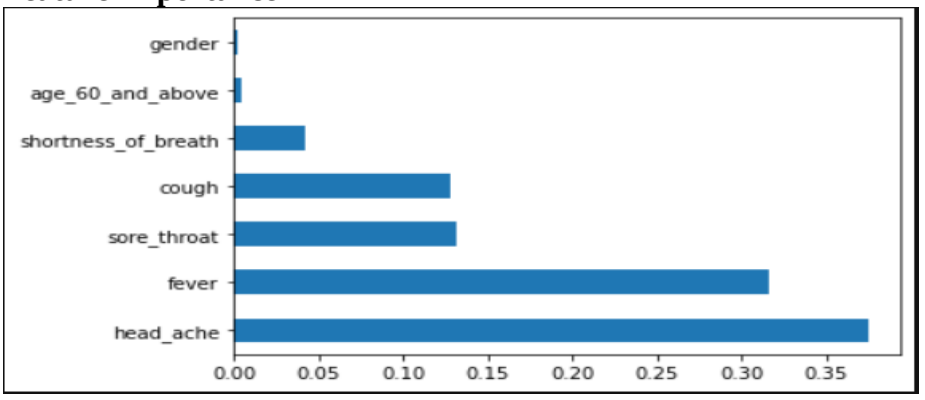
Dataset Description -

1. test_date ==> Date for covid 19 test
2. cught ==> is the patience having cough or not (binary variable yes=1, no=0)
3. fever ==> is the patience having fever or not (binary variable yes=1, no=0)
4. sore_throat ==> is the patience having sore_throat or not (binary variable yes=1, no=0)
5. shortness_of_breadth ==> is the patience having shortness_of_breadth or not (binary variable yes=1, no=0)
6. head_ache ==> is the patience having head_ache or not (binary variable yes=1, no=0)
7. corona_result ==> is the patience having corona_result or not (categorical variable -ve for covid,+ve for covid, other=> no confirmation with covid or not) *self encoded* **(1= +ve, 0= -ve)**

8. gender ==> is the patience having gender or not (categorical variable male,female) *self encoded* **(1= male, 0=female)**
9. test_indication ==> is the patience having gender or not (categorical variable contact_with_confirm=> covid due to contact of covid 19 person, abord=> covid in other country, other=>no idea about covid infection)

- Total data in dataset : 2742596
- Total null values in dataset : 640530
- Total data remains if we remove all null values = 2102066
- 23.35 % of null data is present in the dataset.

Feature Importance -



P-Values of all Features are 0.00

Interpretation

- After performing hypothesis testing by considering p value **gender** i am on the conclusion that all features we have are significant
- same interpretaion is given by RFE(recursive feature selection) and ExtraTreeClassifier feature contribution score.
- if **age** feature is not contributing more we will be dropping that feature.
- VIF score is also in the range.

Feature selected-

1. cough
2. fever
3. sore_throat
4. shortness_of_breath
5. head_ache
6. age_60_and_above (on hold)

Model Selection

Models and Training Accuracy

1) Decision Tree = 0.945463
2) Random Forest = 0.945463
3) Navie bayes= 0.940577
4) Logistics Regression=0.938743
5) Gradient Boosting=0.945463
6) Xgboost= 0.945463
7) ADABoost=0.934402

After performing hyperparameter tunning

1) Navie bayes = 0.940577
2) Logistics Regression=0.938743
3) Gradient Boosting=0.945463
4) Xgboost=0.945463
5) ADABoost=0.934402

3 Heart Cancer Prediction

Assign to = Ritu Kumari (Completed By Achal Mate)

Dataset Information : The dataset contains 76 features from 303 patients, however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, In this project we will be using the dataset consisting of 303 patients with 14 features set.

Features Explanations:

1) age (Age in years)

2) sex : (1 = male, 0 = female)

3) cp (Chest Pain Type): [0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina]

4) trestbps (Resting Blood Pressure in mm/hg)

5) chol (Serum Cholesterol in mg/dl)

6) fps (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes]

7) restecg (Resting ECG): [0: showing probable or definite left ventricular hypertrophy by Estes' criteria, 1: normal, 2: having ST-T wave abnormality]

8) thalach (maximum heart rate achieved)

9) exang (Exercise Induced Angina): [1 = yes, 0 = no]

10) oldpeak (ST depression induced by exercise relative to rest)

11) slope (the slope of the peak exercise ST segment): [0: downsloping; 1: flat; 2: upsloping]

12) ca [number of major vessels (0–3)

13) thal : [1 = normal, 2 = fixed defect, 3 = reversible defect]

14) target: [0 = disease, 1 = no disease] ---> In The dataset we have 303 rows with 14 variables

variables types:

1) Binary: sex, fbs, exang, target

2) Categorical: cp, restecg, slope, ca, thal

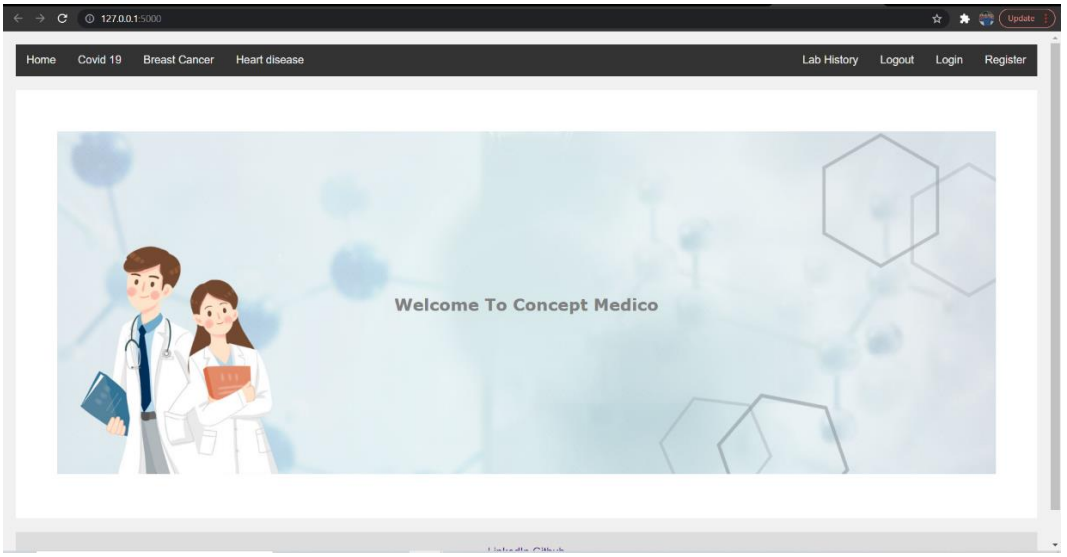
3) Continuous: age, trestbps, chol, thalac, oldpea

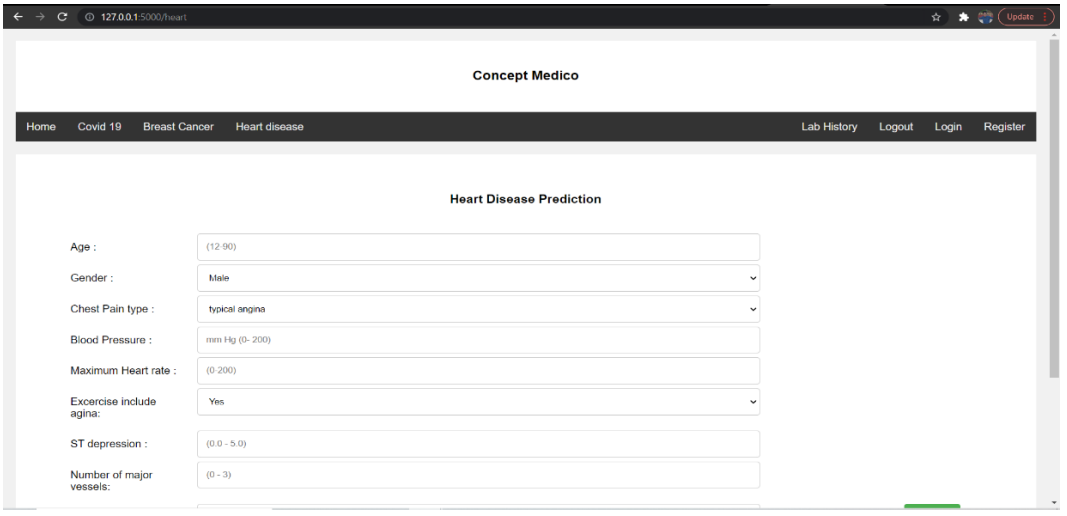
6.A INPUT

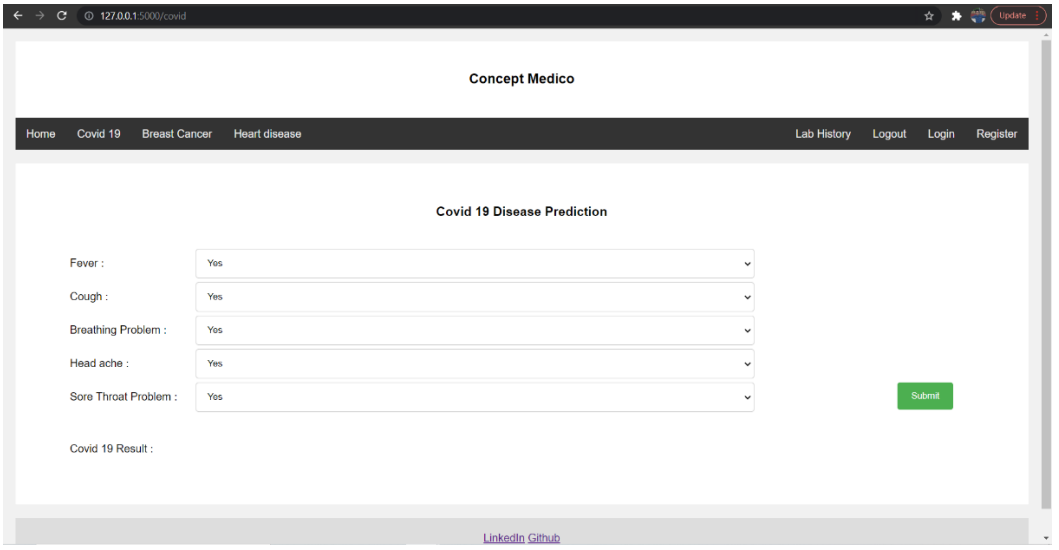
Project Code Link :

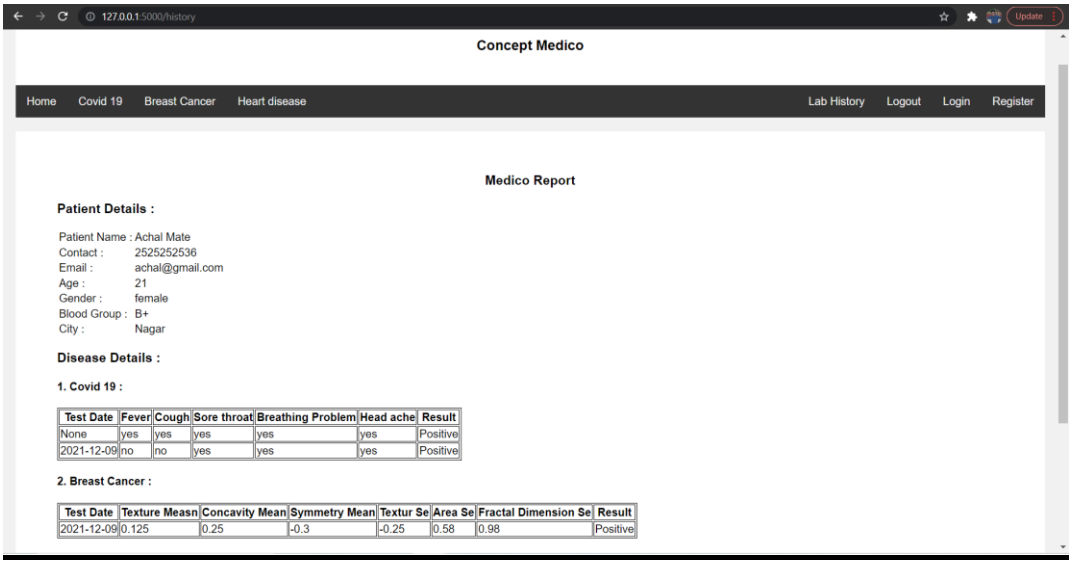
https://github.com/AchalMate/Disease-Prediction-System

6. B OUTPUT






127.0.0.1:5000/covid
Update
Concept Medico
Home
Covid 19
Breast Cancer
Heart disease
Lab History
Logout
Login
Register
Covid 19 Disease Prediction
Fever :
Yes
Cough :
Yes
Breathing Problem :
Yes
Head ache :
Yes
Sore Throat Problem :
Yes
Submit
Covid 19 Result :
LinkedIn Github


127.0.0.1:5000/history
Update
Concept Medico
Home
Covid 19
Breast Cancer
Heart disease
Lab History
Logout
Login
Register
Medico Report
Patient Details :
Patient Name : Achal Mate
Contact : 2525252536
Email : achal@gmail.com
Age : 21
Gender : female
Blood Group : B+
City : Nagar
Disease Details :
1. Covid 19 :
Test Date | Fever | Cough | Sore throat | Breathing Problem | Head ache | Result
None | yes | yes | yes | yes | yes | Positive
2021-12-09 | no | no | yes | yes | yes | Positive
2. Breast Cancer :
Test Date | Texture Measn | Concavity Mean | Symmetry Mean | Textur Se | Area Se | Fractal Dimension Se | Result
2021-12-09 | 0.125 | 0.25 | -0.3 | -0.25 | 0.58 | 0.98 | Positive

7 .SYSTEM TESTING

LEVELS OF TESTING

1. UNIT TESTING

Unit testing is a procedure used to verify that a particular segment of source code is working properly. The idea about unit tests is to write test cases for all functions or methods. Ideally, each test case is separate from the others. Unit testing focuses verification efforts on the smallest unit of software design, the software component or module. Using the component level design description as a guide, important control paths are tested to uncover errors within the boundary of the module. The relative complexity of test and uncovered errors is limited by the constraints scope established for unit testing. The unit test is white box oriented, and the step can be conducted in parallel for multiple components. In this project many aspects are covered under unit testing, because if any of the function does not work properly then system may be fail.

2. INTEGRATED TESTING

Integrated testing is a systematic technique for construction of the whole program structure whole at the same time conduction tests to uncover errors associated with interfacing. The objective is to take unit tested components and build a program structure that has been dictated by design. Integrated testing follows unit testing and precedes system testing. Integration testing takes as its input, modules that have been checked out by unit testing, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output, the integrated system ready for system testing. The purpose of Integration testing is to verify functional performance and reliability requirements placed on major design items.

3 .SYSTEM TESTING

System testing is executing a program to check logic changes made in it and with the intention of finding errors-making the program fail. Effective testing does not guarantee reliability. Reliability is a design consideration.

ACCEPTANCE TESTING

Acceptance testing is conducted by a customer to verify that the system meets the acceptance criteria of the requested application. It generally involves running a suite of tests on the completed system. Each individual test, known as a case, exercises a particular operating condition of the user's environment or feature of the system, and will result in a pass or fail boolean outcome. There is generally no degree of success or failure.

5.ALPHA TESTING

Alpha testing is an actual operational testing done by potential users/customers or an independent test team at the developers' site, but outside the development organization. In other words, alpha testing is a type of acceptance testing carried out at developer's site by users (internal staff). In this type of testing, the user goes on testing the system and the outcome is noted and observed by the developer simultaneously.

6.BETA TESTING

Beta testing comes after alpha testing. Beta testing is considered the second phase of software testing. Beta tests are typically external tests to identify any performances issues or bugs prior to an official release. Beta tests can be open or closed. A closed beta test is used to control the number of users participating. An open test is open to anyone who has an interest in beta testing.

Beta testers are important because it is almost impossible for developers to test their software in all of the various conditions that might occur. Software should never be released without thorough beta testing. It is impossible to predict or test software on all kinds of hardware with other applications. Some developers segment the closed beta into different release stages so they can maximize feedback. Historically the majority of feedback is received from beta testers within the first week of the beta

TESTING METHODOLOGY

1. BLACK BOX TESTING

Black-box test design treats the system as a "black-box", so it doesn't explicitly use knowledge of the internal structure. Black-box test design is usually described as focusing on testing functional requirements. Majority of the application are tested by black box testing method. We need to cover majority of test cases so that most of the bugs will get discovered by black box testing. Test cases can be designed as soon as the functional specifications are complete.

2. WHITE BOX TESTING

White box testing strategy deals with the internal logic and structure of the code. White box testing is also called as glass, structural, open box or clear box testing. The tests written based on the white box testing strategy incorporate coverage of the code written, branches, paths, statements and internal logic of the code etc. In order to implement white box testing, the tester has to deal with the code and hence is needed to possess knowledge of coding and logic i.e. internal working of the code. White box test also needs the tester to look into the code and find out which unit/statement/chunk of the code is malfunctioning.

- Defining the data and control flow in the program
- Uniform representation of the program, language independent
- Simple basic elements: assignment and condition
- **Statement:** each statement executed at least once
- **Branch:** each branch traversed (and every entry point taken) at least once Branch Coverage requires that each branch will have been traversed, and that every program entry point will have been taken, at least once.
- **Condition:** each condition True at least once and False at least once
- **Branch/Condition:** both Branch and Condition coverage achieved
- **Compound Condition:** all combinations of condition values at every branch statement covered (and every entry point taken). It also known as Multiple Condition Coverage.

8.FUTURE SCOPE

User Side

- We can add More Number of Disease for predictions
- Doctor Authentication can be added so doctor can able to login.
- Patient will able to take doctor appointment as the their requirement and doctor specification.

Developer Side :

- Will predict the disease on image dataset.
- Then will perform parallel of disease which can be train on structure and Unstructure dataset.
- Hybridize the model

9.REFERE NCES

https://machinelearningmastery.com/

https://www.python.org/

https://www.anaconda.com/

https://www.youtube.com/

https://www.kaggle.com/

https://archive.ics.uci.edu/ml/index.php

https://towardsdatascience.com/