Sardar Patel Institute of Technology

A Report On

"Data Analytics for Heart Disease Prediction with Covid-19 Possibility"

Second Year Computer Engineering
Sardar Patel Institute of Technology

May 2021

A report submitted in partial fulfillment of the requirement of Mini Project Course

CERTIFICATE

This to certify that the work on the project titled “Data Analytics for Heart Disease Prediction with Covid-19 Possibility” has been carried out by the following students, who are bonafide students of Sardar Patel Institute of Technology, Mumbai, in partial fulfillment of the syllabus requirement in the Mini Project (May 2021)

1. **Sejal Gurkhe (2019130017)**
2. **Sakshi Mahadik (2019130034)**
3. **Ishita More (2019130039)**

Project Guide: _____

(Prof. Jyoti Ramteke)

ACKNOWLEDGMENT

We are pleased to present our project on the topic “Data Analytics for Heart Disease Prediction using Covid-19 possibility” and take this opportunity to express our profound gratitude to all those people who guided us in this project.

First, we are thankful to the Sardar Patel Institute of Technology, Computer Engineering Department, and Professor Jyoti Ramteke for her guidance and support for our project work.

We would like to thank them for their patience, and faith in our capabilities and for giving us flexibility in working.

TABLE OF CONTENT

INTRODUCTION

According to the World Health Organization, every year 12 million deaths are recorded worldwide because of heart disease. It has been observed that the cases of cardiovascular disease have rapidly increased in the past few years. Research has been conducted to find and highlight the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even pinpointed as a silent killer, which leads to the death of the person without any obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn, reduces the complications.

Medical decisions are often made based on a doctor's intuition and experience, rather than on the knowledge hidden in the data. This might lead to errors and high costs that affect medical service quality. This project aims to predict future heart disease by analyzing data of users, which classifies whether they have heart disease or not, using machine-learning algorithms. The basic idea is to predict the possible heart diseases a user might have based on the data filled by them. The input would be the symptoms a user might be experiencing. The possible symptoms are heart rate, heartburn, etc. From this data, the machine will be predicting the possible diseases a user might have. This data is also used to predict whether the user might have a possibility of suffering from COVID-19.

PROBLEM DEFINITION

The major challenge to heart disease is its detection. The instruments available for predicting heart disease are either very expensive or are not efficient for calculating the chance of heart disease in humans. It is better to detect early cardiac diseases, which help to decrease the mortality rate and overall complications. But it is not possible to follow patients every day, in all cases accurately, and also there cannot be the availability of a doctor to consult a patient for 24 hours because it requires more patience, time, and expertise. In today's world, there is a huge amount of data available which can be used for analyzing the data for hidden patterns using various machine learning algorithms and it can be used for diagnosing medical data.

The main idea is to predict the possible Heart diseases a user might have based on the data (Symptoms) filled by the user, along with the possibility of Covid-19 based on Symptoms. If the user has heart disease, then precaution for curing the disease. There are instruments available that can predict heart disease, but they are expensive. Since we have a good amount of data, we can use various machine learning algorithms to analyze the data and find possible results.

OBJECTIVE

The prediction system for heart disease uses a system that contains huge amounts of data, used to extract hidden information for making intelligent medical diagnosis.

The main objectives of developing this project are:

1. To build a heart disease prediction system that predicts heart disease based on the data taken as input from the user by implementing a gradient Boosting Algorithm.
2. To determine significant risk factors based on medical dataset which may lead to heart disease.
3. To find the possibility of COVID-19 using symptoms experienced by users.
4. To reduce the cost of medical tests.
5. To analyze feature selection methods and understand their working principle.

LITERATURE SURVEY

Paper	DOI	Techniques used	Accuracy
Prediction of Heart Disease Using Machine Learning Algorithms. [1]	https://doi.org/10.1109/ICIICT1.2019.8741465	Decision Tree Naive Bayes	91% 87%
Predictive analytics to prevent and control chronic diseases [2]	https://doi.org/10.1109/ICATCCT.2016.7912028	Naïve Bayes, Decision tree, Support vector machine	SVM gives the best accuracy with 95.55%
Efficient heart disease prediction system using decision tree [3]	https://doi.org/10.1109/CCAA.2015.7148346	Decision tree	86.3%

The core concept of this paper [1] is to predict the arising possibilities of heart disease. This was done with the help of data mining techniques like the Decision Tree algorithm and naive Bayes algorithm. These algorithms were applied to the same dataset to see which of the best algorithms for accuracy level. The results showed that the DT has an accuracy of 91%, whereas the NB has an accuracy of 87%. Thus, it was concluded from this paper that DT is better for handling medical datasets.

This paper [2] proposed predictive analytics to prevent and control the chronic disease with the help of machine learning techniques such as naive Bayes, support vector machine, decision tree, and artificial neural network and they have used UCI machine learning repository datasets to calculate the accuracy. Among them, the Support vector machine gives the best accuracy of 95.55%.

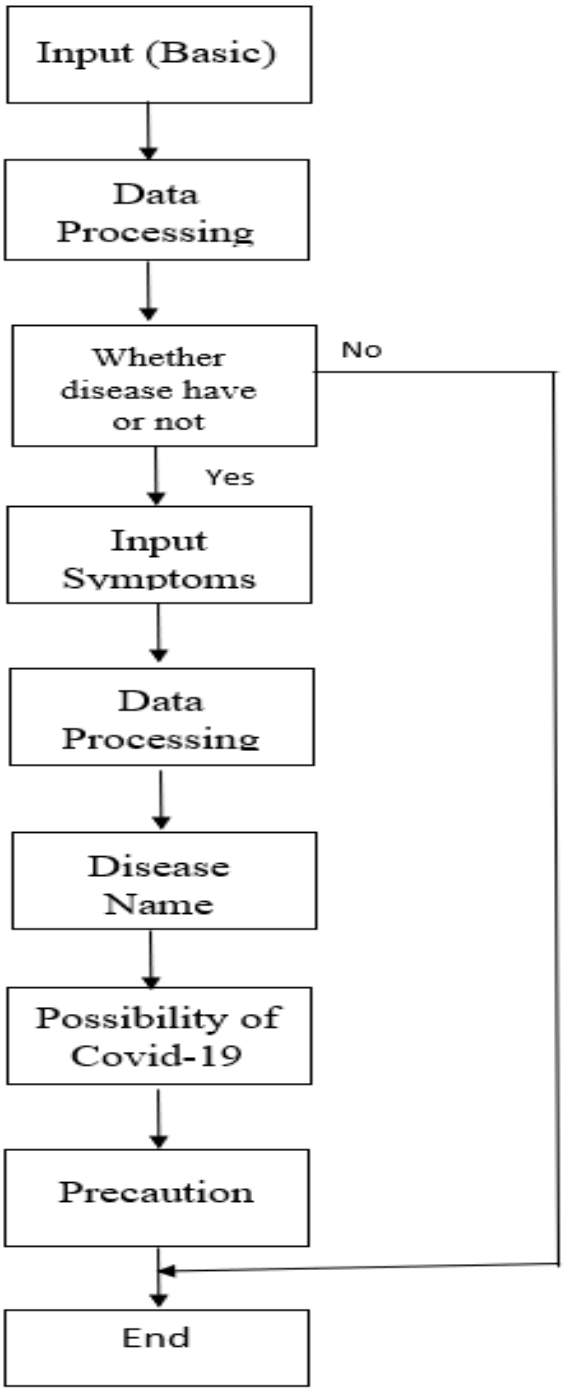
This paper [3] developed a data mining model to predict heart disease efficiently. It mainly helps medical practitioners to make efficient decisions based on the given parameters. The author has used Cleveland dataset from UCI, and they used age, sex, resting blood pressure, chest pain, serum, cholesterol, fasting blood sugar, etc. as attributes. Furthermore, they have divided the datasets into two parts: one is for testing, and the other one is for training. They have used a 10-fold method to find accuracy. The DT has given an accuracy of 86.3%.

SCOPE OF WORK

Here, the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcomes. Data modeling and analysis tools such as data mining have the potential to create a knowledge-rich environment that can help to significantly improve the quality of medical decisions. Doctors and medical institutions can utilize this for diagnosing heart diseases at an early stage. It helps the user to know if he/she might be at risk of having a heart disease, so that the user can get the required medical attention and treatment. Also, if any symptoms are similar to COVID-19 symptoms, then such a result is also found.

Medical diagnosis is considered to be an intricate task that needs to be carried out accurately and efficiently. The automation of the same would be very beneficial. Medical decisions are often made based on the doctor's experience and rather than on the knowledge rich data hidden in the database. This practice leads to unwanted errors, biases, and high medical costs which affects the quality of service provided to patients. Data mining has the potential to generate a knowledge-rich environment which can help to significantly enhance the quality of medical decisions.

BLOCK DIAGRAM



Overview:

While designing the model, we have assumed that the user has a clear idea about the symptoms he/she might be experiencing. The Prediction developed considers 52 symptoms related to heart disease and 13 symptoms related to COVID-19 among which the user can give the symptoms as an input.

Once the system is trained with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model, those symptoms are processed according to the rule set developed, thus making classifications and predicting the most likely disease. After taking input for certain attributes like age, blood pressure, heart rate, etc., our prediction system analyses the data and predicts whether there is a possibility for the user to have a heart disease or not in the future.

If the prediction model says yes then the user has to input the symptoms, he/she might be experiencing to know exactly what disease the user is likely to have or might have. It is considered that if any 3 inputted symptoms match to COVID-19 related symptoms as well then, the person might be suffering from COVID-19 and should get diagnosed by a doctor. The necessary precautions for the resulting disease(s) are also mentioned.

PROJECT PLAN / TIMELINE

Sr. No	Task Name	Start	End	Duration (Days)
01	Topic Selection & Approval from Guide	18/02/2021	21/02/2021	04
02	Project Plan Review	22/02/2021	01/03/2020	08
03	Phase 1	02/03/2020	12/03/2020	11
04	Basic Problem Statement Implementation	13/03/2020	20/03/2020	08
05	Basic GUI Module	21/03/2020	25/03/2020	05
06	Combining with Covid-19 Possibility	26/03/2020	05/04/2020	11
07	Phase 2	06/04/2020	13/04/2020	08
08	Complete GUI and Precaution part	14/04/2020	27/05/2020	40
09	Closeout	28/05/2020	30/05/2020	03

IMPLEMENTATION DETAILS

The Project focuses on whether an individual has been diagnosed with heart disease or not. Once the individual has been diagnosed with heart disease. The input for symptoms will be given. With the help of these input symptoms, we will find out which disease the person is suffering from. While predicting the disease if the symptoms show some symptoms, like the Covid -19 symptoms. The system will result in the possibility of Covid-19 and the precautions for the predicted disease, which will help the person to cure the disease.

Data Collection and Processing-

The Dataset for the implementation of this system was taken from the UCI repository. The Dataset includes 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise, number of major vessels, and target ranging from 0 to 1, where 0 is an absence of heart disease and 1 is of the presence of heart disease.

Sr.no	Attributes	Description	Values
1	Age	Patient's age in years	Continuous value
2	Sex	Sex of Patient	1: Male 0: Female
3	Cp	Chest Pain	Value1: typical angina Value2: atypical angina Value3: non- angina pain Value4: asymptomatic
4	Trestbps	Resting blood Pressure	Continuous value in mg/Hg
5	Chol	Serum cholesterol in mg/dl	Continuous value in mg/dl
6	Fbs	Fasting Blood sugar	1>= 120 mg/dl 0<= 120 mg/dl

7	Restcg	Resting electrocardiographic results	0: normal 1: having ST_T wave abnormal 2: Left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value
9	Exang	Exercise induced angina	1: Yes 0: No
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	Slope	Slope of the peak exercise ST segment	1: Unslopping 2: Flat 3: Down sloping
12	Ca	Number of major vessels colored by fluoroscopy	0-3 value
13	Thal	Defect type	3: normal 6: fixed defect 7: reversible defect
14	Target	Diagnosis of heart disease	No_heart disease Have heart disease

While creating the machine learning model, to predict whether the person has heart disease, we used the target column as the class and other columns as the features for the model. The data is divided into training sets and test sets in which 80% data is used for training and 20% for testing.

The Machine learning algorithm used for analysis are

- Support Vector Machine
- Random Forest
- AdaBoost
- Gradient Boosting

From the above Algorithms, **Gradient Boosting algorithm** gives the best result.

Features are the inputs that are given to the ML algorithm, which will be used to calculate an output value. Mathematically, the dataset is the variable used to resolve the equation. The other part of the equation is the target, which are the classes the instances will be categorized into. Because the target contains the target values for the ML classifier. While training a classifier, it is needed to sort the data into two parts: training and testing sets. The training set will have targets, while the testing set won't

contain these values.

```
classification_report :
              precision    recall  f1-score   support

           1       0.79      0.73      0.76        30
           2       0.69      0.75      0.72        24

    accuracy                           0.74        54
   macro avg       0.74      0.74      0.74        54
weighted avg       0.74      0.74      0.74        54


confusion_matrix :
[[22  8]
 [ 6 18]]
```

The coding portion was carried out to prepare the data, visualize it, pre-process it, build the model and then evaluate it. The experiments and all the model building are done based on python libraries.

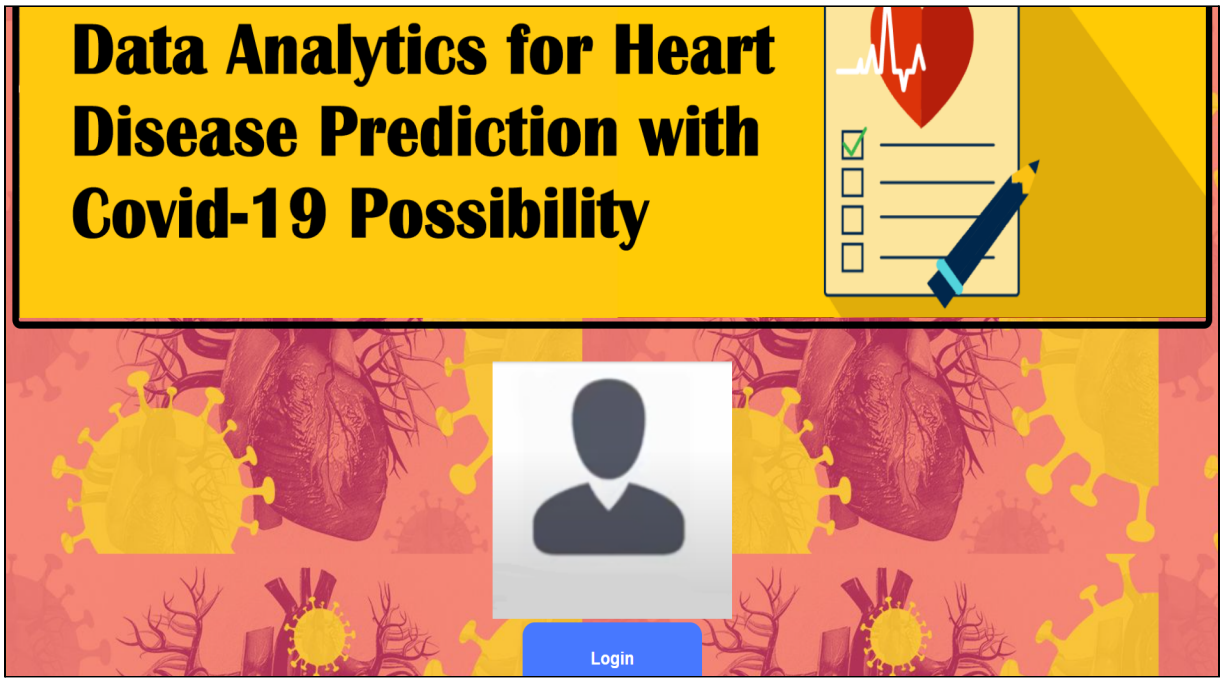
Libraries used:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import re

from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, make_scorer

from plotly.offline import iplot
import plotly as py
import plotly.tools as tls

import pickle
```

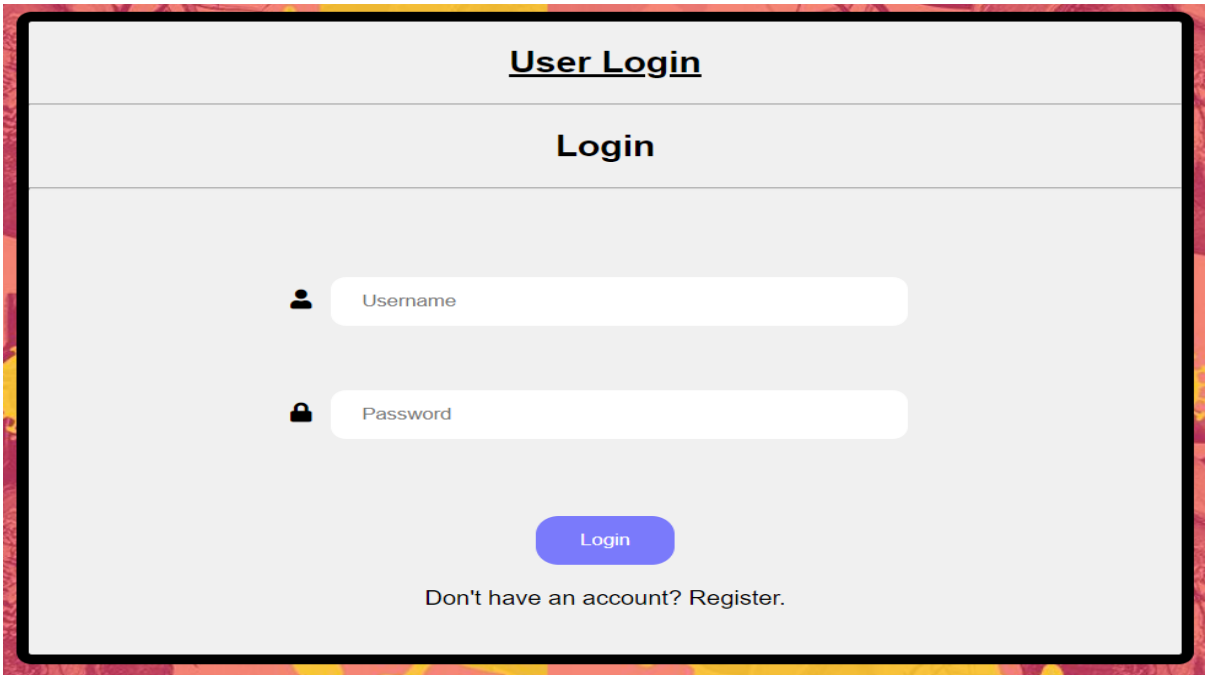
Model Deployment:

The web application is built using the Flask web application framework. We created the web app code (API) to load the model and get user input from the HTML template and made the prediction to return the result.

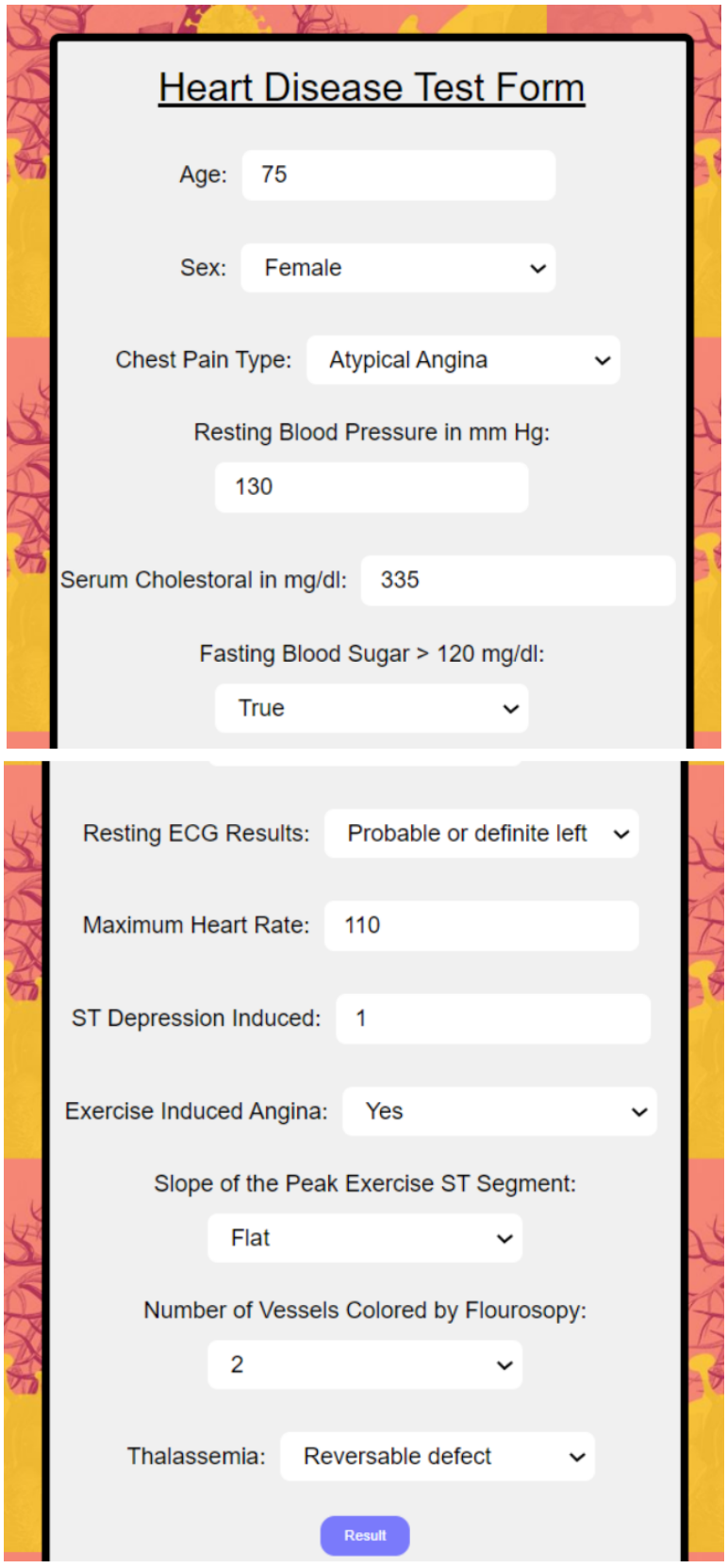
SYSTEM IMPLEMENTATION



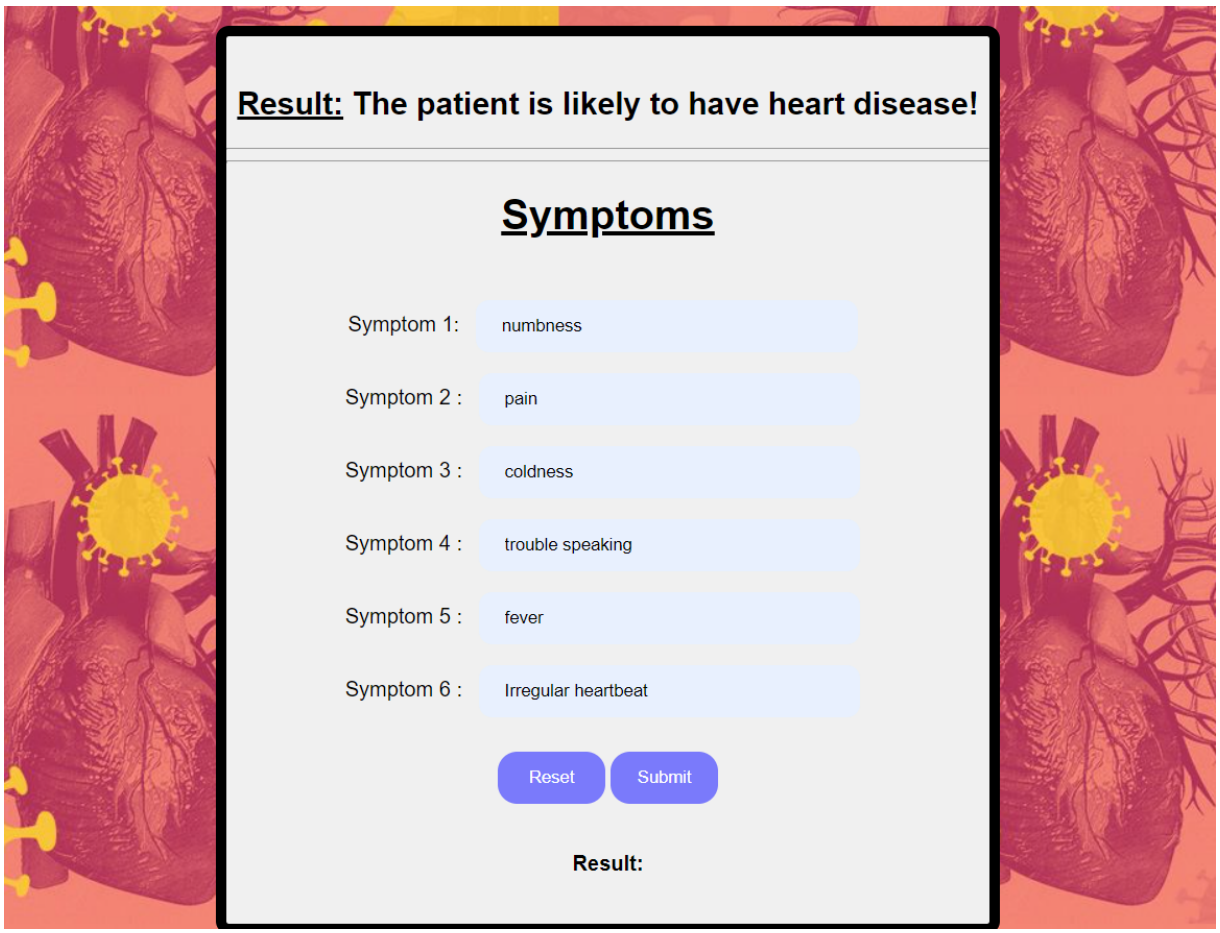
Home Page



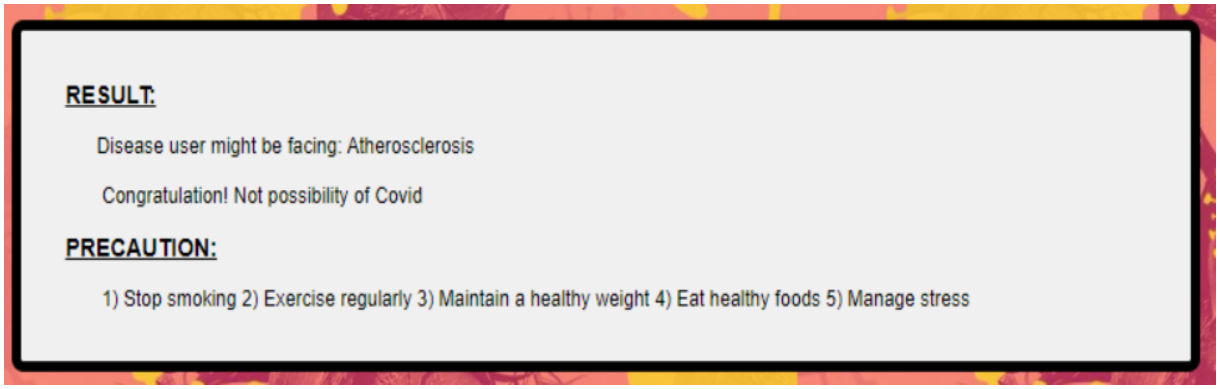
Login Page



Input for predicting the heart disease



Input form for the Symptoms if a person is suffering from heart disease.



The Possibility of covid-19 and Disease name and with Precautions.

CONCLUSION & FUTURE WORK

An efficient system is created which integrates various aspects of Machine learning and produces an accurate result. The algorithm used to predict the analysis helps to give dependable and precise results in the field of health maintenance. The maximum accuracy is achieved through the Gradient Boosting algorithm in terms of Performance.

The Proposed system implementation is user-friendly, GUI-Based, easy to use, reliable, and scalable. The proposed system implementation working model reduces the medical test, treatment costs with the help of providing initial diagnostics in time. The system can also effectively act as a training tool for medical students and can be a lenient diagnostic tool that can be made available for cardiologists, physicians, etc. There can be possible development further by exploring and improving the scalability and accuracy of this predicting system. As we have developed a generalized system where we can identify the heart disease name and the possibility of covid-19 with the precautionary measures, in the future we can try making a new combine dataset for the heart disease and covid-19 and add up all the types of heart-related disease and predict the accuracy for the covid -19 possibility. The performance of the health diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. Due to time limitation, the following research/work needs to be performed for the future. Would like to make use of testing different discretization techniques, multiple classifiers voting techniques and different decision tree types. Similar prediction systems can be built for various other chronic or fatal diseases like Cancer, Diabetes, etc. with the help of recent technologies like machine learning, image processing, and many others.

REFERENCES

1. Kumari Deepika, Dr. S. Seema; "Predictive Analytics to Prevent and Control Chronic Diseases", April 27, 2017 [Online].
   Available: https://doi.org/10.1109/ICATCCT.2016.7912028

2. Prof. (Dr.) Kanak Saxena, Richa Sharma; "Efficient heart disease prediction system using decision tree", July 06, 2015 [Online].
   Available: https://doi.org/10.1109/CCAA.2015.7148346

3. Santhana Krishnan J.,Geetha S; "Prediction of Heart Disease Using Machine Learning Algorithms", June 20, 2019 [Online].
   Available: https://doi.org/10.1109/ICIICT1.2019.8741465

4. Zeeshan Mulla; "Heart Disease Dataset (CSV)"
   Available: https://www.kaggle.com/zeeshanmulla/heart-disease-dataset