Heart Disease Prediction

using

Machine Learning

Contents

Declaration

Certificate

Acknowledgement

Abstract

We are living in the modern Technological era, and we are trying to solve the problem using technology for that we are striving for more knowledge with help of our technology, the more our lifestyle is changing. In the fast growing world everyone is neglecting their Health.Result of this Health diseases are increasing day by day due to our lifestyle and other hereditary reasons. Especially Heart Diseases have become a major issue these days.WHO report that over 11 million deaths are caused each year worldwide due to heart related complications. Using this we can understand that the condition is very serious.There is an urgent need for some technological solutions which can help in educating people and early diagnosis of these disease. This requires a string infrastructure including a large enough work force, which is a slow process. In the recent year machine learning has evolved a lot .To resolve this problem we can use Machine Learning that can help us in this. Sudden increase in the number of patients depict the need for scalability in our medical system. While work force may be limited, we can look for algorithmic solutions for screening and diagnosis stages, due to their correlation with Our project "Heart Health Classification and Early Diagnosis of Heart Disease" is a system that uses predictive capability of machine learning models based on similar data points collected in past. We are going to use the previous medical report data to predict the risk of heart disease.Using our system user can estimate the chance of being suffer from heart disease.Our system takes input data and runs 5 Machine Learning Models on the data and gives the result. We are trying to provide the user friendly platform to user that can be access remotely to screening facilities that will predict the risk of heart disease and that will eliminate the load on the medical system.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Increasing population, changing life patterns, new diseases, pandemics, and change in the natural environment has surely changed the way we take care of our health. Maintaining good health is a major factor associated with any individual's efficiency and It has always been a big concern for all the countries of the world. Providing a good medical service across the whole country proves to be challenging for developing and poor countries.

The condition of medical systems across the world is different for different countries. The ratio of healthcare workers(doctors especially) to the population is a good way to determine the strength of any healthcare system. Developed countries (the USA and Western European) have this ratio to a good level (~25:10000), for Southeast Asian countries like India are way below the world average (India has a ratio of 12.21 to the 10000, whereas the world average is closer to 10:1000) [1]

Medical doctors (per 10,000)

FILTERS

EXPORT DATA in CSV format:
Right-click here & Save link

Last updated: 2022-01-24

Location	Medical doctors (per 10,000)	Medical doctors (number)	Generalist medical practitioners (number)	Specialist medical practitioners (number)	Medic
Thailand	2.24	12,713	17,199	13,492	12,71
Timor-Leste	6.17	712	345	33	367
Togo	0.45	244	219	264	244
Tonga	3.03	100	36	12	30
Trinidad and Tobago	10.06	1,038	4,595	1,098	1,038
Tunisia	10.24	10,554	6,491	7,755	10,55
Turkey	10.22	100,853	10,542	101,998	10,05
Turkmenistan	22.13	11,365	4,156	6,851	20,03
Tuvalu	10.31	10	11	3	0
Uganda	0.82	17,186	17,007	179	2,209
Ukraine	29.92	134,986	12,995	118,737	
United Arab Emirates	13.3	11,630			11,63
United Kingdom of Great Britain and Northern Ireland	16.2	101,803	29,220	101,214	185,6
United Republic of Tanzania	0.23	1,481	2,434	451	1,481
United States of America	24.39	525,070			525,0
Uruguay	37.23	12,384	5,021	8,524	12,38
Uzbekistan	23.74	65,805	10,965	57,279	65,80
Vanuatu	1.14	20			20
Venezuela (Bolivarian Republic of)	17.3	48,000			48,00
Viet Nam	5.24	42,327			42,32
Yemen	2.39	13,560	5,412		3,814
Zambia	0.53	1,499	1,701	325	1,499
Zimbabwe	0.54	1,054	1,002	227	1,054

Fig 1.1 Ratio of Doctors per 10000 population in the USA

Medical doctors (per 10,000)

FILTERS

EXPORT DATA in CSV format:
Right-click here & Save link

Last updated: 2022-01-24

Location	Medical doctors (per 10,000)	Medical doctors (number)	Generalist medical practitioners (number)	Specialist medical practitioners (number)	Medic
Guinea	0.83	1,844	2,649	138	1,844
Guinea-Bissau	0.54	100	197	10	0
Guyana	14.24	1,120			1,120
Haiti	0.85	1,392			1,392
Honduras	2.95	2,454	1,241	1,213	2,680
Hungary	26.8	20,877	6,855	25,615	15,356
Iceland	28.47	1,029	156	692	1,029
India	12.21	1,014,538			1,014,
Indonesia	0.58	11,067	137,920	16,597	0
Iran (Islamic Republic of)	11.29	116,536	51,974	39,127	116,53
Iraq	10.36	19,738	24,586	12,807	1,472
Ireland	15.52	10,270	10,781	2,524	10,270
Israel	32.31	10,140	8,609	13,856	10,140
Italy	34.59	148,101	48,230	147,845	148,10
Jamaica	3.54	1,006	1,044	280	1,103
Japan	16.39	113,214			101,02
Jordan	18.69	10,627			10,627
Kazakhstan	32.8	45,400	1,818	42,480	69,722
Kenya	1.34	4,506	5,602	2,440	4,506
Kiribati	1.43	15			15
Kuwait	14.54	10,000			10,000
Kyrgyzstan	22.13	10,609			10,609
Lao People's Democratic Republic	1.8	1,160			1,160

Fig 1.2 Ratio of Doctors per 10000 population in India

Increasing this ratio of doctors to patients is a complex process and takes a lot of time in order of years or maybe decades and medical buildings equipped with modern machinery are needed.

It is also a general foreseen that people avoid going to doctors for preliminary stages of their diseases and other risks. Designing a suitable system for them so that they can pre-screen themselves can be very helpful hence A basic solution is to design a good scalable system here. Scalability here means that the technological stacks used can expand the scope of health care without needing more workforce i.e. Use software services.

Software services generally are beneficial for the fact that they can detect Early Disease and their Speed & Precision in Medical Diagnostics ,enabling the Provision of Quality Care. It also promotes Improved Patient Participation & Management and improves the chances of safer Treatment Solutions.

A good scalable software service ensures that in case of larger participation need, it can take release loads from its concerned department. Since we know that building smarter and better infrastructure with better equipment, training more workforce, etc. are time taking processes. Once made available, they cannot be rolled back and are limited but they are permanent. They obviously increase the capacity of the medical system but don't contribute to increasing

scalability at the current level which is the need of time. Hence, we are looking at their non-biological counterparts, i.e., computers, algorithms, to help us.

If we consider any diagnosis of a patient, then it is greatly affected by experience, knowledge of its doctor. Some decisions are so intense that it may take suggestions from a few other trained and experienced people of the concerned area to arrive at a decision, it is the stages that need huge amounts of experience with little analysis. It can be solved easily by making use of computer software and various algorithms. The software is able to do a lot of critical tasks in such conditions such as arriving at a suitable decision, diagnosis of disease, early screening of disease based on previous medical record, etc. The reason for this is that scaling-up only needs more computing power which is much cheaper and available than training humans for the same, and can be deployed much more quickly. Before all, a screening phase should exist which can be predicted just by asking simple questions, simple symptoms, if they exist etc.

1.2 Medical Records

Patient records make effective healthcare possible. Professional only after analysing the medical record of an individual can assign any test or treatment. But getting all the results from the medical tests in itself can be costly, it also takes time and costs money, which is a major factor for not taking tests in poor countries or developing countries. They also put an additional load on the medical system from taking samples to using machines. The reason for this complication is also that all the people who are not affected by the disease/ complications will also take tests just to be sure whether they are diseased or not. Obviously They will not be referred for treatment, but the tests they have to go through, the time spent on them, etc. contributes nothing to those who are actually suffering.

In the meantime, part of the resources to be spent on victims will be wasted. The identification of disease could be a bottleneck in the health care system in many cases

1.3 Heart Disease

Heart Disease or cardiac Disease is a broad phrase that refers to any faults or anomalies of the heart. It is a medical condition which is generally associated with heart muscle disease which is also called cardiomyopathy, heart attacks, coronary/peripheral-artery diseases, irregular heartbeats which is also called arrhythmia, congenital heart defects, etc.Although heart disease

can manifest itself in a variety of ways, there are a number of common risk factors that impact whether or not someone is at risk for heart disease.

As further symptoms are considered, they can include:[2]

- Chest pain, chest tightness, chest pressure and discomfort in chest (angina)
- Shortness of breath even during light activities.
- If the blood vessels in your legs or arms are narrowed, you may feel pain, numbness, weakness, or coldness.
- During feedings, infants may feel shortness of breath, resulting in poor weight gain and health ailments.
- Continuous severe discomfort in the neck, throat, and back of the abdomen
- Excess of adipose tissue, high blood pressure, and high cholesterol levels.

These all symptoms can easily be correlated well with increased risk of heart diseases. If we are able to maintain a healthy lifestyle like exercising regularly, eating healthy food, and taking adequate sleep, the risk can be minimized. It means that many of the conditions mentioned here are in turn predictable and are based on type of lifestyle. These questions would be part of a full examination to assist in the process of diagnosis heart originated bodily malfunctioning.

According to World Health Organization data [3], an estimated 17.9 million individuals died from CVDs in 2016, accounting for 31% of all global deaths. Heart attacks and strokes killed 4 out of every 5 people who died from CVDs, and over a third of those who died were under the age of 70. In India, NCDs were responsible for 63% of all deaths, with CVDs accounting for 27% of all deaths. CVDs were also involved with 45 percent of deaths among people aged 40 to 69. A survey revealed that the United States is the worst affected country by heart disease where the ratio of heart disease patients is very high [4].Itcost about $363 billion for the United States of America each year from 2016 to 2017 to treat just heart diseases.

If we talk about less developed countries then, CVDs are responsible for at least three fourth of the world's deaths in such countries. People living in low-income nations do not have easy access to primary health care programmes, nor do they have adequate medical infrastructure or machines for early identification and treatment of people with risk factors for cardiovascular disease or other less detectable ailments. As a result, identification of such diseases is typically

late in the disease's progression, and people die of cardiovascular diseases and other noncommunicable diseases at a younger age.

The poorest individuals in low- and middle-income countries are the ones who bear the burden of the consequences. CVDs and other noncommunicable diseases are a major contributor to poverty at the household level, since they have a lower income and have catastrophic health spending and high out-of-pocket expenses. CVDs wreak havoc on the economies of low- and middle-income countries on a macroeconomic basis. [5]

1.4 Project Description

Upon investigating the need of medical professionals in current scenarios, which seems to be a long aim to achieve and deadliness and seriousness of heart disease this project is developed. The overall objective of this project is to be able to predict the presence or absence of heart disease by utilizing few important medical and common attributes. Attributes considered form the primary basis for tests and give accurate results more or less.

- We have created a user interactive machine learning model along with minimal frontend for interaction with the user, so that based on his/her previous medical records, he/she can check whether they are suffering from heart disease or not.
- Models made are trained with various machine learning algorithms and front end integratedwith the help of flask. Various algorithms used here are measured for their accuracy and suitable algorithms are used in the end project. Our basic aim will be to increase the accuracy of models which are already existing.

Chapter 2

Literature Review

Sushmita Roy Tithi et al discussed about ECG data analysis and heart disease prediction using machine learning algorithms. [6]

In this paper they have used 6 supervised machine learning algorithms to distinguish between normal and abnormal ECG. also they wanted to find a particular disease. They divided there dataset into 2 parts 75% for training and rest 25% for testing.They used- ECG, Machine Learning, Logistic Regression, Decision Tree, Nearest Neighbour, Naïve Bayes , Support Vector Machine, Artificial Neural Network, Right bundle branch block, Myocardial infarction, Sinus tachycardia, Sinus Bradycardia, Coronary Artery disease, Abnormal ECG.

ECG provides us with series of sinus rhythm which defines the condition of heart. Used to detect certain kind of diseases.

Disease Name	Best Algorithm	Score
CAD	Naïve Bayes	94%
Sinus Bradycardia	Decision Tree	95%
Sinus tachycardia	All except NN	95%
Myocardial infarction	Decision Tree	96%

Right bundle branch block	Logistic Regression	96%

Table 2.1 Best Algorithms for Various Disease

Bo Jin et al discussed about predicting the Risk of Heart Failure with EHR Sequential Data Modeling.[7]

Their aim was to provide heart patients an early diagnosis and treatments. Because now a heart failure is really common among people age of 65 , overweight people and those with previous heart attack. This paper develops a new approach to this vital task using and enhanced long short- term memory networks (LSTM) method and a data-driven framework. In this paper they proposed a novel method for diagnosis event modeling that includes one-hit encoding and word vectors and employs LSTM approach for this. This paper used electronic health record (EHR) data from real-world databases regarding congestive heart disease. Dataset had 2 parts A- diagnostic records of 5000 patients who have been diagnosed with heart failure. B- diagnostic records for 15000 patients who have not been diagnosed.

Yar Muhammad et al discussed about Early and accurate detection and diagnosis of heart disease using intelligent computational model.[8]

Early and accurate detection and diagnosis of heart disease using intelligent computational model. They used two datasets that are Cleveland (s1) and Hungarian (s2) heart disease datasets. Ten classification algos were used that include KNN, DT, RF, NB, SVM, AB,ET,GB,LR and ANN and 4 feature selection algos that are FCBF, mRMR, LASSO and relief. The top two accuracies of classification algos were ET and GB with 92.09% and 91.34% respectively. So the ET classifier with relief feature selection algo performs Excellently.

Ashir Javeed et al discussed about an Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection".[9]

They designed an intelligent learning system based on random search algorithm and optimised random forest model for improved heart disease detection. This paper used random search algorithm for factor selection and random forest model for diagnosing the cardiovascular disease. They highlighted the problem of overfitting in models and proposed a novel learning system. System combined two algos i.e. random search algorithm and random forest. And in the end they could improve performance of random forest by 3.3%.

Based on the research above, we will be using 13 attributes:

Age: The probability of heart related issues after 60 years of age is very high, so it's important to include age in the heart disease prediction model.Over 80% of heart disease deaths occur in people over the age of 65, according to empirical estimates.

Sex: Cardiovascular risk factors are more pronounced in women with clinically manifest heart disease than in men.Females are more likely than males to have their first AMI after smoking.

Resting Blood Pressure : (trestbps, Integer, in mmHg)An increased resting heart rate was related to heart disease in both sexes in an annotated study, even after controlling for confounders such as abdominal obesity and general obesity.Resting blood pressure is one of several factors contributing to the risk of heart disease, so it is important to control it.

Angina : Induced during exercise (exang, 0 : no : yes) Reduced blood flow to the heart causes this type of chest pain.The condition is associated with coronary artery disease.Angina pectoris is another name for the symptom.Pain in the chest is commonly described as being squeezing, heavy, or tight.

Highest ST-segment (slope, 1: upward slope, 2: flat, 3: downward slope

Resting ECG (aberration in ST/T-wave): Your heart is affected when the electrical impulses produced by other muscles interfere with the heart's electrical impulses. Heart disease patients' ECGs are characterized by naturalness, making them a good indicator to detect the disease.

Classification of Chest-pain:(cp, Integer, myocardial infarction: {1(typical), 2(atypical)} 3: pain, 4: asymptotic) Shown to be closely related to heart related risk factors

Fasting Blood Sugar (fbs, Integer, 1: if fasting blood sugar over 120mg/dl, 0: otherwise)Heart related disease risks and fasting glucose levels tend to follow J shapecurves.A glucose level of 85 to 99 mg/dL carries the lowest risk.The risk of heart disease, ischemic heart disease, myocardial infarction, and thrombotic stroke increased progressively when fasting glucose levels reached more than 110 mg/dL, but not the risk of hemorrhagicstroke.A fasting glucose level below 70 mg per dL was associated with increased stroke risk in 15 of 17 patients (hazard ratio 1.0, 95% CI 1.01-1.11) in men, and 1.10, 1.05-1.18 for women.

Fluoroscopy colored vessels (ca, Integer, 0,1,2,3)

Serum Cholesterol (chol, Integer, in mg/dl) A correlation between serum cholesterol level and coronary heart disease was found to range from 1.3 (95% CI 0.7-2.3) in those with a level between 4.7 and 5.1.However, the difference was less for those with 6.2 mmol/L or more (95 % CI 1.0, 2.7), compared to those who had 4.7 mmol/L or less.In contrast, more active individuals did not experience this.Individuals who participated in lesser physical activity had significantly inverse relative risks for all levels of cholesterol, including cholesterol levels exceeding 6.2 mmol/L (Relative risk = 0.4) (95% CI 0.2, 0.7).

Thalassemia (thal, Integer, defect: {6(non-fixable), 7(reversible)}, otherwise: 3)Your body has less hemoglobin than normal as a result of this inherited blood disorder.Basically, hemoglobin carries oxygen.If you have thalassemia, you will feel fatigued.

Max heart rate achieved (thalach, Integer)Despite controlling for age in the study, the maximum heart rate is correlated with heart failure in the annotated study.Likewise, max heart rate, negatively correlated with access adipose tissue, is an indicator of fitness.So it is a good tool, presumably associated with an inverse relationship between heart disease risk and exercise.

Exercise induced ST depression (oldpeak, Integer)In an ECG, exercise-induced ST segment depression is defined as a horizontal or downsloping ST depression greater than 1.0 mm at 80 ms after the J point or any ST depression of greater than 1.0 mm.

Highest ST-segment (slope, 1: upward slope, 2: flat, 3: downward slope

Chapter 3
Methodology

3.1 Building Machine Learning Models

3.1.1 Data Collection

The dataset which we have used is available on Kaggle website and is available for public download [10]. It is processed from UCI's dataset and contains valid tuples for further processing.

Dataset Characters	Multivariate
Number of Tuples	1025
Number of Attributes	14
Attribute Datatype	Categorical, Integer, Real
Source	Kaggle

Table 3.1 Dataset Detail

Next after downloading datasets is generally to clean the data if some missing values, noise in data is present. First the data is imported from downloaded csv files using pandas libraries of Python, to RAM in data type called data frame which is optimized adequately to handle two dimensional array data. The dataset does not have any null values. 13 of our attributes are the attributes which are used to predict the result, while the last attribute "target" is the result, i.e., whether or not the individual was suffering from heart disease.
The following are the result of the **heartData.info()** and **heartData.describe()** command, which reflects the statistics of Processed dataset.

```
success
```

```
In [7]: heartData=pd.read_csv("./dataset/heart_final_processed_by_Sudhansu.csv");
        heartData.info()
        heartData.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Fig 3.1 Attributes in Heart Dataset with datatype

Out[7]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000

Fig 3.2 Detailed Description of Heart Disease Dataset -1

exang	oldpeak	slope	ca	thal	target
1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 3.3 Detailed Description of Heart Disease Dataset -2

3.1.2 Checking Data Distribution

Following is the categorisation of dataset based on target class:

```
In [19]: rcParams['figure.figsize'] = 8,6
         plt.bar(heartData['target'].unique(), heartData['target'].value_counts(), color = ['black', 'silver'])
         plt.xticks([0, 1])
         plt.xlabel('Target Classes')
         plt.ylabel('Count')
         plt.title('Count of each Target Class')

Out[19]: Text(0.5, 1.0, 'Count of each Target Class')
```

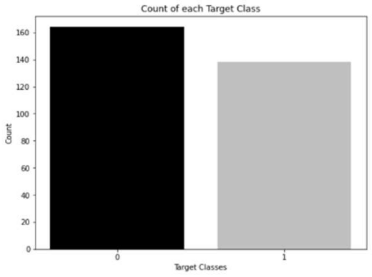


Fig 3.4 Categorization of dataset based on target class

3.1.2 Study of Dataset

We will generate and understand correlation between our attributes and target class. Correlation matrix will be generated and plotted using matplotlib.

For a correlation matrix, the more positive the value of correlation, the more increase in value of one variable causes the other to increase i.e. more directly proportional. The higher a negatively correlated variable gets, the lower the value of the target becomes.

```
corr = heartData.corr()

corr.style.background_gradient(cmap='coolwarm')
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.103240	-0.071966	0.271121	0.219823	0.121243	-0.132696	-0.390227	0.088163	0.208137	-0.169105	0.271551	0.072297	-0.229324
sex	-0.103240	1.000000	-0.041119	-0.078974	-0.198258	0.027200	-0.055117	-0.049365	0.139157	0.084687	-0.026666	0.111729	0.198424	-0.279501
cp	-0.071966	-0.041119	1.000000	0.038177	-0.081641	0.079294	0.043581	0.306839	-0.401513	-0.174733	0.131633	-0.176206	-0.163341	0.434854
trestbps	0.271121	-0.078974	0.038177	1.000000	0.127977	0.181767	-0.123794	-0.039264	0.061197	0.187434	-0.120445	0.104554	0.059276	-0.138772
chol	0.219823	-0.198258	-0.081641	0.127977	1.000000	0.026917	-0.147410	-0.021772	0.067382	0.064880	-0.014248	0.074259	0.100244	-0.099966
fbs	0.121243	0.027200	0.079294	0.181767	0.026917	1.000000	-0.104051	-0.008866	0.049261	0.010859	-0.061902	0.137156	-0.042177	-0.041164
restecg	-0.132696	-0.055117	0.043581	-0.123794	-0.147410	-0.104051	1.000000	0.048411	-0.065606	-0.050114	0.086086	-0.078072	-0.020504	0.134468
thalach	-0.390227	-0.049365	0.306839	-0.039264	-0.021772	-0.008866	0.048411	1.000000	-0.380281	-0.349796	0.395308	-0.207888	-0.098068	0.422895
exang	0.088163	0.139157	-0.401513	0.061197	0.067382	0.049261	-0.065606	-0.380281	1.000000	0.310844	-0.267335	0.107849	0.197201	-0.438029
oldpeak	0.208137	0.084687	-0.174733	0.187434	0.064880	0.010859	-0.050114	-0.349796	0.310844	1.000000	-0.575189	0.221816	0.202672	-0.438441
slope	-0.169105	-0.026666	0.131633	-0.120445	-0.014248	-0.061902	0.086086	0.395308	-0.267335	-0.575189	1.000000	-0.073440	-0.094090	0.345512
ca	0.271551	0.111729	-0.176206	0.104554	0.074259	0.137156	-0.078072	-0.207888	0.107849	0.221816	-0.073440	1.000000	0.149014	-0.382085
thal	0.072297	0.198424	-0.163341	0.059276	0.100244	-0.042177	-0.020504	-0.098068	0.197201	0.202672	-0.094090	0.149014	1.000000	-0.337838
target	-0.229324	-0.279501	0.434854	-0.138772	-0.099966	-0.041164	0.134468	0.422895	-0.438029	-0.438441	0.345512	-0.382085	-0.337838	1.000000

figure 3.5 Maximum positive correlated features is cp and thalach and maximum negative correlated features is exang and old peak.

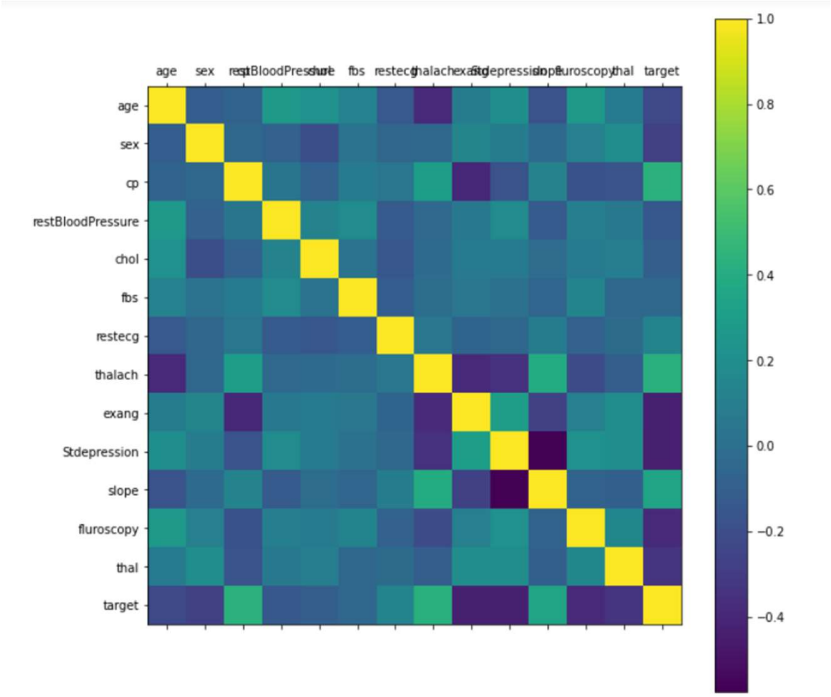


Fig 3.6 Correlation Diagram as heatmap

Using the correlation matrix, we discover the target to be positively correlated to chest pain significantly. This arrival should also be obvious as the greater amount of chest pain means greater risk in the heart. Max Heart rate is also significantly correlated to Target for the reason that healthier hearts do not need to become elevated much for blood supply. It simply means higher heart rate, higher the risk of heart disease. A positive correlation can be observed between those with thalassemia. Since it is a 3 valued ordinal, where 3 indicates normal, 6 to 7 defects. Hence being in the normal category is better. Presence of negative correlation among target and angina can also be observed. This observation also agrees with common sense as exercise causes muscles to crave for more oxygen, in-turn boosting heartbeat, while narrowed-down arteries would act as blockage.

Based upon these studies, selecting features, models are trained and their performances are observed and noted down. The models then are checked against dataset and other values which will predict the presence or absence of heart disease. Various models are mentioned in the next chapter.

3.2 User Interactive Front End

There will be a multipage website containing a homepage, a page for the user in which he enters details to predict the presence or absence of heart disease and a page about us. Flask has been used to connect the frontend with trained models of the backend.

Flask is a Python web framework that was created with a philosophy in mind. Armin Ronacher conceived and developed Flask as an April Fool's Day hoax in 2010. Despite its comedic beginnings, the Flask framework has grown in popularity as a viable alternative to Django projects' monolithic structure and dependencies.

There are some advantages of flask over other frameworks as per requirement of our project. They are,

Flask is a Python web framework built for rapid development of small projects**,**Flask offers a diversified working style while Django offers a Monolithic working style.

3.3 HARDWARE REQUIREMENT

The hardware requirements for running this website and model are:

- RAM – 512 MB
- Operating System – Windows XP/7/8/10/11, MacOS , Ubuntu
- Processor – Intel(R) Core(TM) i3
- Processor speed – 3.60 GHz

3.4 SOFTWARE REQUIREMENTS

The programming language used to develop this application is Python and the IDE used is JupyterNotebook.Front end is made using HTML, CSS and is integrated with flask.

- Programming Language – Python
- Python IDE – Jupyter Notebook
- Python Libraries: Flask

Chapter 4

Experimental Results

The models used are the following:

- Logistic Regression
- K Nearest Neighbours
- Support Vector Machine
- Decision Tree
- Random Forest

4.1 Logistic Regression

This is the one of the most common model used in ML,Logistic Regression is often applied in the actual manufacturing context the fields such as data mining, automatic disease diagnosis and economic prediction.

For our model,we use Logistic regression to know the risk factors for heart disease and forecast the probability of disease occurrence based on risk factors.This model is most frequently applied for classification, primarily two-category issues (that is, there are only two types of output, each representing one category), and can indicate the probability of occurrence of each classification event. Logistic regression model is shown below:

This technique used is also known as sigmoid function .Sigmoid function helps in the easy representation in graphs. Logistic regression also provides better accuracy. By using equation the logistic regression algorithm is represented in the graphs showing the difference between the attributes.

$$prob(Y=1)=\frac{e^{z}}{1+e^{z}}$$

Where Y refers to binary dependent variable (Y is equal to 1 if event happens; Y=0 otherwise), e stands for the foundation of natural logarithms and Z means

$$Z=\beta_0+\beta_1X_1+\beta_2X_2+\ldots\ldots\ldots+\beta_pX_p$$

with constant β0 ,coefficents β j and predictors X j , for p predictors(j=1,2,3,.....,p)

The process of modeling the probability of a discrete outcome given an input variable is known as the Logistic Regression. The most common logistic regression, as its name suggests is not regression rather it is a classification algorithm that classifies something that can take two values such as true/false, yes/no, and so on.Logistic regression identifies a hyperplane in a manner that when it is passes through a function whose value ranges between 0 and 1 (typically we use sigmoidal), it optimizes cost function. Bases on closeness to 0 or 1, it predicts a Boolean output. Here vector parameters is used for training. σ(.) is usually a sigmoid function, with output between 0 and 1.

Features of Logistic Regression:

Multinomial logistic regression is the type of regression which uses the softmax function to compute probabilities.

- We use loss function to learn weights(vector w and bias b) from a labeled training. we perform such activity to minimize the cross-entropy loss.
- Iterative algos like gradient descent are used to get the weight(optimal).while minimizing the loss function the the type of convex optimization problem.
- To avoid overfitting regularization is used.
- Logistic regression has the ability to transparently study the importance of individual features.

Advantages

- This technique is perform well and fast where we have to classify unknown records.
- This is not limited to binary classification we can easily extend it to multinomial regression.

Disadvantage

- Logistic regression will not perform well If the number of observations is lesser than the number of features in such condition it may lead to overfitting.
- Logistic regression constructs the linear boundaries.

In the logistic function equation, *x* is the input variable. Let's **feed in values** −20 to 20 into the logistic function. As illustrated in Figure the inputs have been transferred to between 0 and 1.

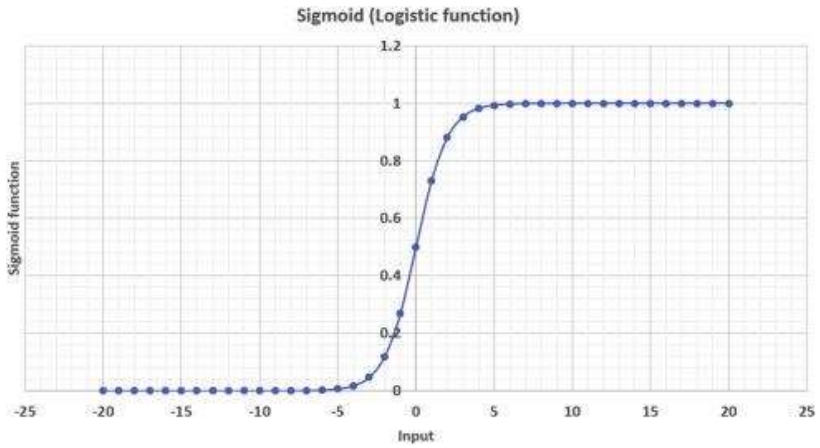


FIgure. 4.1 Sigmoid Graph

```
In [146]: logistic_model = LogisticRegression()
          logistic_model.fit(X_train.values, y_train.values)
          logistic_model_prediction=logistic_model.predict(X_test.values)
          print(accuracy_score(y_test.values,logistic_model_prediction))
          print(classification_report(y_test.values,logistic_model_prediction))

0.8131868131868132
              precision    recall  f1-score   support

           0       0.85      0.71      0.77        41
           1       0.79      0.90      0.84        50

    accuracy                           0.81        91
   macro avg       0.82      0.80      0.81        91
weighted avg       0.82      0.81      0.81        91
```

Figure.4.2 Logistic Model Result

Accuracy: 82%

4.2 **KNN**

K-nearest neighbours is an online processing algorithm having capabilities of both classification and regression. When the input datapoint is provided, it calculates its distance from all the available training data, using any distance metric like Euclidean, Manhattan, etc. It finds k examples closest to the input. Based on majority vote, it predicts the target result for the input. It is nonparametric algorithm. The training is done by passing train set to the function fit () of the object KNeighborsClassifier from sklearn.neighbors. For prediction, we use predict () function of the same object. The value of k, i.e., the number of neighbours has to be experimentally checked. The one with the best results is selected. After increasing the number of neighbors to a certain extent, the accuracy either starts dropping or the increase is negligible. This point of called knee

point. Since the larger number of neighbors means more calculation for evaluation, we tend to choose the knee point, as it provides the minimum number of neighbors that will give good enough results. In practical scenarios, if k turns out to be less than 5, we choose 5, which is also the default value provided by sklearn.

Feature of KNN

The KNN algorithm has the following features:

- KNN is a Supervised Learning algorithm that uses labeled input data set to predict the output of the data points.
- It is one of the most simple Machine learning algorithms and it can be easily implemented for a varied set of problems.
- Feature similarity is the basis of KNN.
- In this algorithm checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

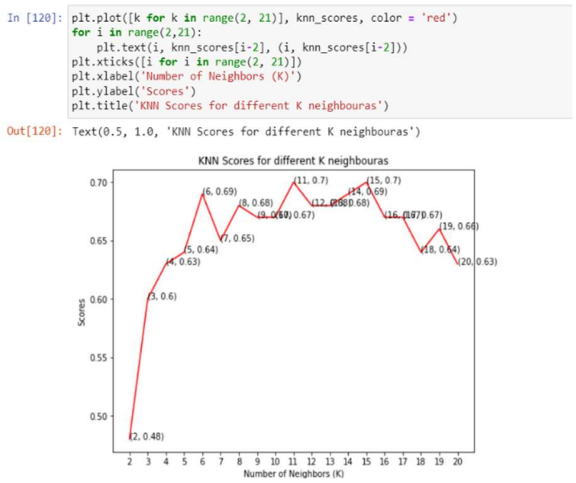


Figure. 4.3 KNN output graph

Since we are getting the best value at k = 11 and k = 15, it was decided in favour of k = 11, which is usually the default value. Finally, prediction is done on test data and the result table is prepared

```
#dated::10/0/2022
knn_scores = []
for k in range(2,21):
    knn_classifier = KNeighborsClassifier(n_neighbors = k)
    knn_classifier.fit(X_train.values, y_train.values)
    knn_score=round(knn_classifier.score(X_test.values, y_test.values),2)
    knn_scores.append(knn_score)


knn_classifier = KNeighborsClassifier(n_neighbors = 5)
knn_classifier.fit(X_train, y_train)
knn_score=knn_classifier.predict(X_test)
print(classification_report(y_test,knn_score))
```

```
              precision    recall  f1-score   support

           0       0.62      0.49      0.55        41
           1       0.64      0.76      0.70        50

    accuracy                           0.64        91
   macro avg       0.63      0.62      0.62        91
weighted avg       0.64      0.64      0.63        91
```

Fig 4.4 Classification report of KNN

Accuracy: 64%

4.3 SVM

It belongs to the supervised machine learning algorithm. This is the type of algorithm that can be used for both classification and regression challenges. SVM is mostly used in classification problems. In the SVM algorithm, we set each data object as a point in the n-dimensional space (where n is the number of attributes you have) by the value of each element which is the value of a particular combination. Then, we do the splitting by finding a hyper-plane that separates the two sections very well. Support Vectors are simply links to individual points.

SVM features:

- Use missing as level
- Include iterations report
- Penalty in the svm we can specify the penalty value.The default value we use is 1.
- Kernel —we can use various type of kernel in SVM.
- Polynomial Degree — The default value is 2.
- Tolerance — specifies the minimum number at which the iteration stops. The default value is 0.000001.
- Maximum iterations — In svm this indicate maximum number of iterations that is allowed with each try. The default value is 25.

```
In [136]:
colors = rainbow(np.linspace(0, 1, len(kernels)))
plt.bar(kernels, svc_scores, color = colors)
for i in range(len(kernels)):
    plt.text(i, svc_scores[i], svc_scores[i])
plt.xlabel('Kernels')
plt.ylabel('Scores')
plt.title('SVM scores Activation function wise...')

Out[136]: Text(0.5, 1.0, 'SVM scores Activation function wise...')
```

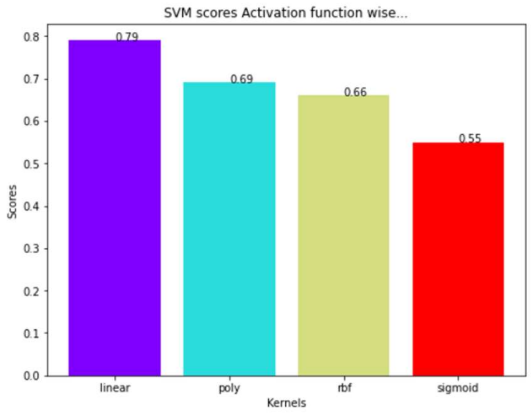


Fig 4.5 SVM Scores against various kernels

Accuracy :79% with linear kernel

KERNEL	ACCURACY
Linear	**79**
Poly	**69**
RBF	**66**
Sigmoid	**55**

Table 4.1 SVM Scores against various kernels

4.4 Decision Tree

A decision tree contain a flowchart-like structure.In the structure of DT(decision Tree) in test on an attribute is represent using internal node.The result of the test is represent using the branch.To represent a class label leaf node is used.To represent classification rules paths from root to leaf is used.

This is the type of analysis in which for A visual and analytical root , closely related influence diagram are used. where the expected values of competing alternatives are calculated. Architecture of Decision Tree.

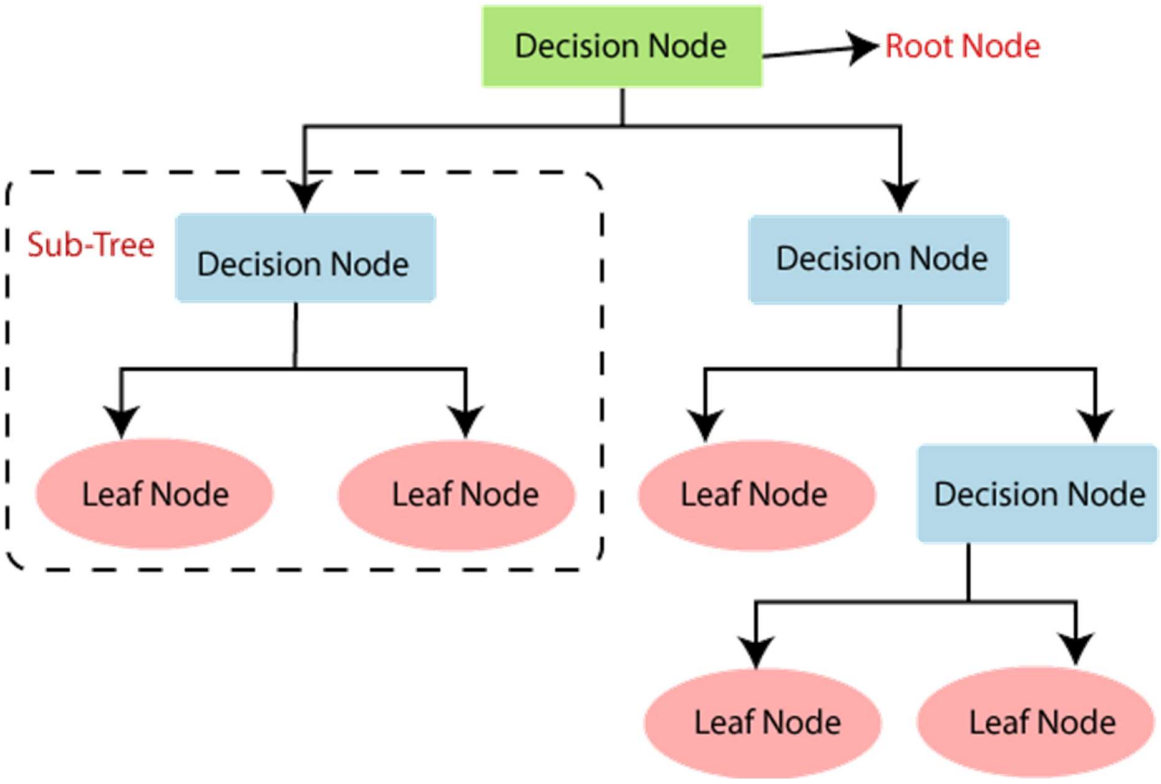


Fig 4.6 Decision Tree image

```
In [142]: plt.plot([i for i in range(1, len(X.columns) + 1)], dt_scores, color = 'green')
          for i in range(1, len(X.columns) + 1):
              plt.text(i, dt_scores[i-1], (i, dt_scores[i-1]))
          plt.xticks([i for i in range(1, len(X.columns) + 1)])
          plt.xlabel('Max features')
          plt.ylabel('Scores')
          plt.title('Decision Tree Classifier scores for different number of maximum features')

Out[142]: Text(0.5, 1.0, 'Decision Tree Classifier scores for different number of maximum features')
```

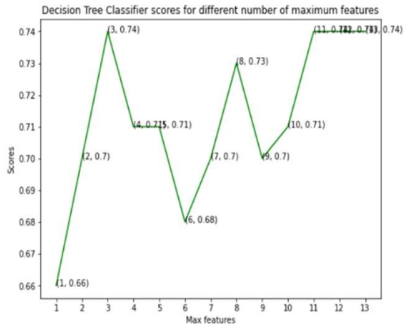


Fig 4.7 Decision tree Result Graph

Accuracy: 74%

4.5 Random Forest Classifier

It is based on the decision Tree.It has a group of various Decision Tree.Various decision trees are used internally.When we try to make a classification for the given Input data we feed same same data into all Decision trees.Now we collect all the votes from Decision trees and majority of votes is going to the result for input.

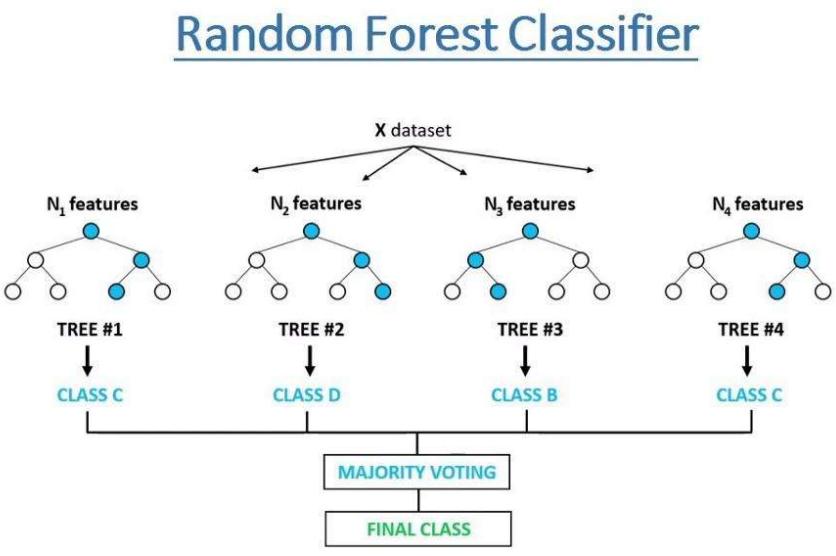


Fig 4.8 Random Forest Image

Random Forest Features

- It runs Efficiently in scenarios where database is very huge.
- We can perform classification on Thousands of Input variables.
- Using this we can find which variable is useful for our classification.
- Using this we can easily calculate the missing data and also maintain the good accuracy.Even though it has missing data.

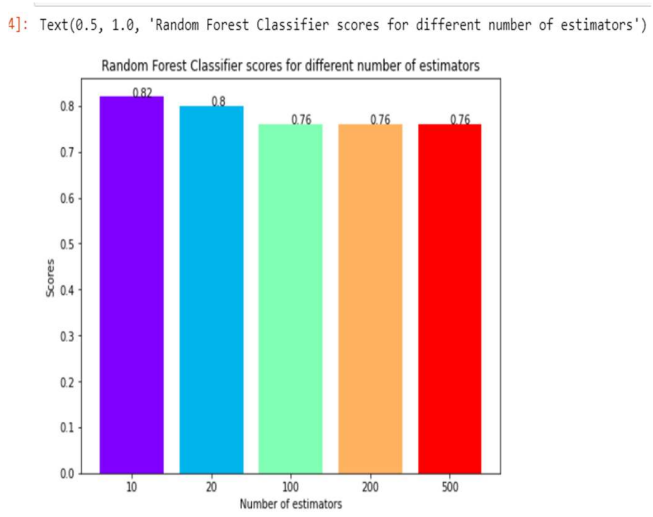


Fig 4.9 Random Forest Result Image

Accuracy:82% with 100 estimators. Increase in number of estimators will increase the calculation complexities significantly.

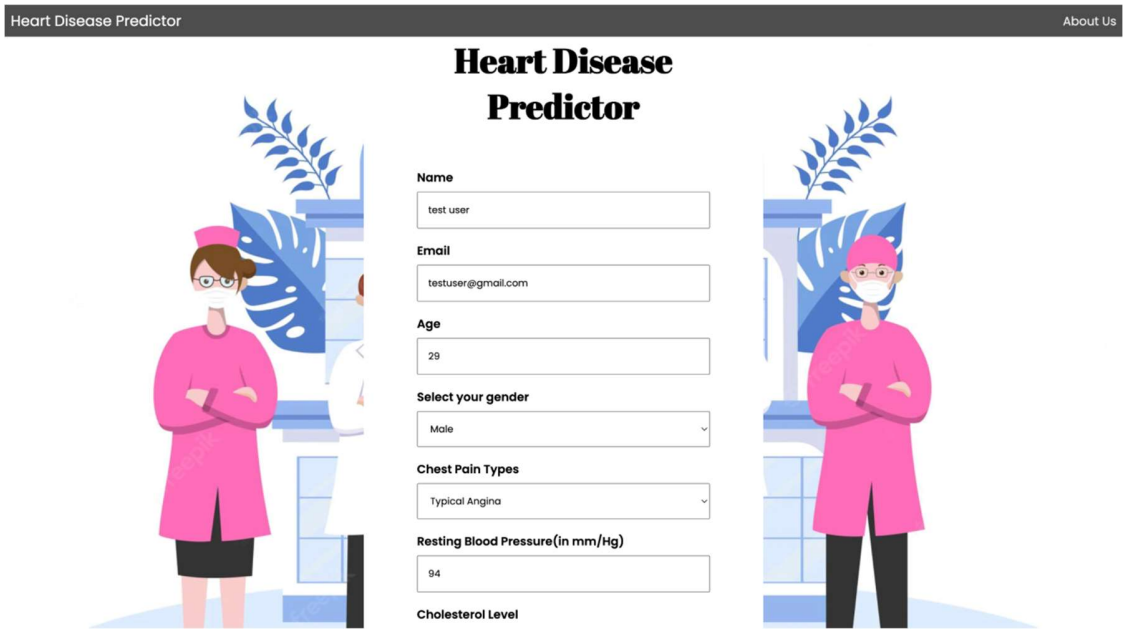
Chapter 5

Conclusions

5. 1 Working Model Diagram

The working front End has a page which contains form where user will enter all the required medical information

The form page screenshot is attached as:



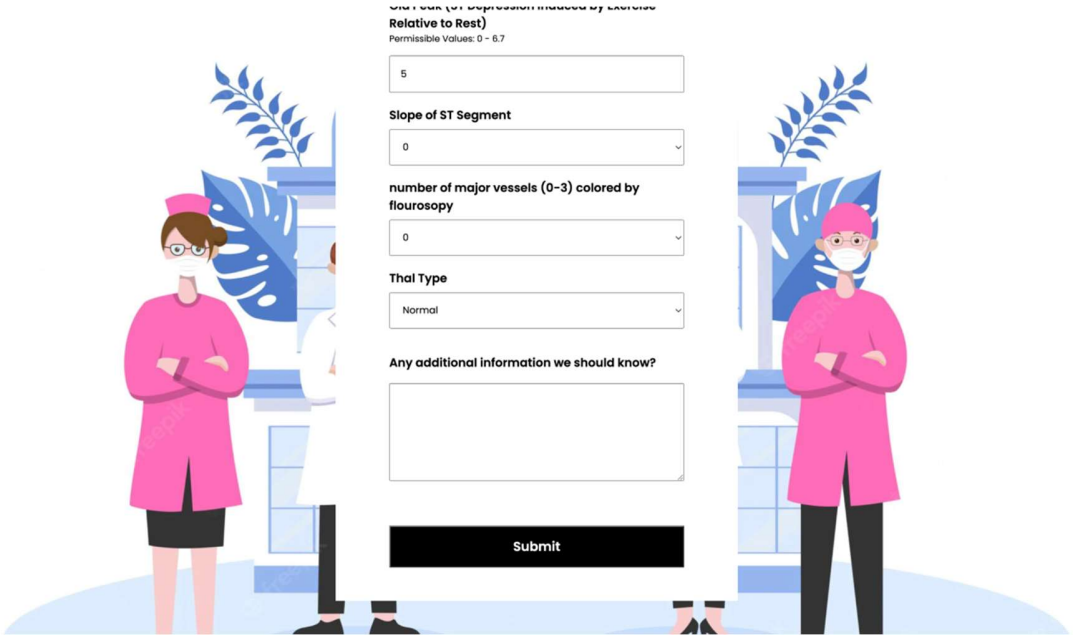


Figure 5.1 Input Form

The result page screenshot:

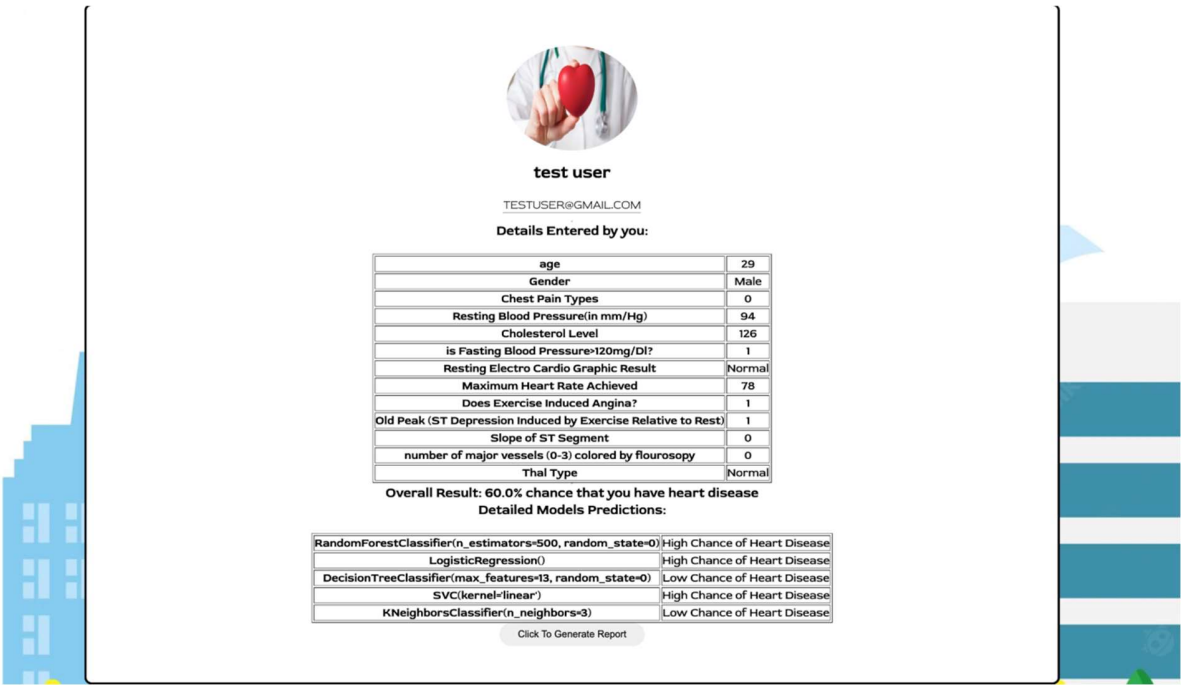


Figure 5.2 Result Page

It also has a button to print the report generated so that user can have a record of data entered by him/her and the respective result.

5.2 Advantages of project

Below are the advantages one can get from this project:

- Anyone can check the presence or absence of heart disease right from their devices only with help of medical records.
- Medical professionals can use this tool to directly determine the results.
- Decrease in engagement of medical professionals for decision taking.
- Project can be launched on large scale and be adapted in every hospitals related to heart diagnosis centres.

5.3 Project Limitations

This project models require 13 attributes for their prediction. If we analyse the attributes closely ,then most of the attributes are not available to any normal person until he/she takes some medical tests which will cost them more money, time and for medical professionals ,equipment.
The attributes required are also more in a medical term than in general language which almost everyone can understand. It would be more friendly and easy for a user if the attributes which require more medical tests, can be decreased significantly and more common attributes which are responsible for heart disease can be included. Some attributes which can work for this are whether the user is a smoker, alcoholic, exercise frequency etc.

5.4 Future Works

Future work for this project can be included as deploying after modifying the parameters. Models can also be trained with reduced parameters .One can increase the accuracy if possible, as there is always a chance of improvement.One can also use other classification algorithms with reduced parameters so that in case of absence of some attributes, users will be able to check ,however with reduced accuracy.

References

1. https://www.who.int/data/gho/data/themes/topics/health-workforce
2. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
3. https://www.who.int/healthtopics/cardiovasculardiseases
4. https://www.cdc.gov/heartdisease/facts.htm
5. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
6. Sushmita Roy Tithi ,AfifaAktar , Fahimul Aleem , Amitabha Chakrabarty ''ECG data analysis and heart disease prediction using machine learning algorithms''. Proceedings of 2019 IEEE Region 10 Symposium
7. Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, XiaomengYin,AndXiaopeng Wei, "Predicting the Risk of Heart Failure WithEHR Sequential Data Modeling" ,IEEE Access Volume 6 2018.
8. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim,Adeeb Noor, Redhwan Nour4, Samad Wali And Abdul Basit ,"An Intelligent Learning System based on Random SearchAlgorithm and Optimized Random Forest Model forImproved Heart Disease Detection" , IEEE Access 2017.This work is licensed under a Creative Commons Attribution 4.0 License Volume 4 2016.
9. Muhammad, Y., Tahir, M., Hayat, M. *et al.* Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci Rep* 10, 19747 (2020). https://doi.org/10.1038/s41598-020-76635-9.[https://www.nature.com/articles/s41598-020-76635-9]
10. **https://www.fullstackpython.com/flask.html**