Visualizing and Predicting Heart Diseases with an Interactive Dash Board

Git Repo - IBM-Project-10542-1659186133

Team ID : PNT2022TMID14797

Batch: B8-2A4E

Team Leader:

Team Lead - Pumenitha S T - 19BEC009 - KCT

Team Members:

Team Member 1 - Geffrey N - 19BEC046 - KCT

Team Member 2 - Karthikeyan P - 19BEC008 - KCT

Team member 3 - Shivani T U - 19BEC012 - KCT

Table of Contents

1.INTRODUCTION

1.1 Project Overview:

The leading cause of death in the developing world is heart disease. Therefore, there needs to be work done to help prevent the risks of having a heart attack or stroke. Our work is to develop an Interactive Dashboard for visualising and prediction the heart disease with the user data.

Now a days, heart disease is spread world wide due to the unhealthy food habits and avoiding regular exercises. People are suffering from heart disease without any knowledge. Hence, an Interactive Dashboard can help people to visualise and prdict the disease in advance and thus take necessary actions to avoid it at the beginning stage.

1.2 Purpose:

An attempt to develop an Interactive Dashboard for the vilsualisation and prediction of heart disease which helps user to take care of their health in an efficient way. This will help the users to predict their diseases in advance and take necessary actions to avoid the disease.

The main motive of the project is to help people to take care of their health in an easy and efficient way by using an Interactive Dashboard. Users can able to visualise and predict the disease without wasting much time and money

2. LITERATURE SURVEY

2.1 Existing Problem :

Healthcare industries generate enormous amount of data, so called big data that accommodates hidden knowledge or pattern for decision making. The huge volume of data is used to make decision which is more accurate than intuition. Exploratory Data Analysis (EDA) detects mistakes, finds appropriate data, checks assumptions and determines the correlation among the explanatory variables. In the context, EDA is considered as analysing data that excludes inferences and statistical modelling. Analytics is an essential technique for any profession as it forecast the future and hidden pattern. Data analytics is considered as a cost effective technology in the recent past and it plays an essential role in healthcare which includes new research findings, emergency situations and outbreaks of disease. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analysing data.

2.2 References:

“Heart Disease Prediction using Exploratory Data Analysis” R. Indrakumari, T.Poongodi, Soumya Ranjan Jena

In this paper, the risk factors that causes heart disease is considered and predicted using K-means algorithm and the analysis is carried out using a publicly available data for heart disease. The dataset holds 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain. To predict the heart

disease, Kmeans clustering algorithm is used along with data analytics and visualization tool. The paper discusses the pre-processing methods, classifier performances and evaluation metrics. In the result section, the visualized data shows that the prediction is accurate.

Prediction of heart disease at early stage using data mining and big data analytics: A survey N. K. Salma Banu, Suma Swamy
Several studies have been carried out for developing prediction model using individual technique and also by combining two or more techniques. This paper provides a quick and easy review and understanding of available prediction models using data mining from 2004 to 2016. The comparison shows the accuracy level of each model given by different researchers. Into practice.

2.3 Problem Statement Definition

Who does the problem affect?
People with unhealthy lifestyles, stress, depression, age above 40 and when their ancestors got heart disease (since heart disease is hereditary).

When does the issue occur?
The issue occurs for people with unhealthy lifestyles and age above 40.

Where is the issue occurring?

The issue is originating from an unhealthy lifestyle. It mostly occurs in the blood valves of the heart.

What would happen if we didn't solve the problem?

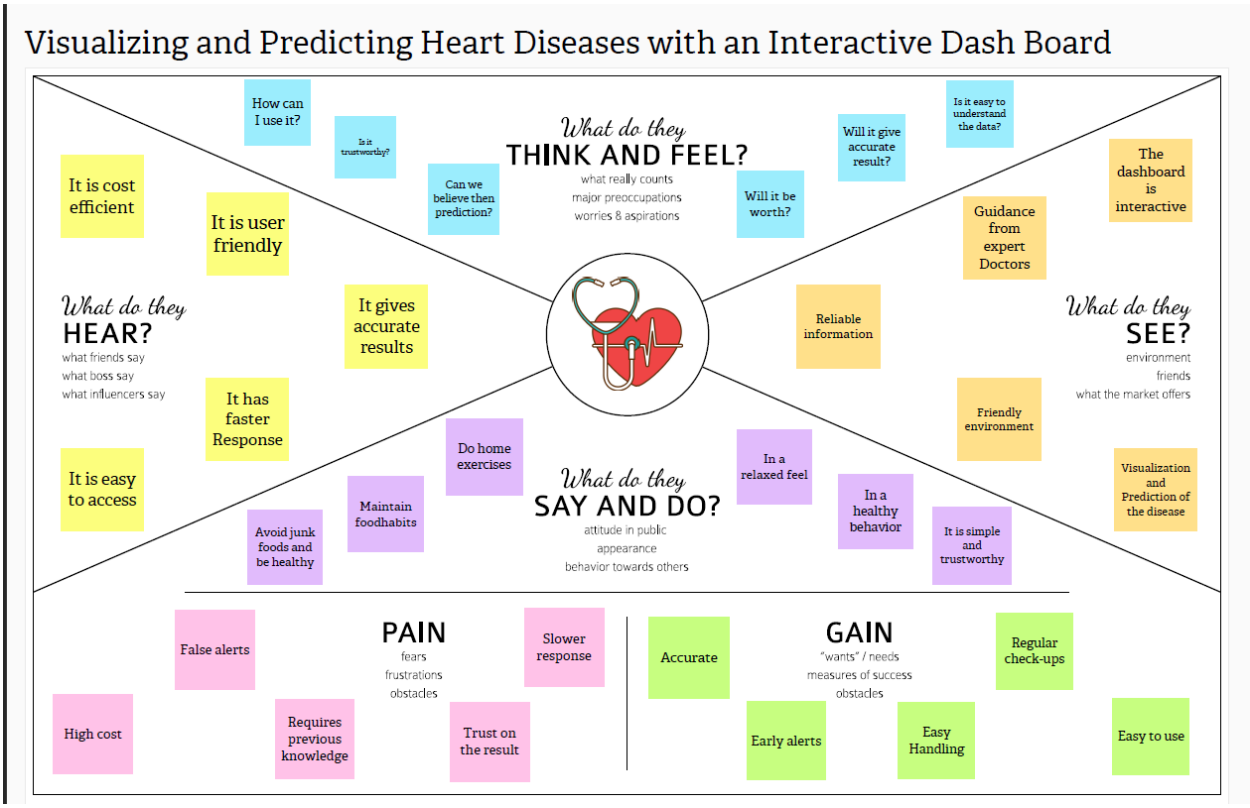
If we don't solve the problem, many people will die at a young age. The death rate due to heart disease will increase rapidly.

Why is it important to fix the problem?

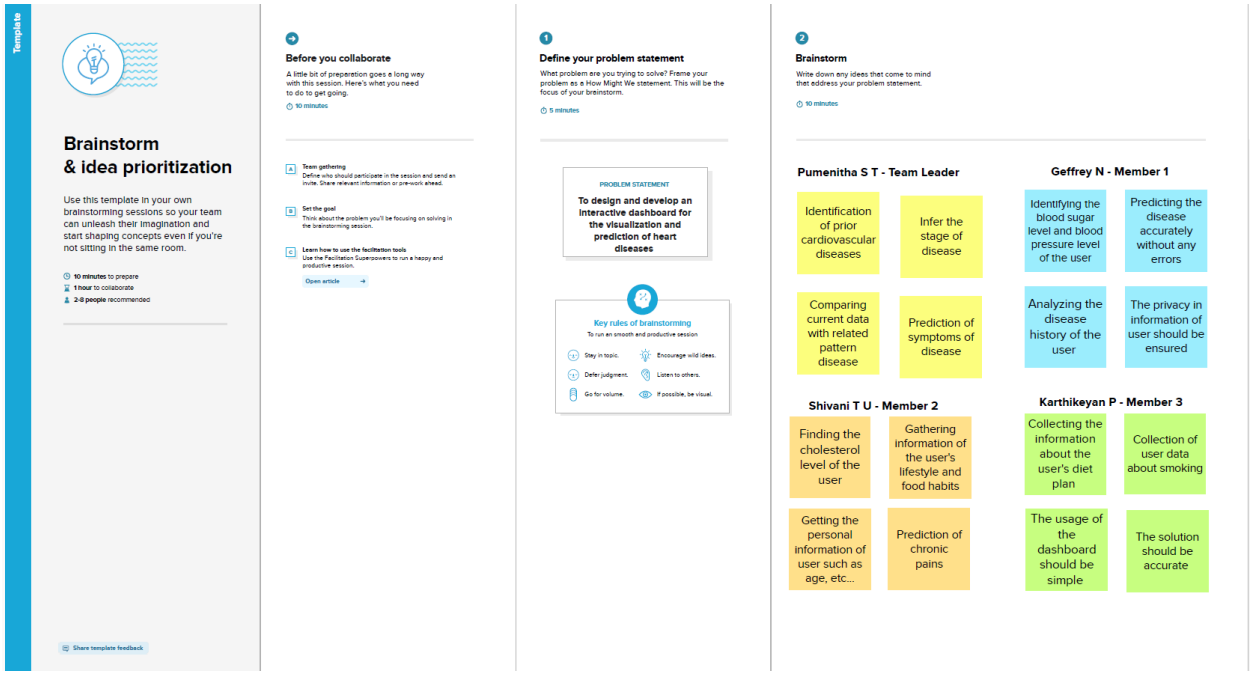
We should predict the problem before giving treatment to the patients. As the problem is predicted early, we can solve it easily and early.

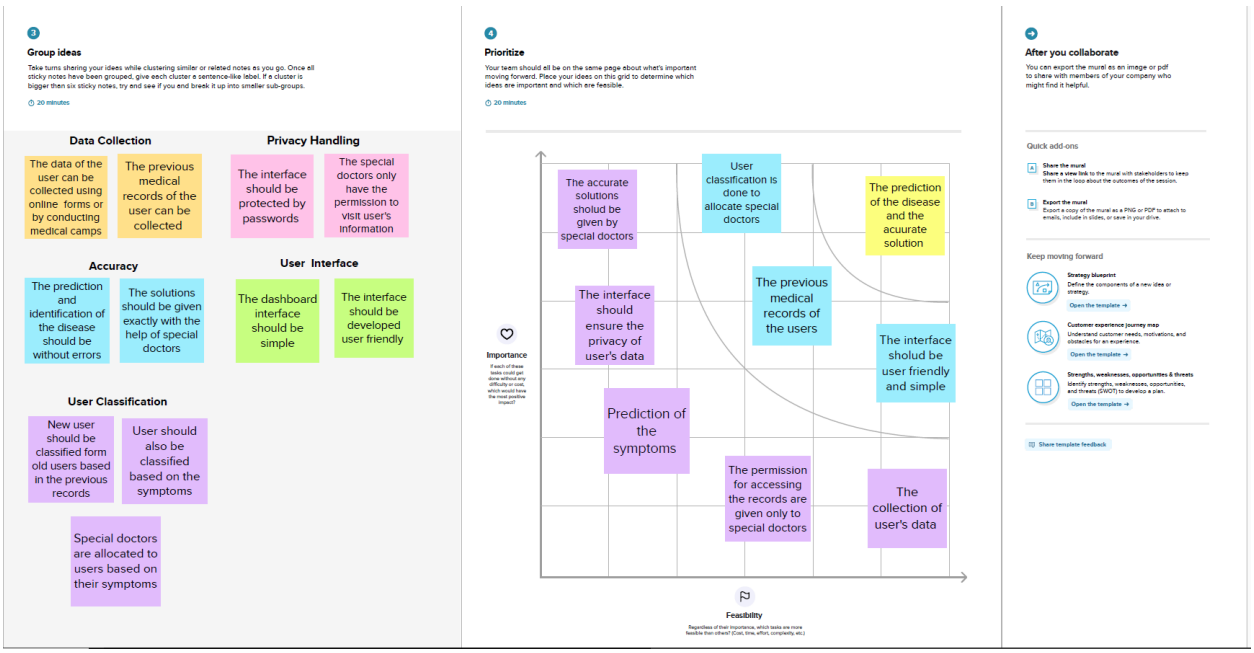
3. IDEATION AND PROPOSED SOLUTION

3.1 Empathy Map:



3.2 Ideation and Brainstorming:



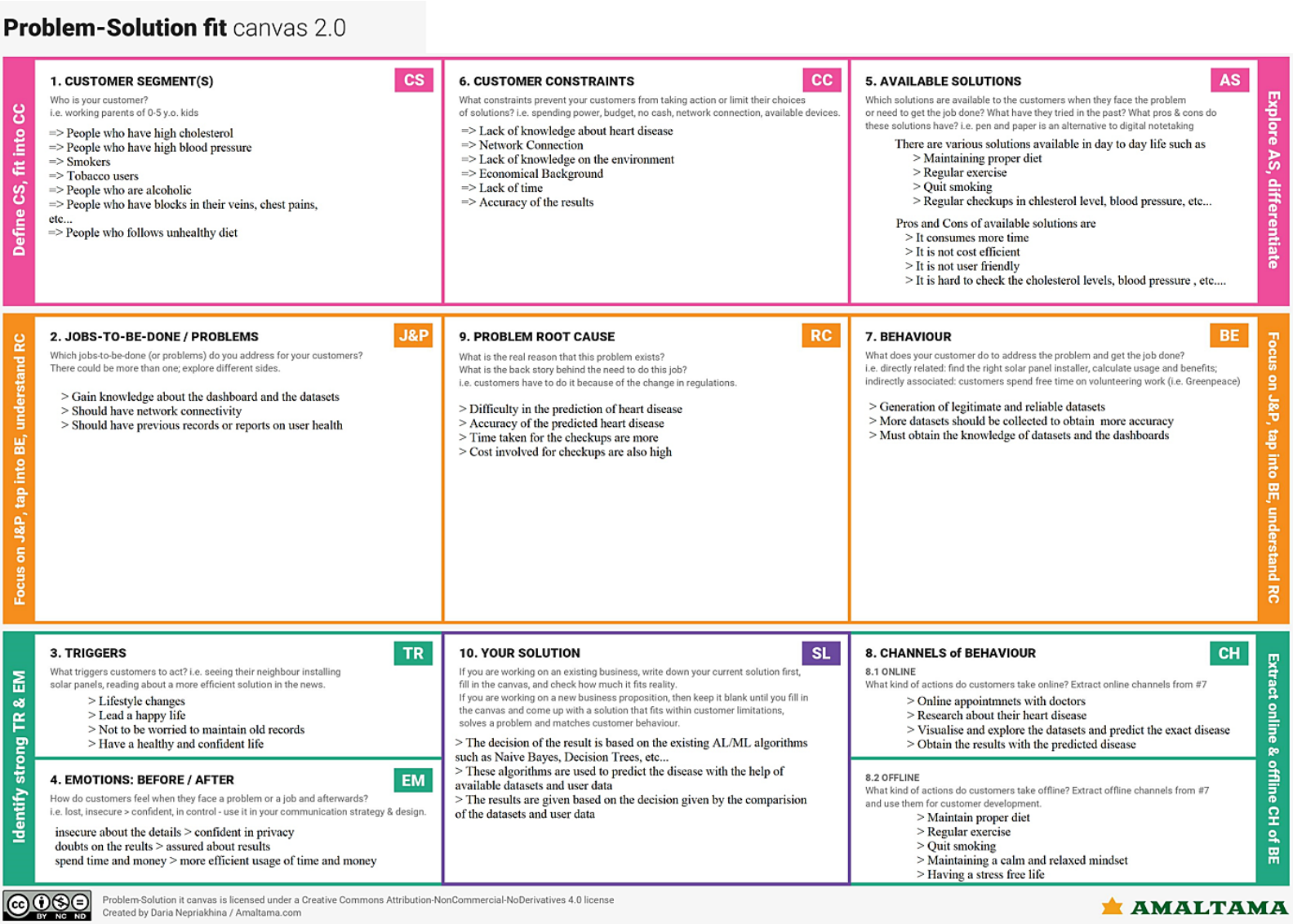


3.3 Proposed Solution:

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Heart disease is one of the most frequently affected diseases. It is infeasible for people to undergo tests for checking heart problems because it is expensive. The motive of this project is to make an interactive dashboard which helps people to visualise and predict diseases based on their symptoms. This dashboard should be reliable, user friendly and cost efficient.
2.	Idea / Solution description	The idea or solution for the problem statement is to create an interactive dashboard which visualises and predicts the heart disease. With the help of this dashboard, user can view his/her medical report and the analysis of the report. He can also predict the result based on the reports. The dashboard is created with the help of Cognos Analytics and the algorithm behind the solution is Naïve Bayes Algorithm.
3.	Novelty / Uniqueness	The novelty behind the proposed solution is to provide preventive measures to prevent heart disease. It also provides the periodical report of the user. It also stores all the report and it can be retrieved for every periodic check.
4.	Social Impact / Customer Satisfaction	This proposed solution will help people to check periodically and it also reduces the cost spent for checking of disease. This is also helpful for doctors to easily predict and give solution to the users based on their problem. It can predict the disease in its earlier stage and makes the user alert about his condition.

5.	Business Model (Revenue Model)	The dashboard is very interactive and is user friendly. Due to these factors, more people will be attracted to this dashboard which will make more profit when compared to hospitals. These dashboards can be usable everywhere easily. Hence the revenue will be more.
6.	Scalability of the Solution	The dashboard will work efficiently and accurately in both small scale and large scale datasets in simpler and faster way.

3.4 Problem Solution Fit:



4. REQUIREMENT ANALYSIS

4.1 Functional Requirement:

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Enable user to register for the application through Gmail (or) mobile number to collect their data
FR-2	User Confirmation	After registration, the confirmation of the user will be done through OTP via phone number (or) Gmail
FR-3	Visualisation and Prediction	User can visualise the heart disease with the help of dashboards created using Cognos Analytics and prediction can be done with the help of datasets and user data
FR-4	Report creation	With the help of visualisation and prediction, user can receive their report and with the help of reports, user can undergo treatments
FR-5	Treatment process	With the help of the report, corresponding treatments and their details such as cost, time will be given. User can choose the treatment based on the report

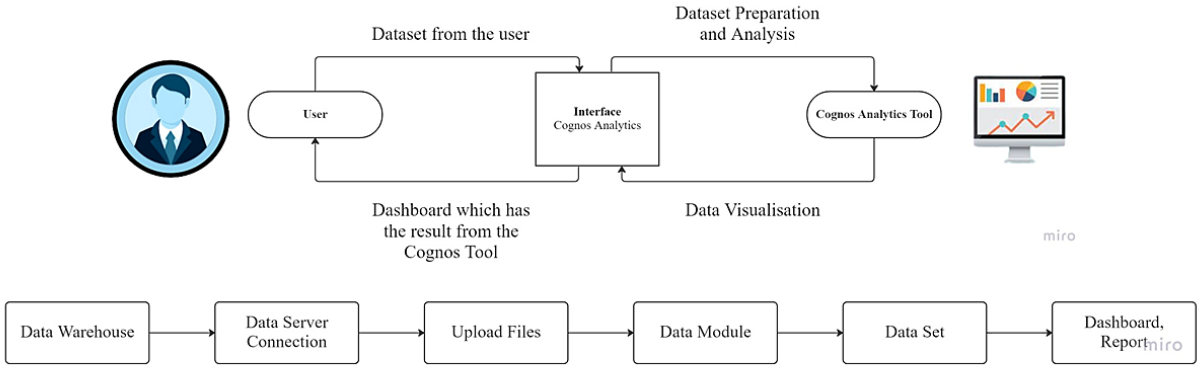
4.2 Non-Functional Requirement:

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The user can easily understand the platform and use all the features provided by the application because it is user friendly and simple.
NFR-2	**Security**	The privacy is the most important feature provided by the application with the help of high level encryption and decryption. It also provides some additional features for backup and recovery of data.
NFR-3	**Reliability**	The application should be more reliable at every situation and the result should be more accurate. The application should work without failures.
NFR-4	**Performance**	Performance of the application is mainly based on the speed and response time taken for the given input data. The speed and the response time of the application should be fast in order to visualise and predict the disease quickly and efficiently.
NFR-5	**Availability**	The tool or application should be available free at any time. It should work 24*7 without any congestion in data.
NFR-6	**Scalability**	The application should work efficiently for more number of users with high speed and performance. It should also be reliable for the users.

5. PROJECT DESIGN

5.1 Data Flow Diagram:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



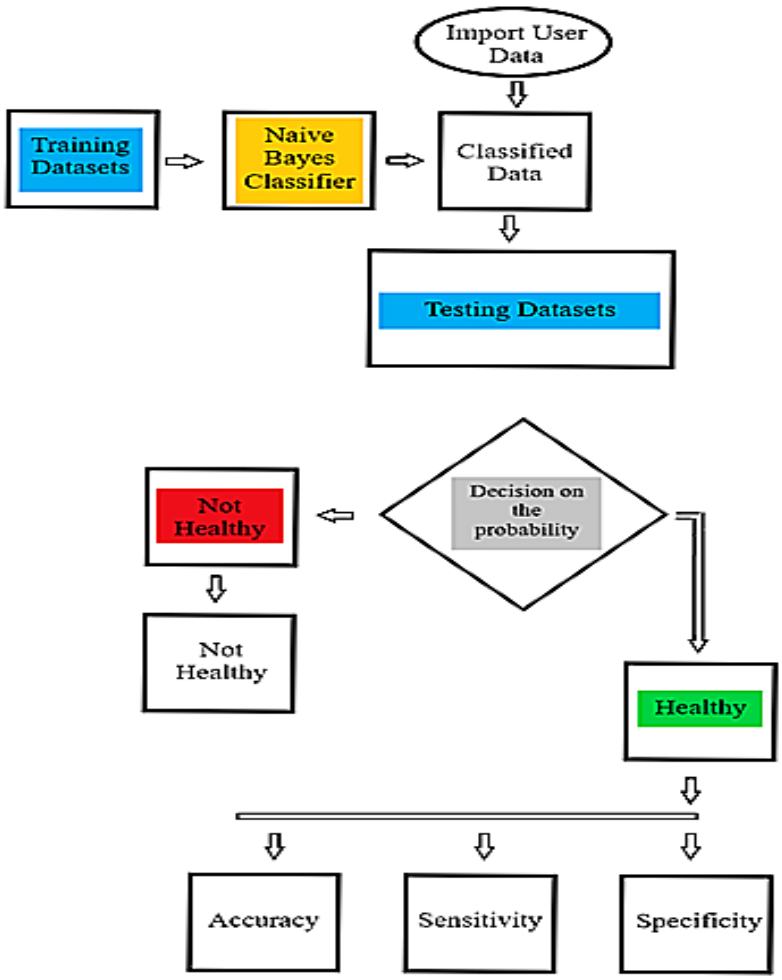
Data Flow Explanation:

1. User creates an own login/account in the application with their username and password which is encrypted and known only to the user
2. After the successful creation, user can upload his/her medical records to the data warehouse which is connected by Data server connection
3. When the medical records are uploaded, data are processed and prepared using Cognos Tool

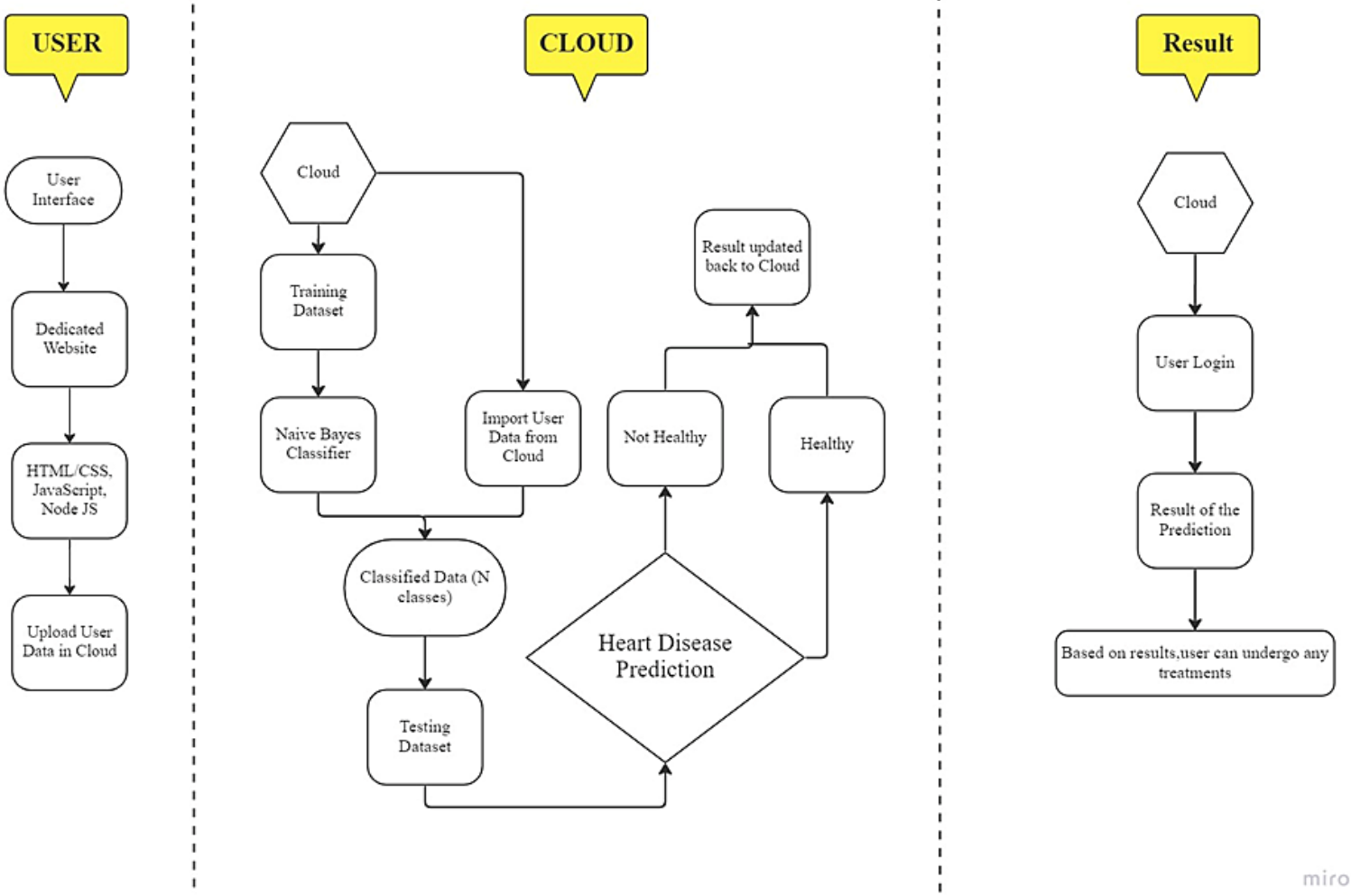
4. User can visualise his/her record with the help of prepared data by the tool
5. By the visualisation, user can view the accuracy of probability of the occurrence of the heart disease in the dashboard
6. After the result, user can take treatments based on the result

5.2 Solution and Technical Architecture:

Solution Architecture:



Technical Architecture:



Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Frameworks are about more than just creating a development environment. They help to define a set of standards that programmers can follow when working collectively. When programmers choose a certain framework, they adopt the specific tools and methodologies associated with that framework. This also means they must be mindful of your choice, as they may end up with processes that don't fit the needs of their project or the developers involved.	Metadata modeling tool
2.	Security Implementations	IAM Controls and Encryptions are implemented to improve security of the application.	SHA-256, Encryptions, IBM Cognos security
3.	Scalable Architecture	Scalable operations are implemented using APIs like HTTP, HTTPS.	Planning Services, API Gateway
4.	Availability	To ensure high availability and optimal service, the load balancer performs continual health checks of each server in the cluster, using probes to determine its eligibility for requests. Also it can user FIFO method to serve user based on their login	Server Load Balancers, FIFO
5.	Performance	Performance of the system is increased using caching methodology. Caching mainly helps in storing the data without any data loss.	Caching

Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	The way of interaction between user and the application	HTML, CSS, JavaScript , NodeJS
2.	Cloud	User data are stored in Cloud, it is storage area for the data	IBM Cloud
3.	Data cleaning and pre-processing	Data cleaning is a process by which inaccurate, poorly formatted, or otherwise messy data is organized and corrected it. Data pre-processing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure	Cognos Analytic Tool
4.	Training Dataset	Training data is the subset of original data that is used to train the machine learning model.	Python
5.	Naive Bayes classifier	Algorithm used for the prediction of the disease	Python
6.	Testing dataset	Test data is data which has been used to check the accuracy of the ML model. Testing is done with classified user data based on Naive Bayes classifier	Python
7.	Result of the application	The result is based on the decision from the testing data and gives output as healthy or unhealthy	Python
8.	Result to the user	The result obtained from the testing data is uploaded again back to the cloud	IBM Cloud
9.	User result access	The user can view his/her result in their cloud login	IBM Cloud

5.3 User Story:

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
	Login	USN-3	As a user, I can log into the application by entering email & password	I can access my account/ Dashboard when logged in	High	Sprint-1
Customer (Web user)	Dashboard	USN-4	User can view his/her complete medical analysis and accuracy of disease prediction	I can view my medical analysis in the dashboard	High	Sprint-2
Customer Care Executive	Helpdesk	USN-5	As a customer care executive, he/she can view (or) answer the customer queries.	I can post my queries and get support form the helpdesk	High	Sprint-3
Administrator	User Profile	USN-6	As an admin, he/she can update the health details of users.	I can view my updated health details.	High	Sprint-4
		USN-7	As an admin, he/she can add or delete users and manage the user data	I can access my account / Dashboard and view the organized data of myself	High	Sprint-4

6. PROJECT PLANING AND SCHEDULING:

6.1 Sprint Planning and Estimation:

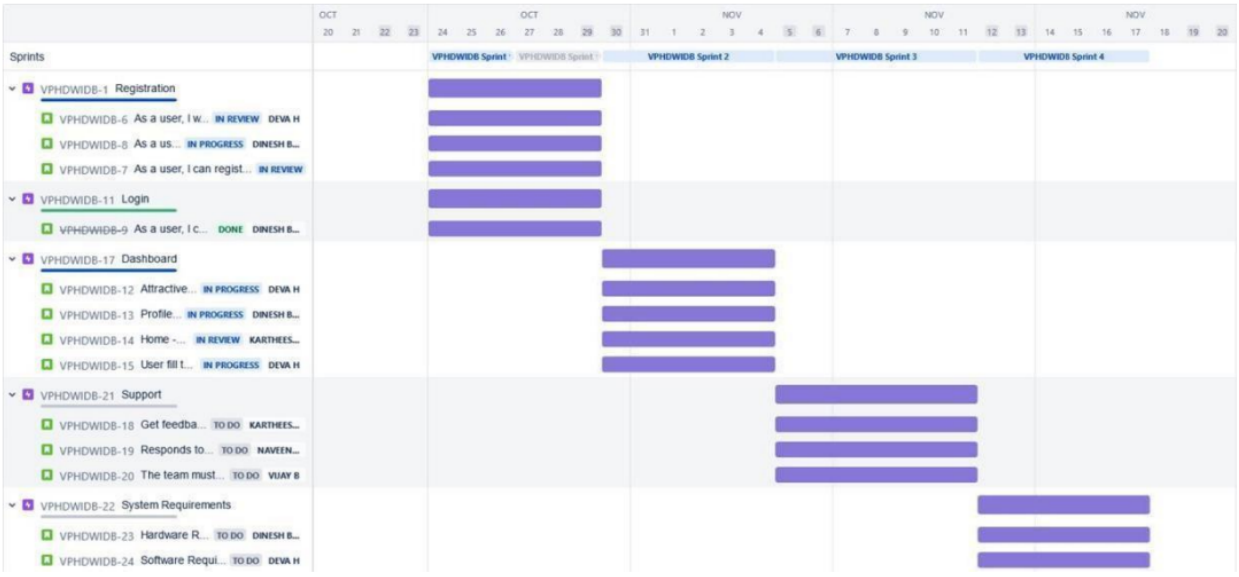
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Story points	Priority	Team Members
Sprint- 1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password	I can access my account / dashboard	10	high	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	5	High	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
		USN-3	As a user, I can register for the application through Gmail	I can register & access the dashboard with Gmail Login	5	High	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
Sprint-2	Login	USN-4	After Registration Login page will appear, The user will login using the login credentials	I can register & access the Dashboard with Gmail login	20	High	Pumenitha S T Shivani T U Karthikeyan P Geffrey N

		USN-7	The user can change password	I can able to change the password.	10	Medium	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
Sprint-4	Classified result	USN-8	Home - Analyze your Heart	I can detect the heart condition from where ever I want.	5	High	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
		USN-9	The user will have to fill in the 13 required fields for the system to predict a heart disease	This will prevent the user to predict whether I has heart disease or not based on the values I entered	10	High	Pumenitha S T Shivani T U Karthikeyan P Geffrey N
		USN-10	The report is generated based on the condition	The user can able to view/download the report if needed	5	Medium	Pumenitha S T Shivani T U Karthikeyan P Geffrey N

6.2 Project Delivery Schedule:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	01 Nov 2022	06 Nov 2022	20	03 Oct 2022
Sprint-2	20	6 Days	08 Nov 2022	13 Nov 2022	20	10 Nov 2022
Sprint-3	20	6 Days	15 Nov 2022	20 Nov 2022	20	17 Nov 2022
Sprint-4	20	6 Days	22 Nov 2022	27 Nov 2022	20	24 Nov 2022

6.3 Reports from JIRA:



7. CODING AND SOLUTIONING

7.1 Machine Learning:

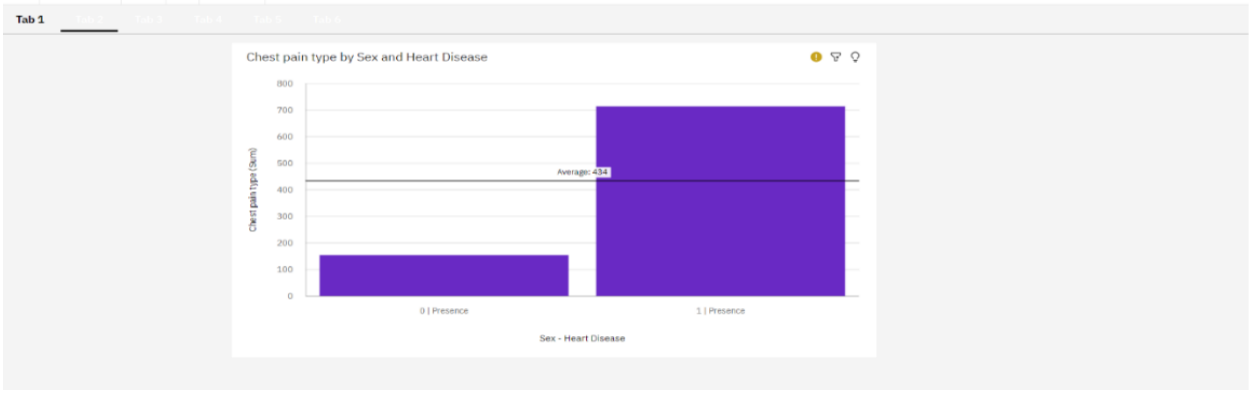
Out[ ]:

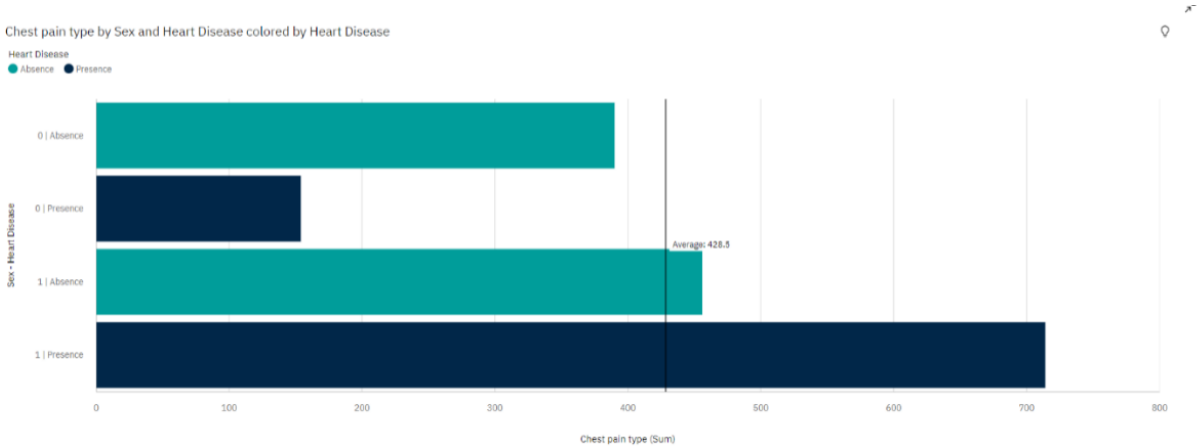
	Estimators	Accuracy
0	Linear Regression	0.565830
3	K-Nearest Neighbor	0.729167
4	Random Forest	0.854167
5	Bagging Decision Tree	0.854167
6	Hard coting classifier	0.854167
2	Gaussian Naive Bayes	0.875000
1	Logistic Regression	0.895833

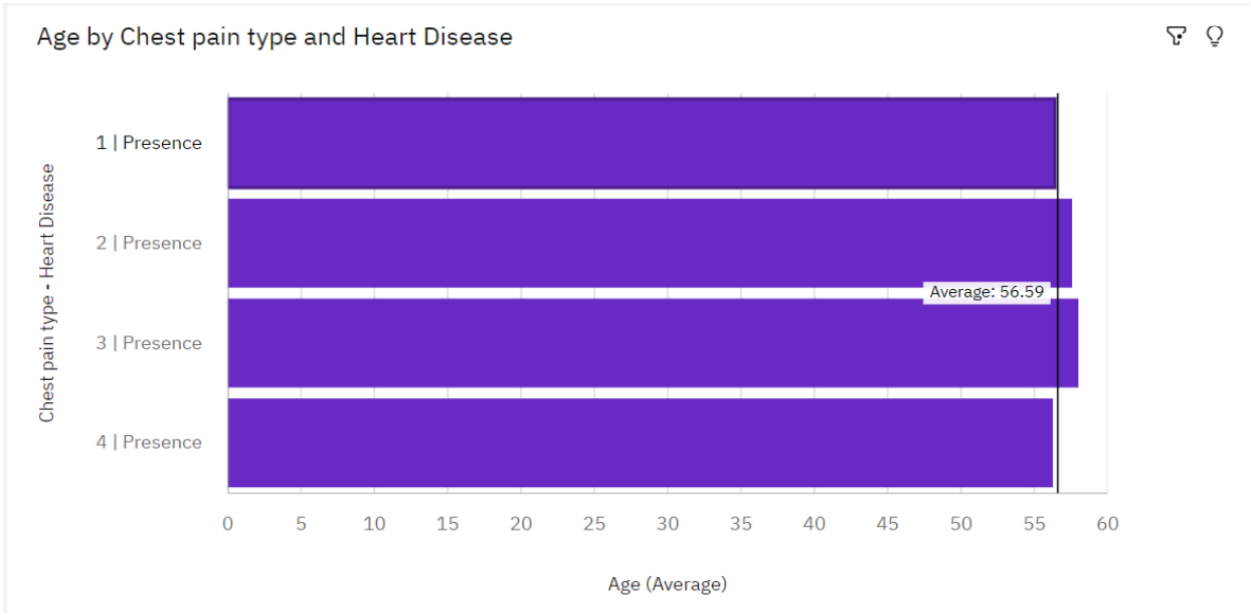
From the above result we can conclude that Logistic Regression has the hisgest accuracy for this particular dataset.

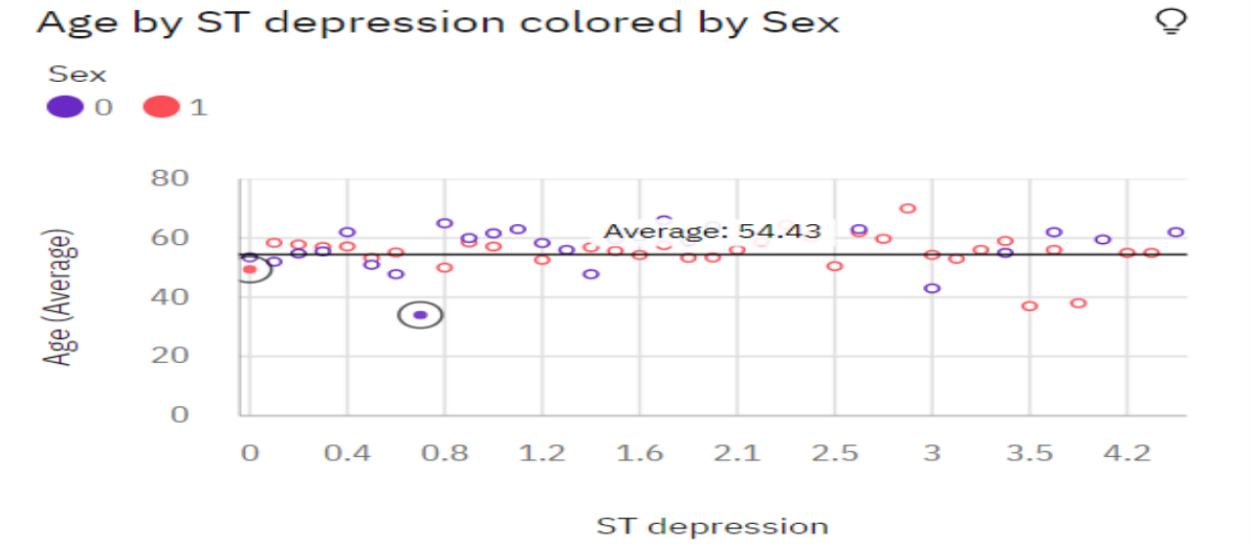
```
Testing Accuracy for Decision Tree: 0.9264705882352942
Testing Sensitivity for Decision Tree: 0.8888888888888888
Testing Specificity for Decision Tree: 1.0
Testing Precision for Decision Tree: 1.0
```

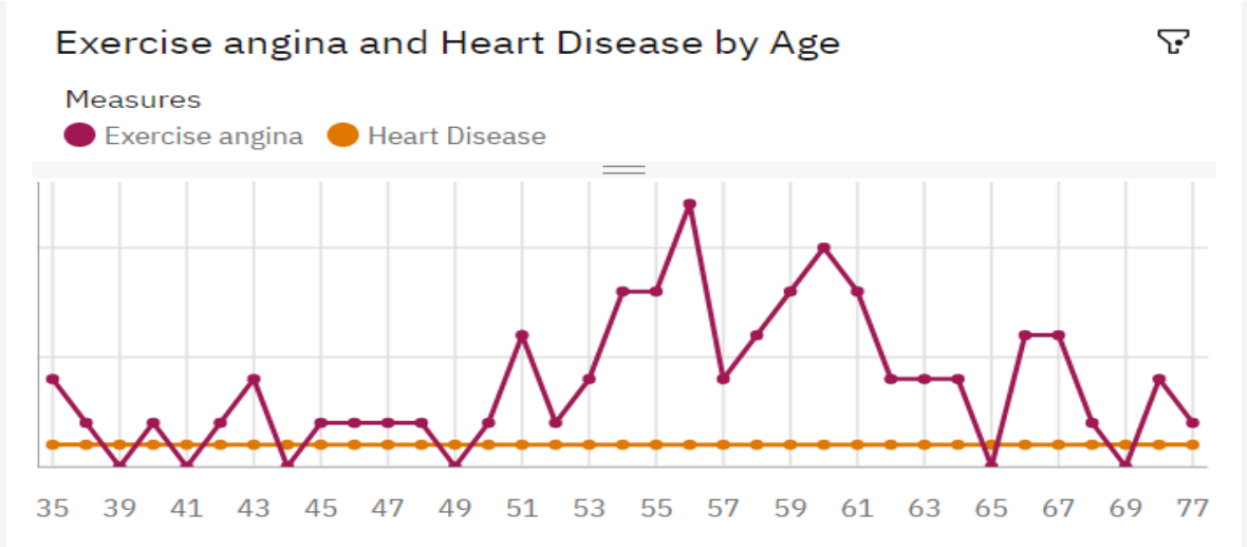
7.2 Dashboard:

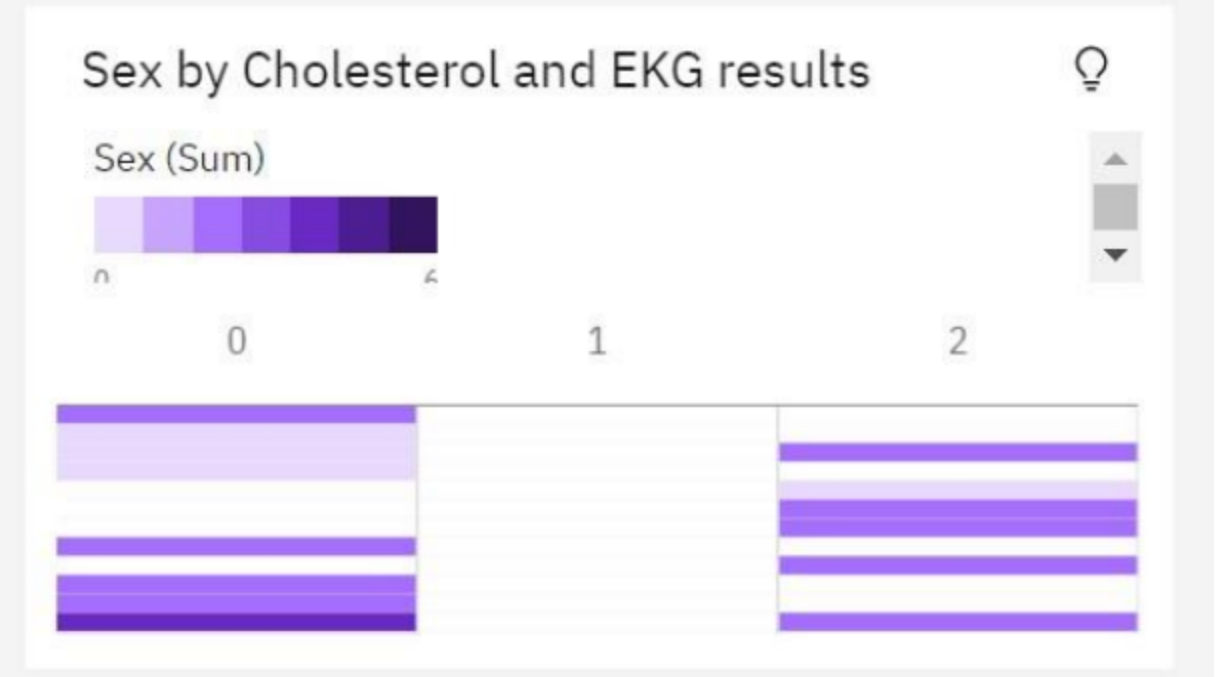


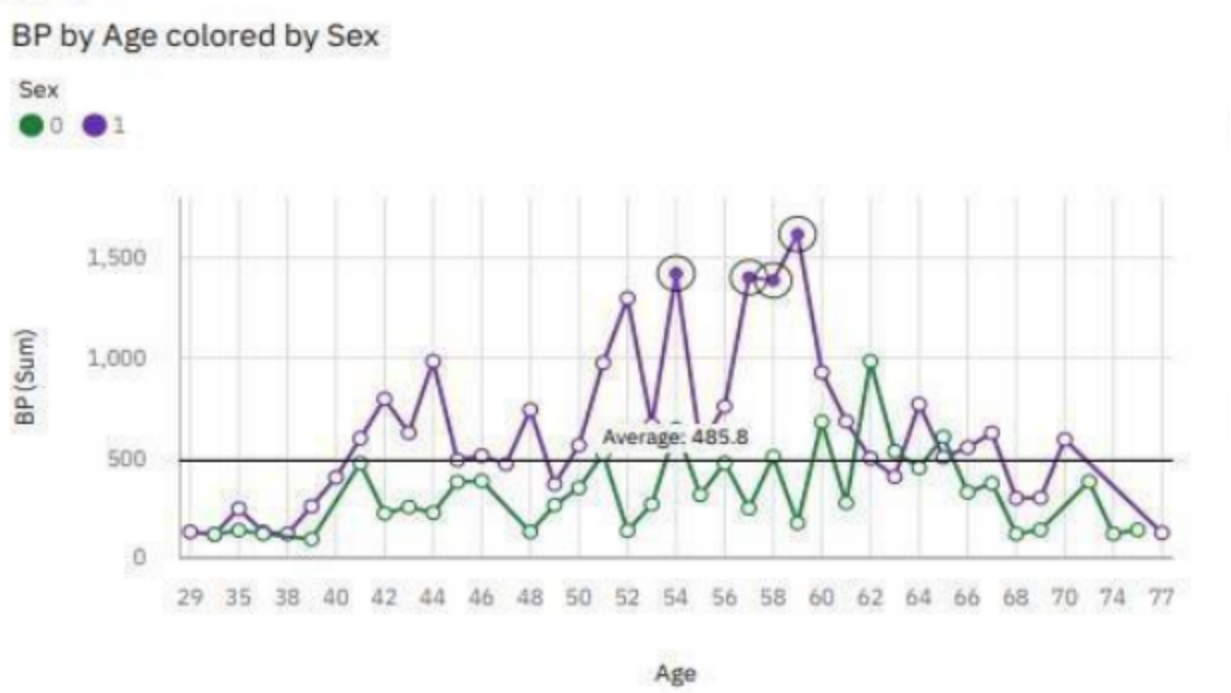



Age by Chest pain type and Heart Disease
Chest pain type - Heart Disease
1 | Presence
2 | Presence
3 | Presence
4 | Presence
Average: 56.59
0
5
10
15
20
25
30
35
40
45
50
55
60
Age (Average)


Age by ST depression colored by Sex
Sex
0
1
Age (Average)
80
60
40
20
0
Average: 54.43
0
0.4
0.8
1.2
1.6
2.1
2.5
3
3.5
4.2
ST depression


Exercise angina and Heart Disease by Age
Measures
Exercise angina
Heart Disease
35 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 77


Sex by Cholesterol and EKG results
Sex (Sum)
0
6
0
1
2



8. TESTING

8.1 Test Cases:

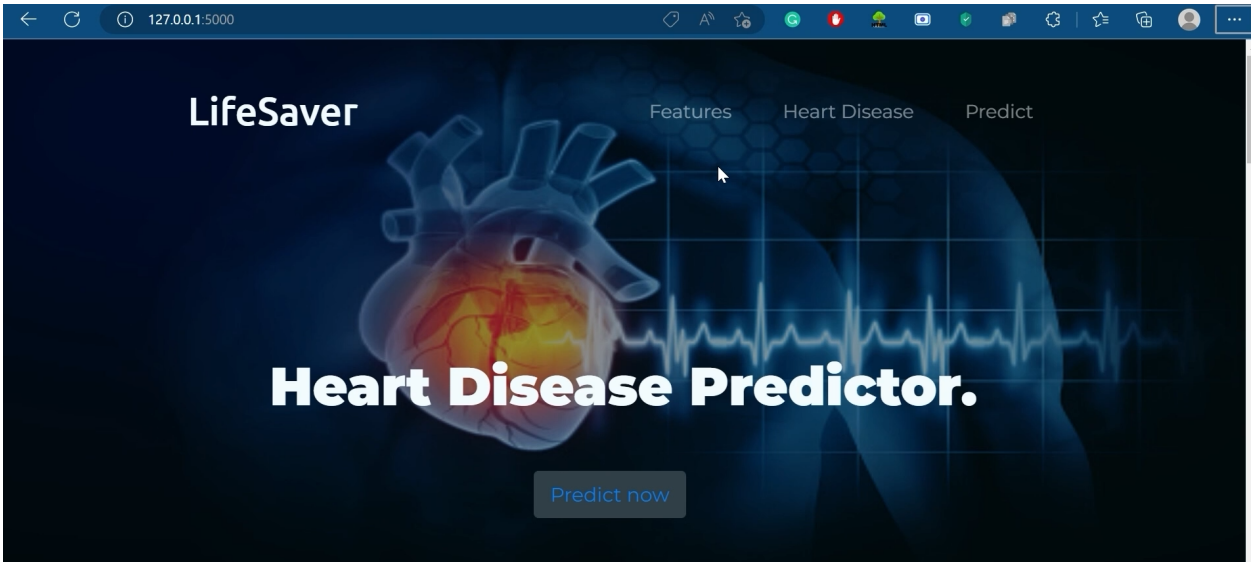
In [ ]:
```
from sklearn.metrics import accuracy_score
input=(63,1,3,145,200,150,98,0,0,0,0,0,0)
input_as_numpy=np.asarray(input)
input_reshaped=input_as_numpy.reshape(1,-1)
pre1=tree_model.predict(input_reshaped)
print(pre1)
a1 = accuracy_score(pre1,model1.predict(input_reshaped)) * 100
print(a1)
```

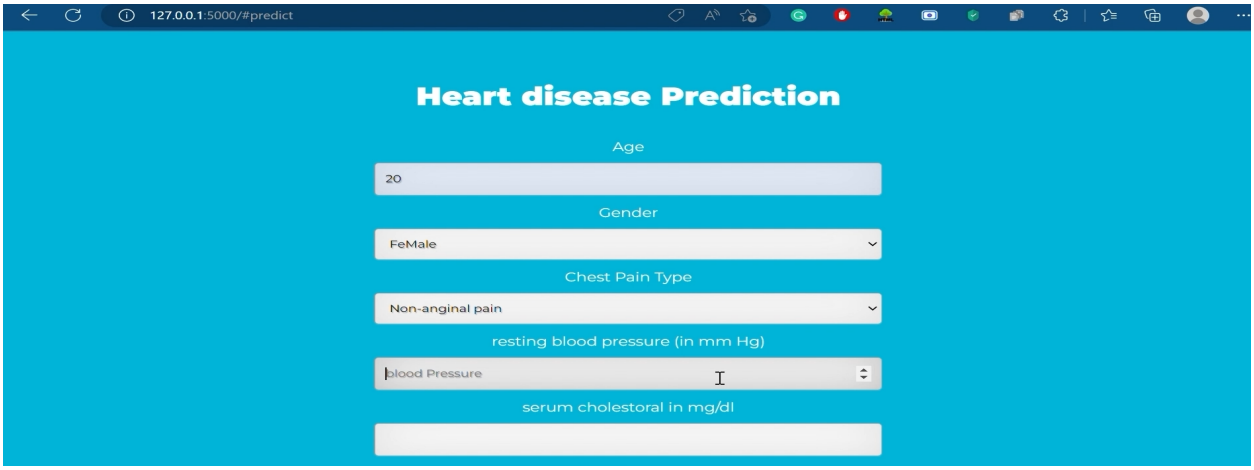
```
['Absence']
100.0
```

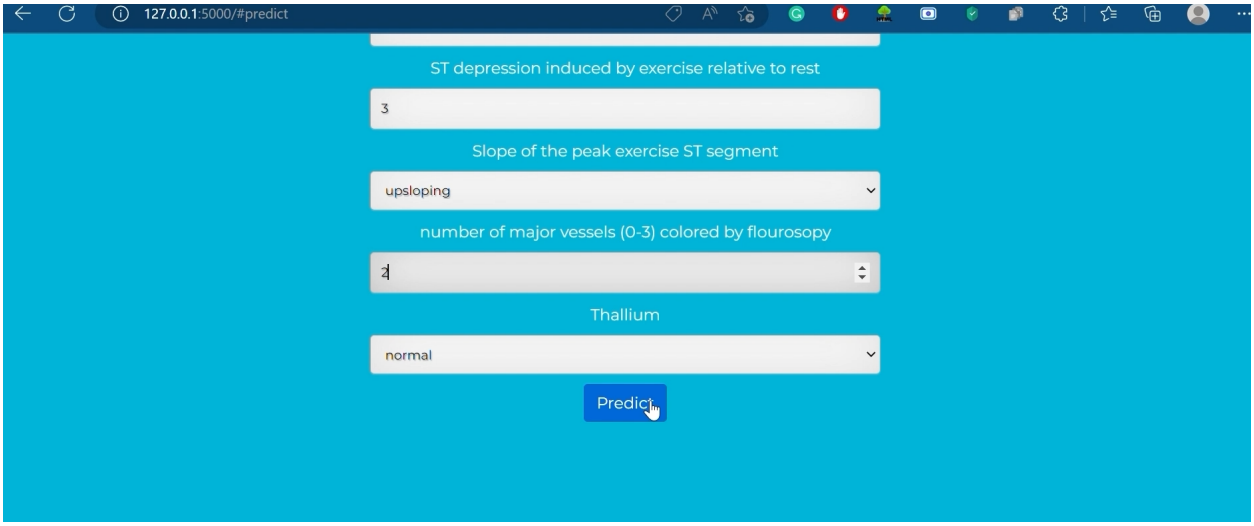
In [ ]:
```
from sklearn.metrics import accuracy_score
input=(70,1,4,130,322,0,2,109,0,2.4,2,3,3)
input_as_numpy=np.asarray(input)
input_reshaped=input_as_numpy.reshape(1,-1)
pre1=tree_model.predict(input_reshaped)
print(pre1)
a1 = accuracy_score(pre1,model1.predict(input_reshaped)) * 100
print(a1)
```

```
['Presence']
100.0
```

8.2 User Acceptance Testing:







9. RESULTS

9.1 Performance Metrices:

In [ ]:

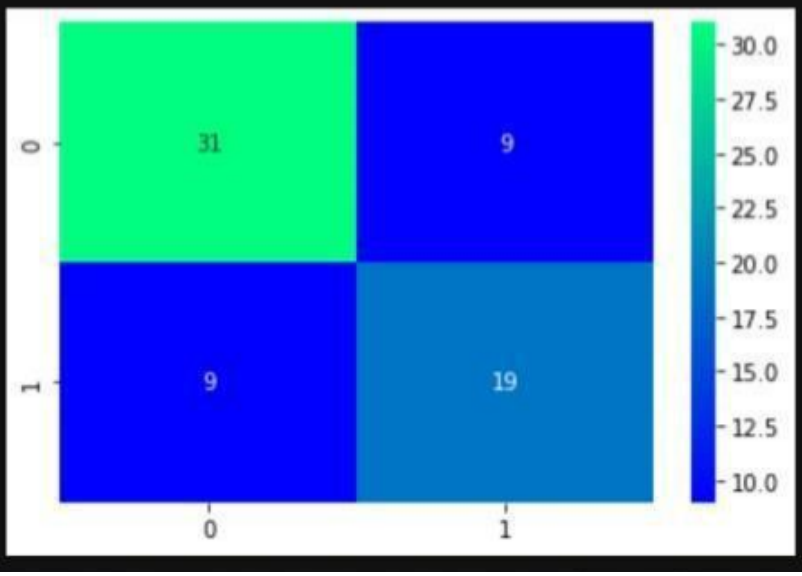
```
from sklearn.tree import DecisionTreeClassifier

dtclas=DecisionTreeClassifier()
modeldt=dtclas.fit(X_train,y_train)
predictiondt=modeldt.predict(X_test)
cmdt= confusion_matrix(y_test,predictiondt)
sns.heatmap(cmdt, annot=True,cmap='winter')
print(classification_report(y_test, predictiondt))
```

```
              precision    recall  f1-score   support

     Absence       0.78      0.78      0.78        40
    Presence       0.68      0.68      0.68        28

    accuracy                           0.74        68
   macro avg       0.73      0.73      0.73        68
weighted avg       0.74      0.74      0.74        68
```



10. ADVANTAGES AND DISADVANTAGES

10.1 Advantages:

- Easy to understand
- Faster way to predict the disease
- Low cost
- User Friendly
- More security
- Dashboards are interactive

10.2 Disadvantages:

- User needs previous knowledge about the fields
- Afraid of false results
- Need to understand the platform
- User needs his/her report

11. CONCLUSION

Heart disease includes heart attack and stroke. The risk of the disease can be reduced by the analysis in the early stages and taking necessary treatments. Hence, the website will help users to predict his/her health and helps user to take care of their health in an healthy manner.

Visualising and predicting Heart Disease using an Interactive Dashboard is an efficient way of identifying his/her health at the early stages with the help of ML Algorithms

12. FUTURE SCOPE

The main objective of the project is just to predict the heart disease but the necessary solutions for the disease can be given so that user can easily maintain his/her health condition according to the solutions given based on the prediction.
This will have a huge impact on the society and it will prevent people suffering from heart disease.

13. APPENDIX

Source Code:
https://github.com/IBM-EPBL/IBM-Project-10542-1659186133

Demo Link:
https://drive.google.com/file/d/1j4J7gR1KC4qsirL6RGv6sQr3XQ1l4xFV/view?usp=share_link

Thank You