Visualizing and Predicting Heart Diseases with an Interactive Dashboard

-

PROFESSIONAL READINESS FOR INNOVATION,EMPLOYABILITY AND ENTREPRENEURSHIP - 2022

Team ID: PNT2022TMID12681

TEAM LEADER: SRIKANTH (19Z349)

TEAM MEMBER: GOWTHAM S (19Z316)

TEAM MEMBER: CHANDRAPRAKASH J (20Z461)

TEAM MEMBER: RITHISH B (20Z464)

FACULTY MENTOR: PRAKASH J

Project Report Format

1. **INTRODUCTION**
   - 1.1 Project Overview
   - 1.2 Purpose
2. **LITERATURE SURVEY**
   - 2.1 Existing problem
   - 2.2 References
   - 2.3 Problem Statement Definition
3. **IDEATION & PROPOSED SOLUTION**
   - 3.1 Empathy Map Canvas
   - 3.2 Ideation & Brainstorming
   - 3.3 Proposed Solution
   - 3.4 Problem Solution fit
4. **REQUIREMENT ANALYSIS**
   - 4.1 Functional requirement
   - 4.2 Non-Functional requirements
5. **PROJECT DESIGN**
   - 5.1 Data Flow Diagrams
   - 5.2 Solution & Technical Architecture
   - 5.3 User Stories
6. **PROJECT PLANNING & SCHEDULING**
   - 6.1 Sprint Planning & Estimation
   - 6.2 Sprint Delivery Schedule
   - 6.3 Reports from JIRA
7. **CODING & SOLUTIONING (Explain the features added in the project along with code)**
   - 7.1 Feature 1
   - 7.2 Feature 2
   - 7.3 Database Schema (if Applicable)
8. **TESTING**
   - 8.1 Test Cases
   - 8.2 User Acceptance Testing
9. **RESULTS**
   - 9.1 Performance Metrics
10. **ADVANTAGES & DISADVANTAGES**
11. **CONCLUSION**
12. **FUTURE SCOPE**
13. **APPENDIX**
    - Source Code
    - GitHub & Project Demo Link

CHAPTER 1

INTRODUCTION

1.1 Project Overview

Heart disease is the main reason people die in the industrialised world. Work must thus be done to reduce the chances of suffering a heart attack or stroke. In order to forecast which patients are most likely to develop a cardiac ailment in the near future using the provided attributes, this project seeks to develop an interactive dashboard utilising the IBM Cognos Tool and dataset.

1.2 Purpose

In today's culture, heart disease is a leading cause of death. Medical diagnosis is a crucial yet challenging process that needs to be completed precisely and effectively. Due to a number of risk factors, such as high blood pressure, cholesterol, and an abnormal pulse rate, cardiovascular disease is difficult to detect. Using analytics, we can determine which patients are most likely to develop heart disease in the near future, and using patient information, we can decide how best to treat them. Heart disease is one of many illnesses that can be fatal, and it has received a lot of attention in medical studies. Heart disease diagnosis is a difficult undertaking that can provide automated predictions about the patient's heart state to improve the effectiveness of subsequent therapy. Heart disease is often diagnosed based on the patient's physical examination, signs, and symptoms. The risk of heart disease is influenced by a number of variables, including smoking, body cholesterol, family history of the disease, obesity, high blood pressure, and inactivity. The provision of high-quality services at reasonable prices is a significant problem for healthcare institutions, including hospitals and medical facilities. The provision of high-quality care necessitates accurate patient diagnosis and efficient treatment delivery. Both numerical and categorical data are present in the heart disease database that is accessible. These entries are cleaned and filtered to remove any extraneous data from the database before being subjected to further

processing. From a historical heart illness database, the suggested method may extract precise hidden information, i.e., patterns and associations related to heart disease. It may also respond to difficult questions about heart disease diagnosis, which can assist medical professionals make wise clinical judgments. The suggested system has its own special efficacy, according to the results.

CHAPTER 2

LITERATURE SURVEY

2.1 Existing Problem

The "big data" used in the healthcare sector is large and contains patterns or information that may be used to make decisions. Decisions made using the vast amount of data are more accurate than those made using intuition. Exploratory Data Analysis (EDA) looks for the right information, verifies presumptions, and establishes the relationships between the explanatory variables. EDA is viewed in this context as data analysis without statistical modelling or inferences. Analytics is a crucial skill for any career because it may predict the future and reveal hidden patterns. Data analytics is now seen as a low-cost tool that is crucial in the healthcare industry, helping to identify disease outbreaks, emergency situations, and new research findings.

2.2 References

1. Ashir Javeed, Shijie Zhou et al. (2017) designed “An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”

2. Bo Jin, Chao Che et al. (2018) proposed a “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling”.

3. Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”

4. L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, “An optimized stacked support vector machines based expert system for the effective prediction of heart failure,”

5. A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,”

6.M.Satish, D Sridhar, “Prediction of Heart Disease in Data Mining Technique”.

2.3 Problem Definition

The leading cause of death in the developed world is heart disease. Therefore, there needs to be work done to help prevent the risks of having a heart attack or stroke.

Content: Use this dataset to predict which patients are most likely to suffer from a heart disease in the near future using the features given.

CHAPTER 3

IDEATION AND PROPOSED SOLUTION

3.1 Empathy Map

The empathy map's main objective is to foster understanding between users and developers. The empathy map for the vehicle performance analyser based on machine learning is shown in Figure 3.1.

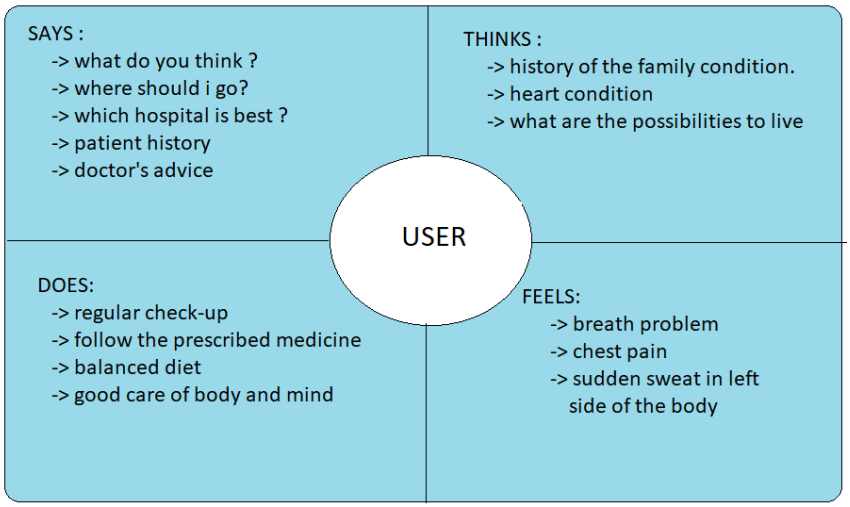
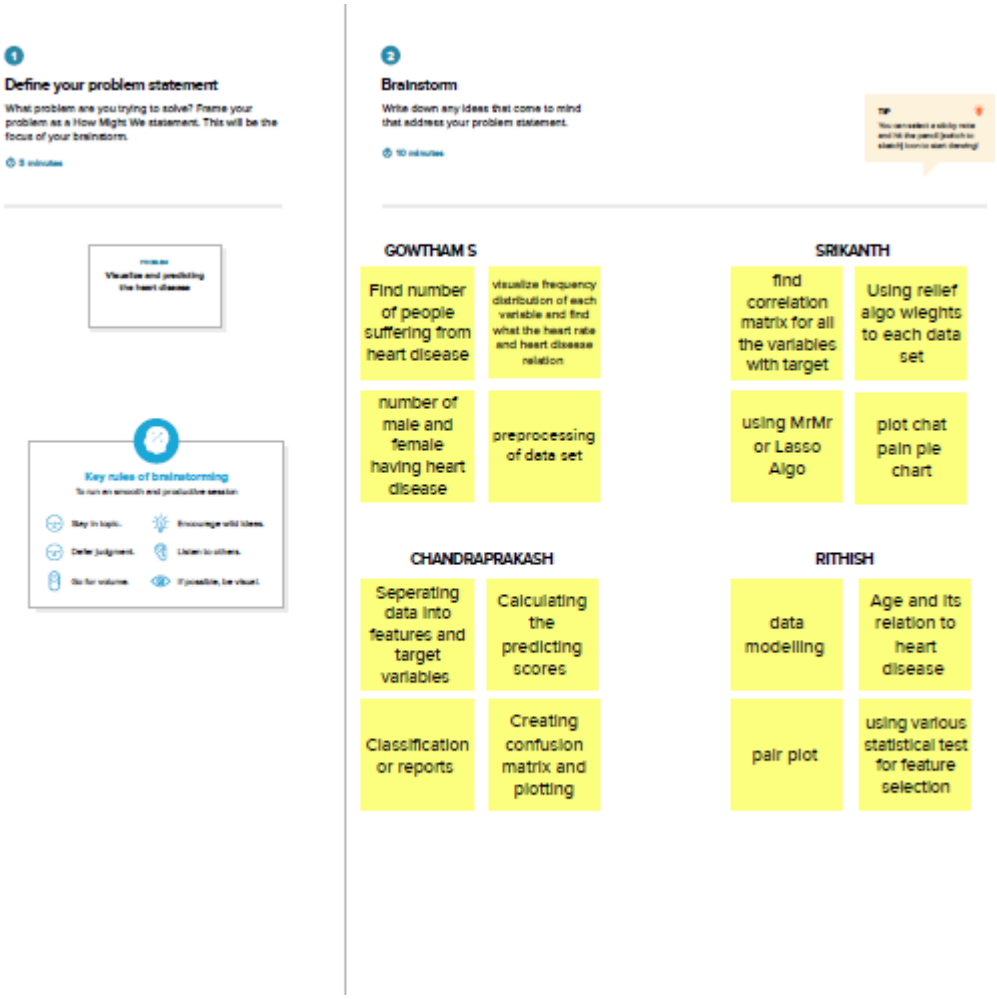


Figure 3.1 – Empathy Map

3.2 Ideation and Brainstorming

The goal of ideation and brainstorming is to generate a lot of ideas, which the team can then sort through and narrow down to the best, most useful, or most creative ones to inspire new and improved design solutions and products. As a result, this is frequently the most exciting phase of a project.



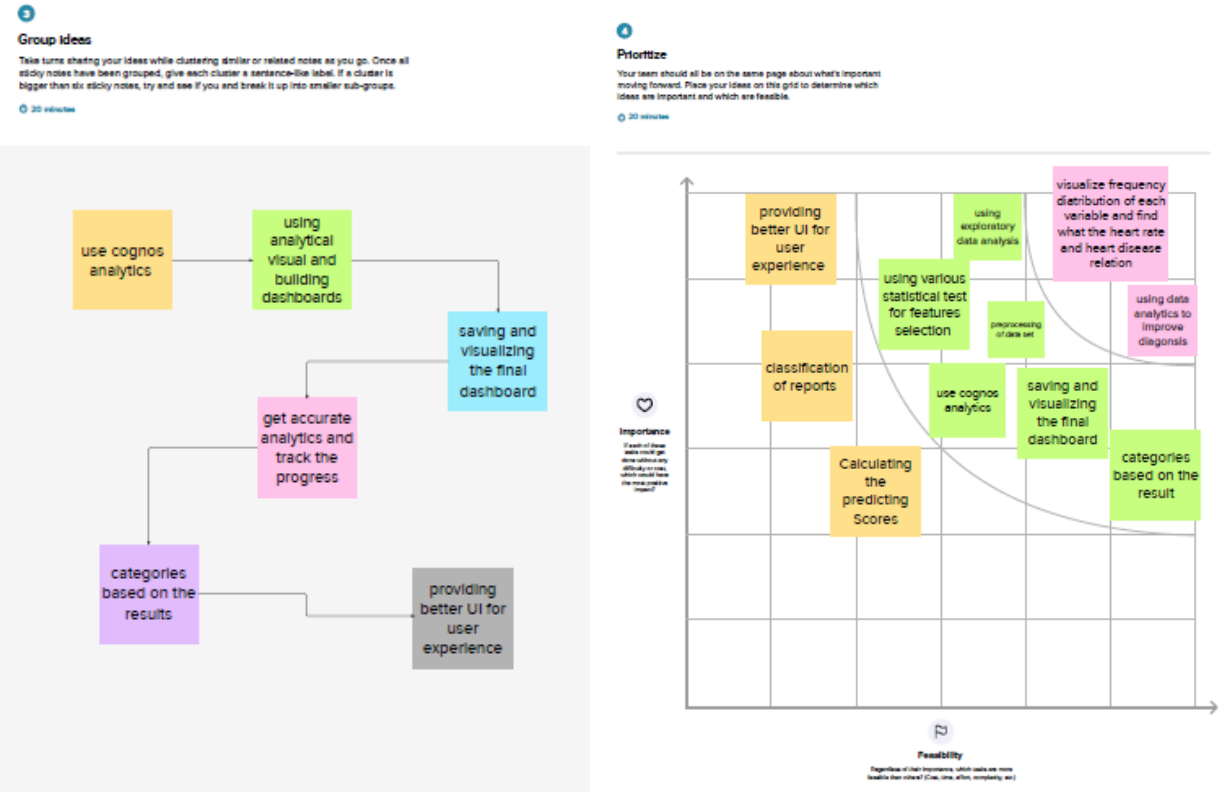


Figure 3.2 – Ideation & Brainstorming

3.3 Proposed Solution

The leading cause of death in the developed world is heart disease. Therefore, there needs to be work done to help prevent the risks of having a heart attack or stroke.

This database contains of 14 fields. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

S.No Field Name

1. 1 Age
2. 2 Sex
3. 3 Chest pain type
4. 4 BP
5. 5 Cholesterol
6. 6 FBS over 120
7. 7 EKG results
8. 8 Max HR

9. 9 Exercise angina
10. 10 ST depression
11. 11 Slope of ST
12. 12 Number of vessels fluro
13. 13 Thallium
14. 14 Heart Disease

Use this dataset to predict which patients are most likely to suffer from a heart disease in the near future using the features given.

Novelty:

Heart diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of the patient and we are going to create the interactive dashboard through which we can analyse the heart diseases based on age, sex, blood pressure of a person,etc.Dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross-validation.

Feasibility of the idea:

Know fundamental concepts and can work on IBM Cognos Analytics, Gain a broad understanding of plotting different visualizations to provide a suitable solution. Able to create meaningful Visualizations and Dashboard(s). We consider a dataset which is having 14 fields by using that we are going to do explorations and building visualizations so that we can analyse the heart diseases of the patient.

Business Models :

Scalability Of The Solution:

We are going to do explorations and visualizations ,Exploration of bp versus chest pain type and gender ,Exploration of max heart rate during the chest pain, Exploration of BP by age, Exploration of cholestrol by age and gender these are the explorations we are going to use. Average age for different chest pain types, Average exercise angina during chest pain,BP variation with respect to age,Effort of existing heart disease on average of exercise angina , Average age for different types of chest pain in existing heart diseases , Maximum heart rate in existing heart disease by exercise angina these all are the visualizations.

3.4 Problem Solution Fit

The answer one has discovered to solve the client's issue is the problem solution fit. The fit of the solution is shown in Figure 3.4.

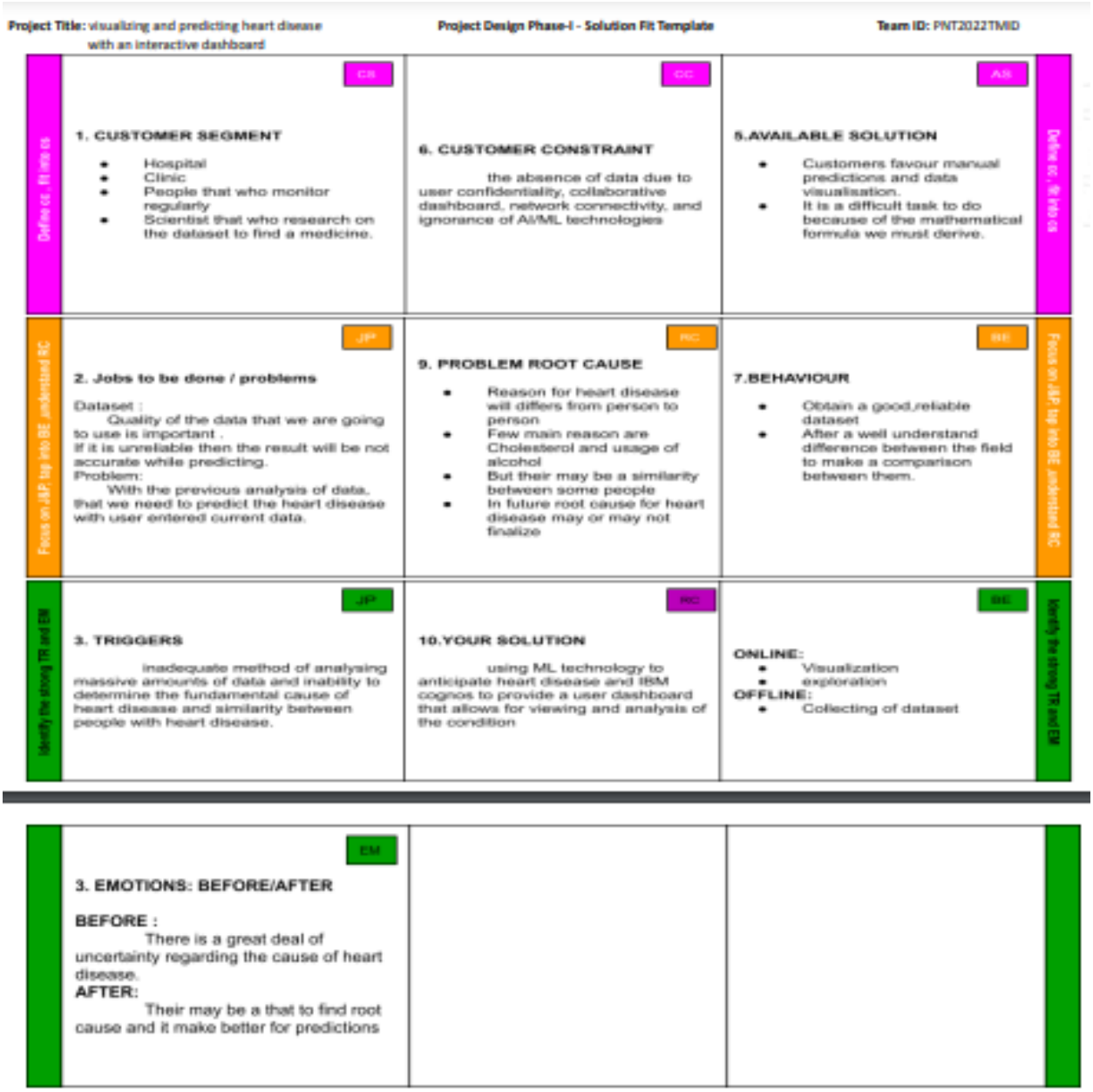


Figure 3.4 – Problem Solution Fit

CHAPTER 4

REQUIREMENT ANALYSIS

4.1Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration for application through Gmail
FR-2	User Confirmation	User gets Confirmation via Email
FR-3	Visualizing data	User can view the Visualization of the available data
FR-4	Predicted Result	User can view the Predicted result based on the medical details

4.2 Non-functional Requirements:

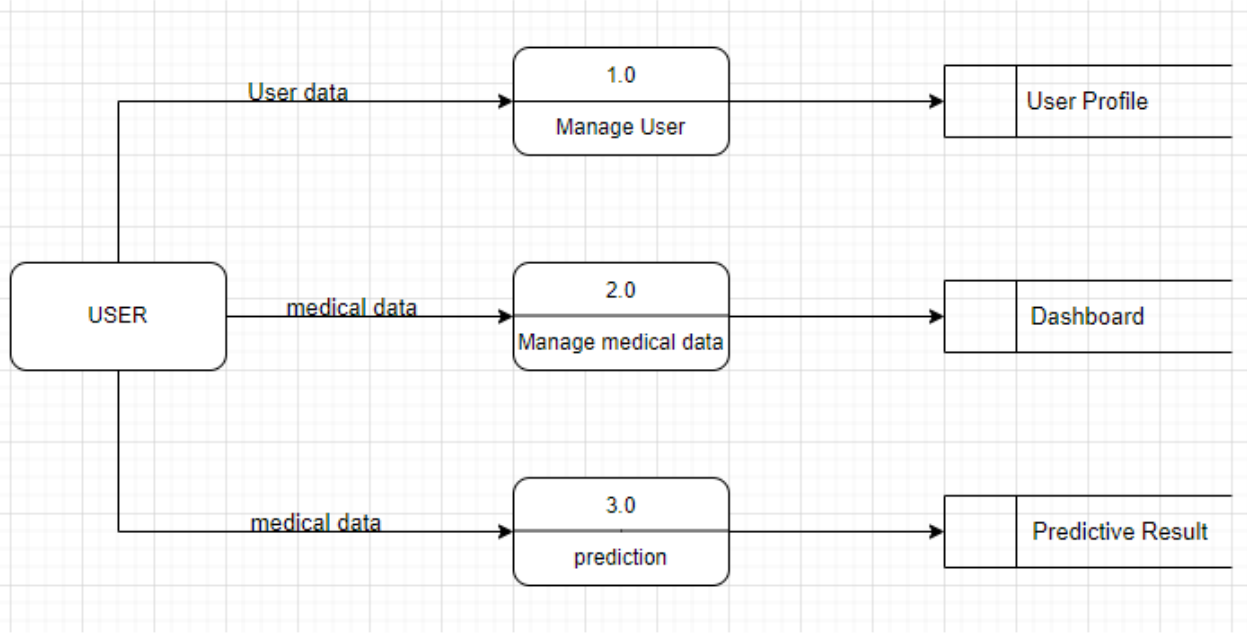
Following are the non-functional requirements of the proposed solution.

NFR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The program will have a straightforward and user-friendly graphical user interface. The application's functionality will be simple for users to comprehend and utilize.
NFR-2	**Security**	Data replication
NFR-3	**Reliability**	The application must be reliable in any environment and consistent in every scenario.
NFR-4	**Performance**	Response time and data submission speed affect how well an application performs.
NFR-5	**Availability**	The application has availability of 24x7 for users
NFR-6	**Scalability**	The program can handle an increase in the number of users.

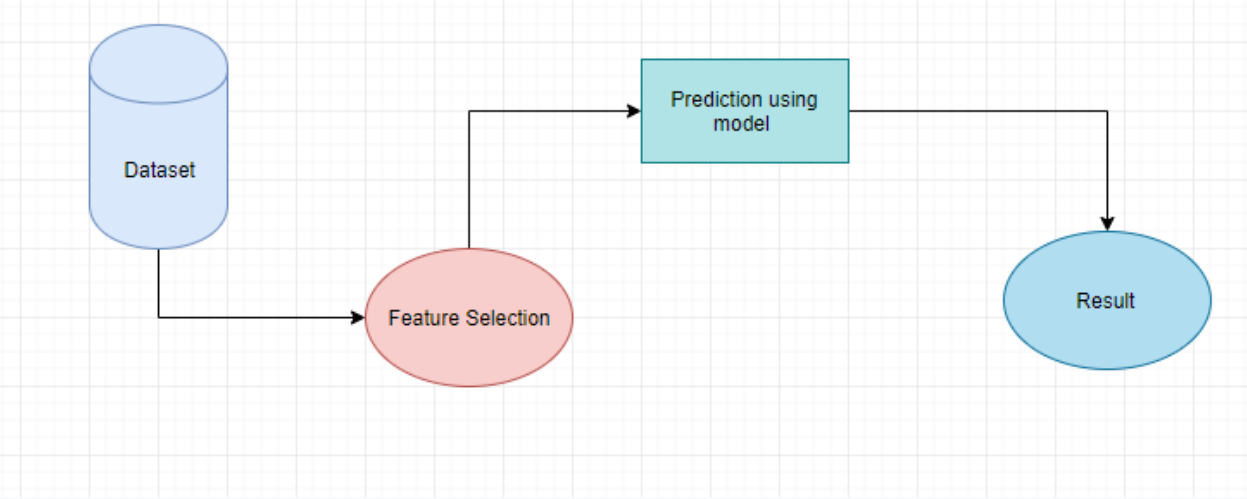
CHAPTER 5

PROJECT DESIGN

5.1 Data flow diagram :



5.2 Solution and Technical architecture:



CHAPTER 6

PROJECT PLANNING AND SCHEDULING

6.1 Sprint planning and Estimation :

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	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.	5	High	Gowtham S, Srikanth , Rithish, Chandrapr akash
Sprint-1		USN-2	As a user, I will receive a confirmation email once I have registered for the application	5	High	Gowtham S, Srikanth , Rithish, Chandrapr akash

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Login	USN-3	As a user, I can log into the application with the user-id and password	10	High	Gowtham S, Srikanth , Rithish, Chandrapr akash
Sprint-2	Dashboard	USN-4	As a user, I can give medical data in the application	20	High	Gowtham S, Srikanth , Rithish, Chandrapr akash
Sprint-3	Helpdesk	USN-5	As a staff, I can view the customer queries	10	Low	Gowtham S, Srikanth , Rithish, Chandrapr akash
Sprint-3		USN-6	As a staff, I can answer the customer queries	10	Low	Gowtham S, Srikanth , Rithish, Chandrapr akash
Sprint-4	Profile	USN-7	As an admin, I can add or delete Users	5	High	Gowtham S, Srikanth , Rithish, Chandrapr akash

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-4		USN-8	As an admin, I can manage user details	15	High	Gowtham S, Srikanth , Rithish, Chandrapr akash

6.2 Sprint Delivery Schedule:

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duratio n	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	19 Nov 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	19 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	19 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

CHAPTER 7

CODING AND SOLUTION

7.1 Feature 1:

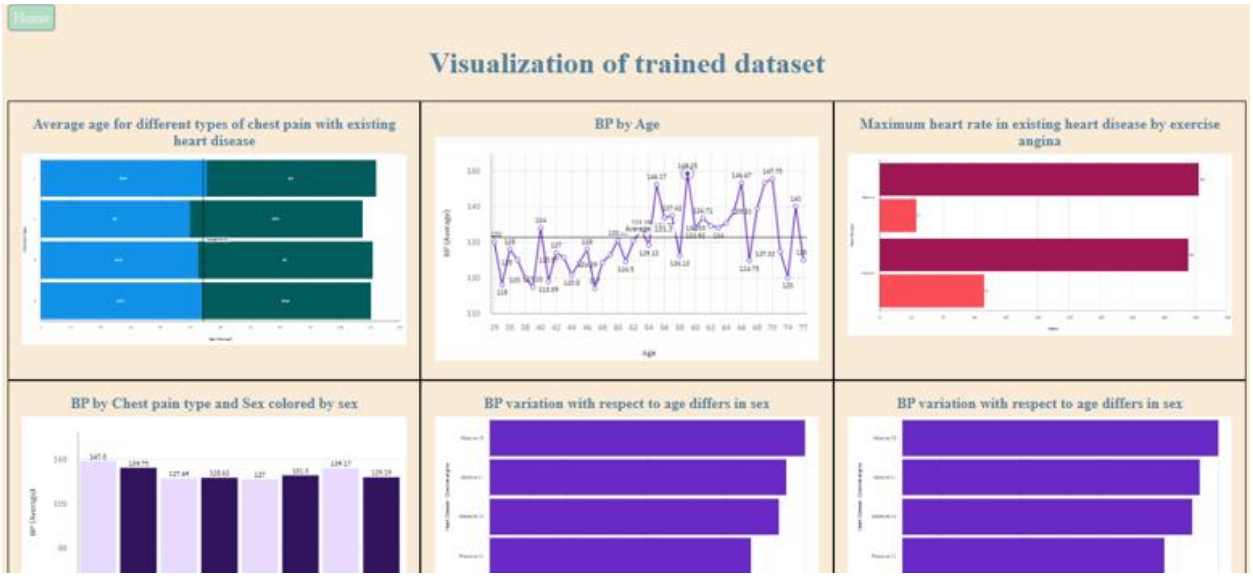
Finding the best model that is suitable for the problem ,

Out[ ]:

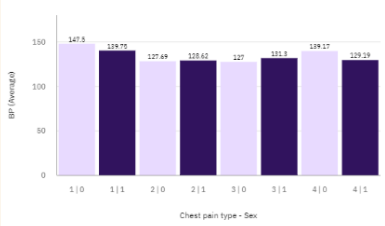
	Estimators	Accuracy
0	Linear Regression	0.565830
3	K-Nearest Neighbor	0.729167
4	Random Forest	0.854167
5	Bagging Decision Tree	0.854167
6	Hard coting classifier	0.854167
2	Gaussian Naive Bayes	0.875000
1	Logistic Regression	0.895833

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

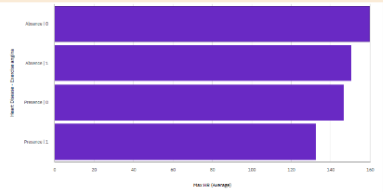
7.2 Feature 2:



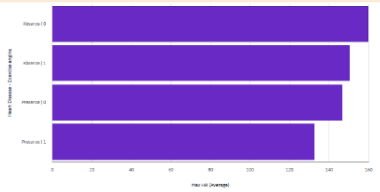
BP by Chest pain type and Sex colored by sex



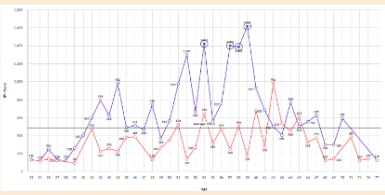
BP variation with respect to age differs in sex



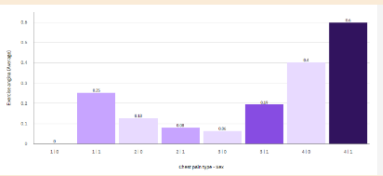
BP variation with respect to age differs in sex



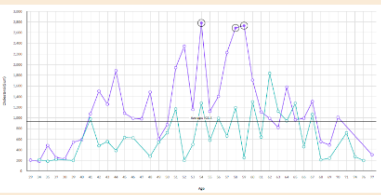
BP variation with respect to age differs in sex



Average Exercise angina during a chest pain



BP variation with respect to age differs in sex



CHAPTER 8

TESTING

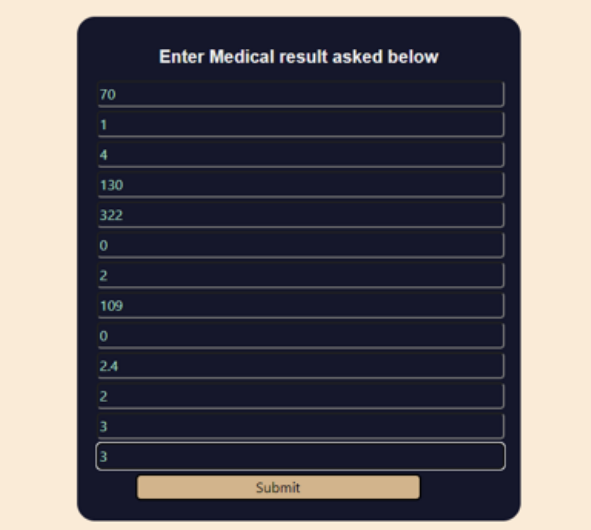
8.1 Test case:

Testing the data with various input value.

```
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
```

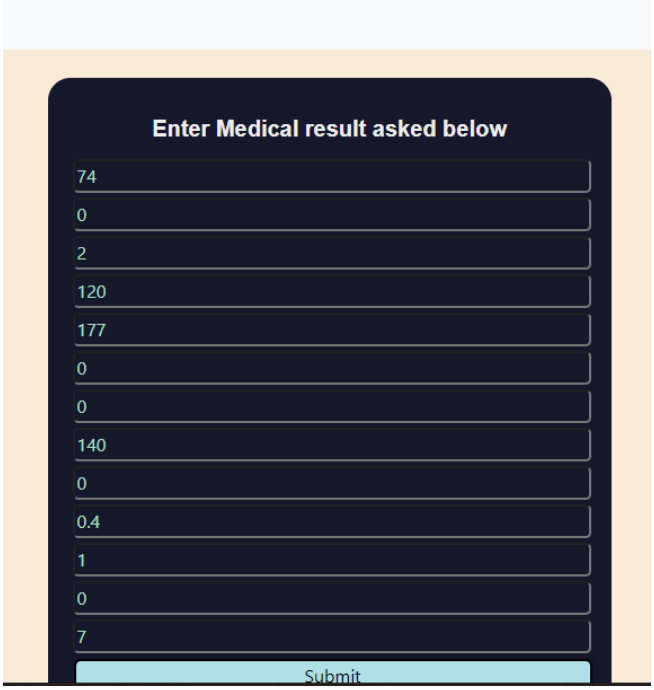
8.2 User Acceptance test case:

For positive response =>



The Predicted result for the enter medical details : Presence

For negative response =>



The Predicted result for the enter medical details : Absence

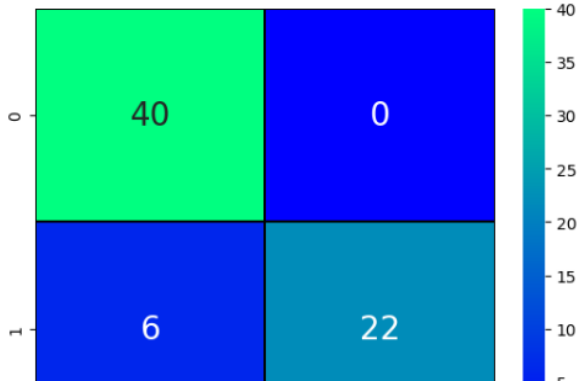
CHAPTER 9

RESULT

9.1 Performance Metrices:

```
from sklearn.model_selection import RandomizedSearchCV
from sklearn.tree import DecisionTreeClassifier

tree_model = DecisionTreeClassifier(max_depth=5,criterion='entropy')
cv_scores = cross_val_score(tree_model, x, y, cv=10, scoring='accuracy')
m=tree_model.fit(x, y)
prediction=m.predict(X_test)
cm= confusion_matrix(y_test,prediction)
sns.heatmap(cm, annot=True,cmap='winter',linewidths=0.3, linecolor='black',annot_kws={"size": 20})
print(classification_report(y_test, prediction))

TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing Accuracy for Decision Tree:',(TP+TN)/(TP+TN+FN+FP))
print('Testing Sensitivity for Decision Tree:',(TP/(TP+FN)))
print('Testing Specificity for Decision Tree:',(TN/(TN+FP)))
print('Testing Precision for Decision Tree:',(TP/(TP+FP)))
```

```
              precision    recall  f1-score   support

     Absence       0.87      1.00      0.93        40
    Presence       1.00      0.79      0.88        28

    accuracy                           0.91        68
   macro avg       0.93      0.89      0.91        68
weighted avg       0.92      0.91      0.91        68

Testing Accuracy for Decision Tree: 0.9117647058823529
Testing Sensitivity for Decision Tree: 0.8695652173913043
Testing Specificity for Decision Tree: 1.0
Testing Precision for Decision Tree: 1.0
```



CHAPTER 10

ADVANTAGES AND DISADVANTAGES

Advantages:

- This is one of the fastest ways to determine if a person is likely to suffer from heart disease or not.
- Useful for medical practitioners to easily classify their patients.
- User Friendly
- Easy to understand
- Secure
- Dashboard provides insightful information

Disadvantages:

- Needs work
- Users need to know all the fields
- Does Not take a null value as input
- Does not provide suggestions to the user

11. Conclusion:

Heart attack and stroke are two heart disease side effects. With early identification and treatment, the risk of problems might be decreased. So, the advice we receive from the website could potentially save lives. Heart disease should always be treated when it is young.

12. Future Scope:

"Prevention is better than cure," the proverb goes. Instead of only anticipating heart disease in its early stages, we need to look into ways to avoid it completely. We must pass a tonne of checks before using this website. Therefore, it would be preferable if we could get the same results with less requirements.

13. Appendix:

Source code: https://github.com/IBM-EPBL/IBM-Project-30383-1660145585/tree/main/final%20deliverable

Demo video: https://drive.google.com/file/d/10FetHcsY7SdIAZaQ3FLGvNTw3gJU4rYm/view?usp=share_link