

Project Report

Visualizing and Predicting Heart Diseases with an Interactive Dash Board

Team ID: PNT2022TMID35649

Team members:

- Adhetya Narayan J.M(Team leader)
- Saikrishna R
- Paavendhan K.S
- Kavın B

Industry Mentor(s) Name: Mahidhar, Saumya

Faculty Mentor(s) Name: Varalakshmi P

1. INTRODUCTION

1.1 Project Overview

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmia); and heart defects you're born with (congenital heart defects), among others.

The main aim of this project is to use IBM Cognos analytics tools to create an interactive dashboard to understand whether a patient has chance of getting heart disease based on certain parameters.

1.2 Purpose

The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease refers to conditions characterized by narrowed or blocked blood vessels, which can result in a heart attack, chest pain (angina), or stroke. Other heart conditions, such as those affecting your heart's muscle, valves, or rhythm, are also classified as heart disease. any types of heart disease can be avoided or treated by adopting a healthy lifestyle. Data analytics here can help users understand the risk of getting heart disease. In this project, a dataset is imported which has the most common attributes used for predicting heart disease.

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

S No.	Paper Title	Key Points
1	Design And Implementing Heart Disease Prediction Using Naive Bayesian	An application based Smart heart prediction system is proposed using Naïve Bayes and this model has provided accuracy of 89.77%
2	Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare.	Feature selection algorithm is designed with SVM to identify heart disease. Optimization methods are used to further increase the performance of a predictive system for HD diagnosis.
3	Prediction of Heart Disease by Mining Frequent Items and Classification Techniques.	Data mining classification methods are used for prediction and Naive Bayes has given highest accuracy.
4	An Intelligent Clinical Decision Support System Based on Artificial Neural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas.	Proposed Correlation-based feature selection (CFS) and Multilayer Perceptron classifier for prediction.
5	Survey on Prediction of Heart Disease Using Data Mining Techniques.	Measured the accuracy using different accuracy parameter of data mining algorithms and proposed that Support Vector Machine technique is an efficient method for predicting heart disease.
6	Intelligent Cardiovascular Disease Risk Estimation Prediction System.	Used K-Nearest Neighbour algorithm and achieved an accuracy of 92.30% and uses less number of attributes for the prediction

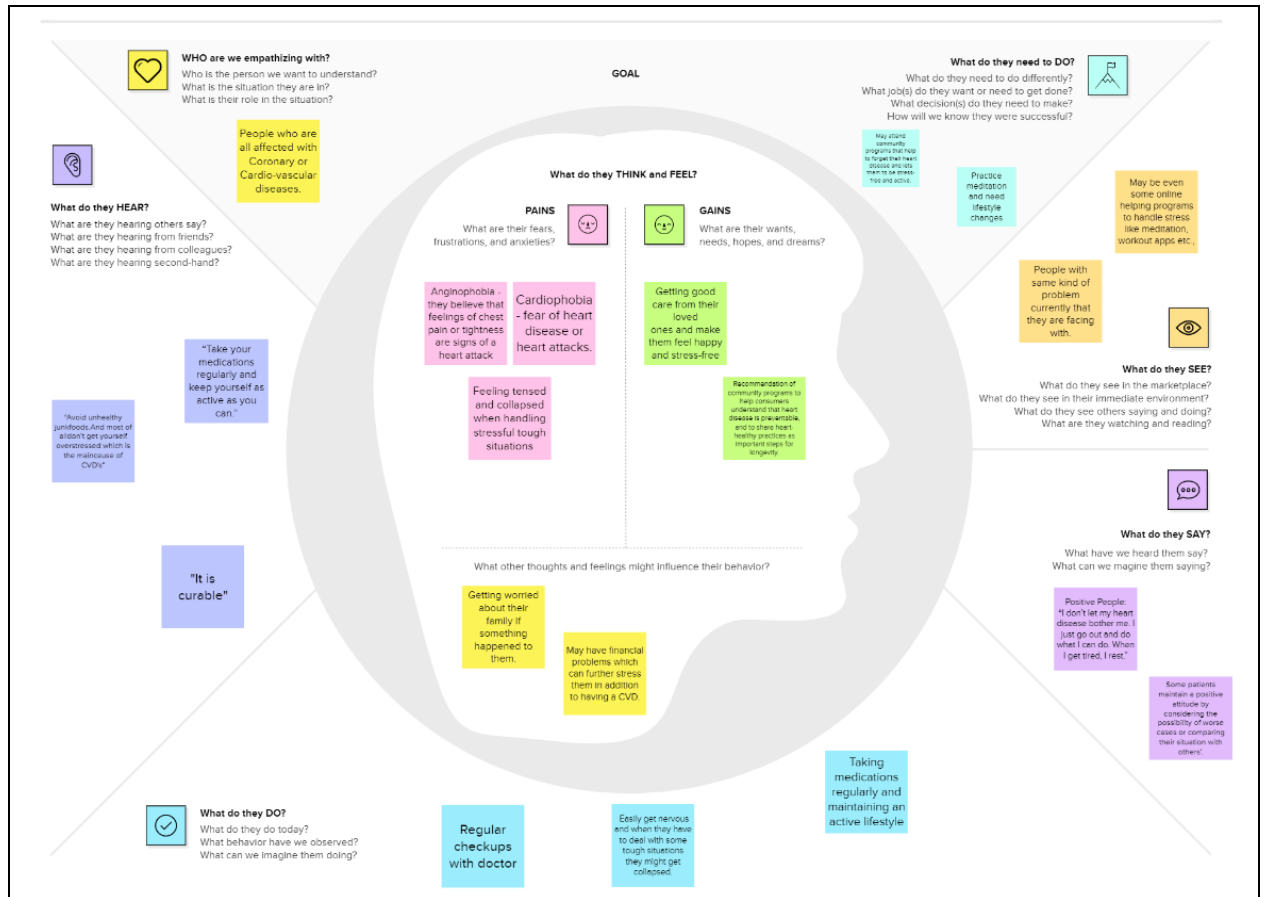
7	Heart diseases prediction with data mining and neural network techniques.	Hybrid techniques are incorporated and various data mining techniques are compared and achieved higher accuracy
8	Heart Disease Prediction using Machine Learning.	Compares the accuracy score of various machine learning algorithms and proposed that random forest algorithm has given highest accuracy score of 90.16%.
9	Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics.	Combines KNN, ANN and SVM using Voting Technique. As ensemble method end up acquiring highest accuracy, more models will increase scope of the trend.
10	An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection.	Two algorithms were hybridized in this paper, Random search algorithm and Random forest model have achieved the highest accuracy of 92.33%.

2.3 Problem Statement Definition

People with unhealthy lifestyles, stress, depression, age above 40 and when their ancestors got heart disease since heart disease is hereditary they are getting the heart disease. The disease is originating from an unhealthy lifestyle. It mostly occurs in the blood valves of the heart. If we don't solve the problem, many people will die at a young age. The death rate due to heart disease will be increased rapidly. We should predict the problem before giving treatment to the patients. As if the problem is predicted early, we can solve it easily and early.

3. IDEATION & PROPOSED SOLUTION


3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming

Step-1: Team Gathering, Collaboration and Select the Problem Statement

Template




Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

- 10 minutes to prepare
- 1 hour to collaborate
- 2-8 people recommended

[Share template feedback](#)



Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

10 minutes

A

Team gathering

Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B

Set the goal

Think about the problem you'll be focusing on solving in the brainstorming session.

C

Learn how to use the facilitation tools

Use the Facilitation Superpowers to run a happy and productive session.

[Open article](#) →

1

Define your problem statement


What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

5 minutes

PROBLEM







How might we [your problem statement]?

Heart disease is said to be a big threat to the people above the age of 40. But now a day even the younger people under the age of 40 between 20-40 might have a high chance of getting coronary artery disease(CAD). This condition occurs when fatty substances called plaque build up inside your coronary arteries. Other reasons are due to higher tension due to stress. In India, people below 60 (20-30 years) get heart failure due to abnormality of heart beat which is due to sudden cause of blockage in valve in heart. It can be treated mostly using ECG. But when someone is technology field for identifying and providing a solution in the field of medicine must undergo several ideas to innovate things that make use of individuals who are all undergo these problems.



Key rules of brainstorming

To run a smooth and productive session

-  Stay in topic.
-  Encourage wild ideas.
-  Defer judgment.
-  Listen to others.
-  Go for volume.
-  If possible, be visual.

Step-2: Brainstorm, Idea Listing and Grouping

2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

🕒 10 minutes

TIP

You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

SAIKRISHNA R

Must have to control their cholesterol level.

Should have to practice self control to overcome hyper tension.

ADHETYA NARAYAN J M

Must maintain their bloodsugar level.

Dont get overstressed and have a happy mindset.

KAVIN B

Monitor the oxygen level in blood.

Practice to meditate regularly for better healthy improvement.

PAAVENDHAN K S

Check and control pulse rate.

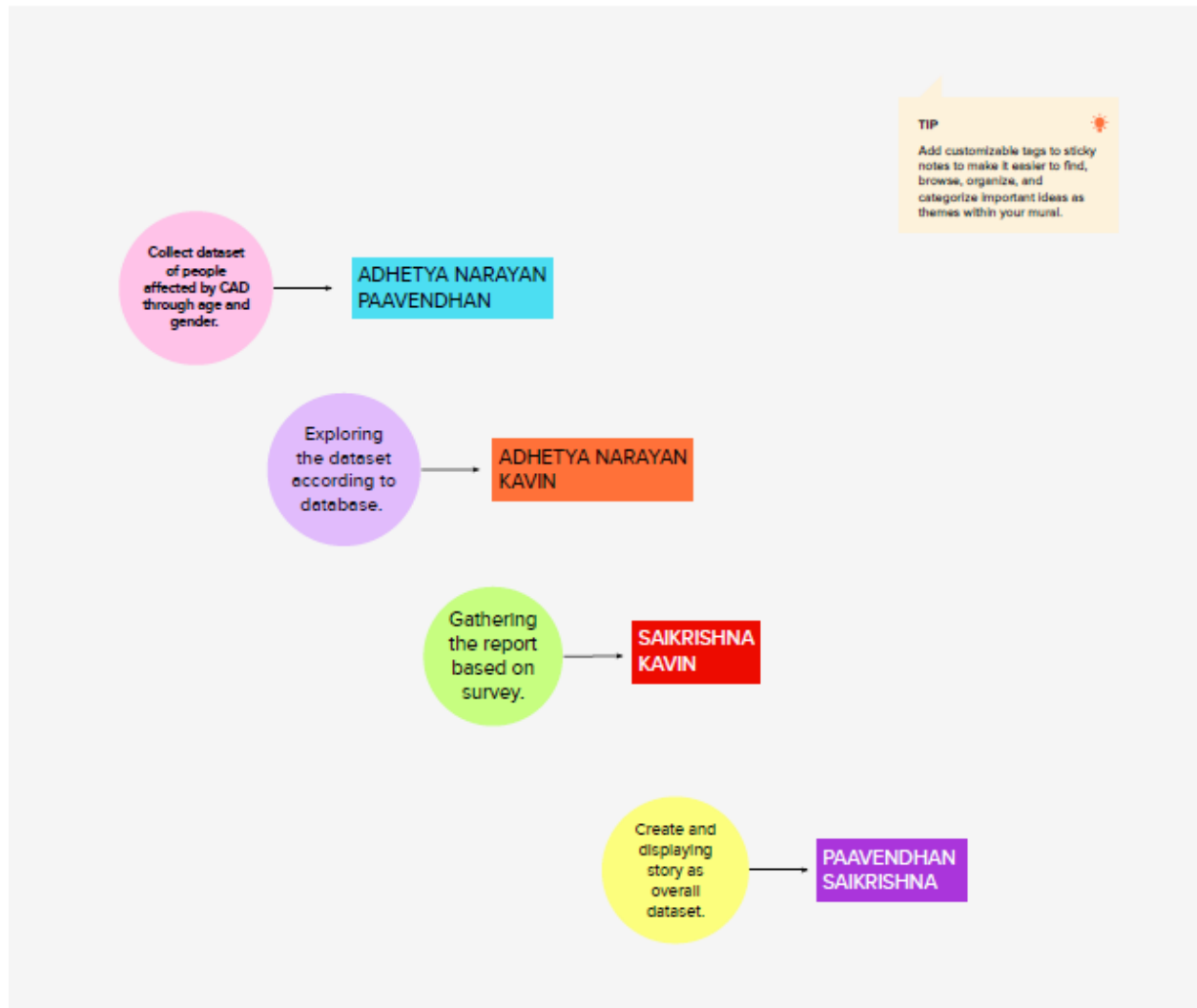
Do physical exercises regularly.

3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

🕒 20 minutes



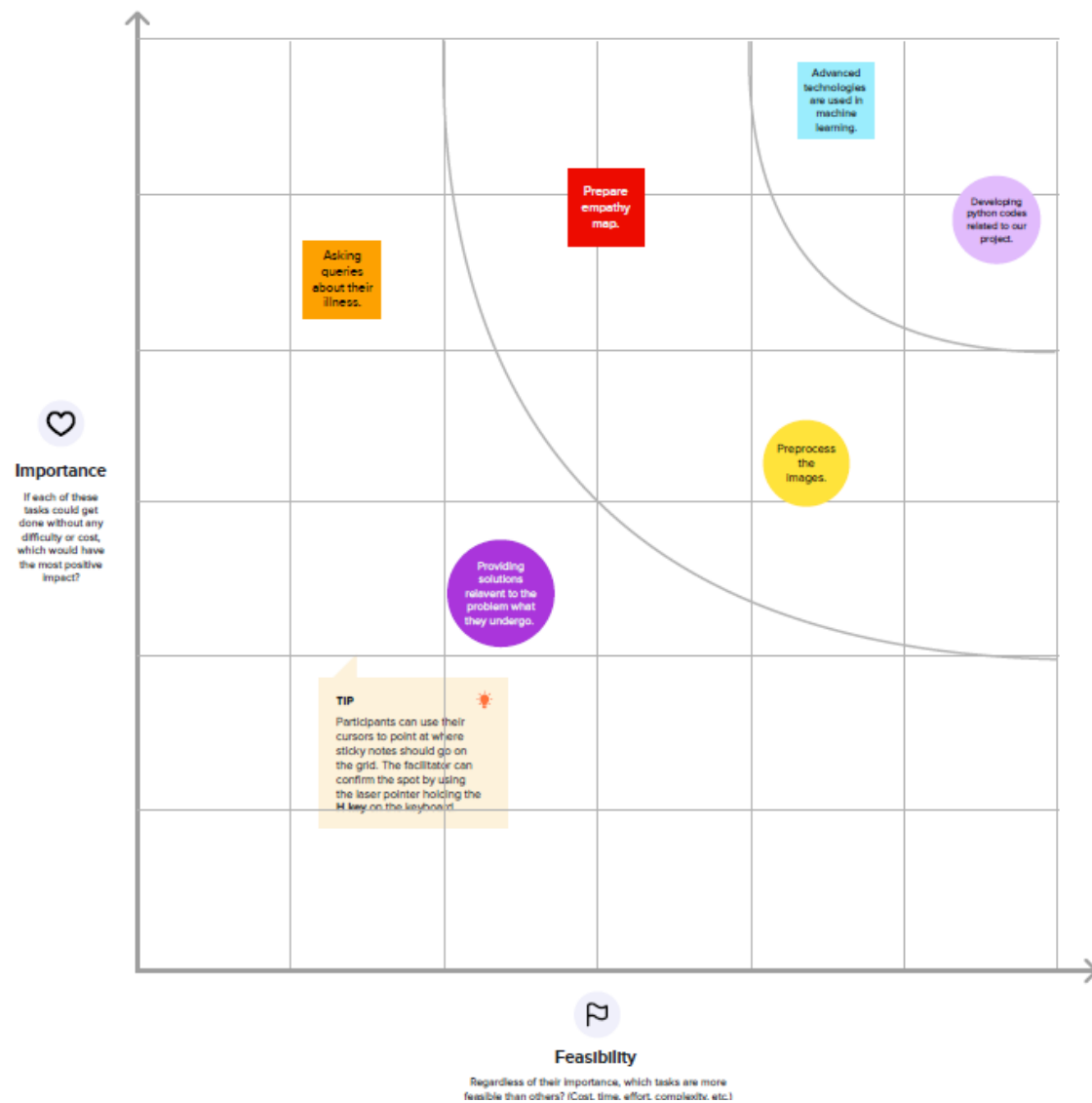
Step-3: Idea Prioritization

4

Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

⌚ 20 minutes



3.3 Proposed Solution

S. NO.	Parameter	Description
1	Problem Statement (Problem to be solved)	To analyse which patients are most likely to suffer from heart disease based on given parameters
2	Idea / Solution description	Creating an interactive dashboard so that people can easily understand whether they have a chance of getting heart disease.
3	Novelty / Uniqueness	Many tests are taken by doctors to detect presence of heart disease. The parameters used are often understood only by medical professional. During the analysis phase, data is extracted and better decisions can be made to diagnose a patient.
4	Social Impact / Customer Satisfaction	Using the power of data analytics, this method will develop awareness among people regarding the risks associated with heart disease as it gives them valuable insights
5	Business Model (Revenue Model)	This project can be converted to an software kit, webpage or even an application which users can interact with. Hospitals could potentially use this for analysis and prediction of heart disease
6	Scalability of the Solution	Based on number of users, performance will vary. So, the application must be memoryefficient and dynamically allocate resources to ensure smooth performance. This is a lightweight application which can be embedded into other projects.

3.4 Problem Solution fit

Define CS, fit into CC	1. CUSTOMER SEGMENT(S) Who is your customer? CS <div> <p>Doctors in hospitals</p> <p>E.g.: Doctors can use this along with the patients' medical data to analyze the risk of heart disease.</p> </div>	6. CUSTOMER CONSTRAINTS CC What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices. <div> <p>Since we are dealing with sensitive medical data, it is not recommended for customers to self-diagnose as it is very risky. It can however be used as a tool to increase awareness regarding this issue.</p> </div>	5. AVAILABLE SOLUTIONS AS Which solutions are available to the customers when they face the problem <div> <p>or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking</p> <p>Customers can go to the doctor for a medical checkup. Based on the test results, doctors will advise them.</p> </div>	Explore AS, differentiate
	2. JOBS-TO-BE-DONE / PROBLEMS J&P Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides. <div> <p>Visualizations give doctors very good insights on the potential chances for a patient to get heart disease. It is also very useful to explain to patients so that they can easily understand the risk factor and take care of themselves to reduce the likelihood of getting heart disease.</p> </div>	9. PROBLEM ROOT CAUSE RC What is the real reason that this problem exists? What is the back story behind the need to do this job? i.e. customers have to do it because of the change in regulations. <div> <p>Not storing and analyzing data properly to help doctors make informed decisions</p> </div>	7. BEHAVIOUR BE What does your customer do to address the problem and get the job done? i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace) <div> <p>Ensure data is stored in an organized and sequential order like an excel sheet for example right from the start so that is ready to be used for analysis.</p> </div>	
3. TRIGGERS TR What triggers customers to act? i.e., seeing their neighbor installing solar panels, reading about a more efficient solution in the news. <div> <p>Patients who have a history with heart disease or those patients who are currently experiencing similar symptoms to those who have heart disease.</p> </div>	10. YOUR SOLUTION SL If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality. If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behavior. <div> <p>To clean data and provide visualizations to help doctors in their diagnosis of patient as well as make customers more aware of this issue.</p> </div>	8. CHANNELS of BEHAVIOUR CH 8.1 ONLINE What kind of actions do customers take online? Extract online channels from #7 8.2 OFFLINE What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development. <div> <p>ONLINE: Users look at the data and compare it with their test results</p> <p>OFFLINE: Doctors use it as a tool to diagnose patients and make accurate predictions.</p> </div>		
4. EMOTIONS: BEFORE / AFTER EM How do customers feel when they face a problem or a job and afterwards? i.e. lost, insecure > confident, in control - use it in your communication strategy & design. <div> <p>Feeling afraid and depressed. Develop a feeling of awareness which mean people</p> </div>				

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

Following are the functional requirements of the proposed solution.

FR No	Functional Requirement(Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Gmail
FR-2	User Confirmation	Confirmation via Email
FR-3	User input	Uploading dataset to platform i.e. IBM Cognos
FR-4	Data pre-processing	Data is prepared and processed by cleaning and checking information
FR-5	Data analysis	Data is analysed to find patterns, relationships and trends
FR-6	Data visualization	Data is converted to various visualizations based on user requirements

4.2 Non-Functional requirements

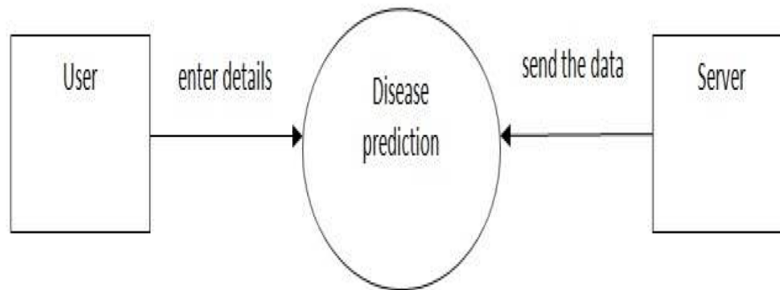
Following are the non-functional requirements of the proposed solution

NFR No.	Non-Functional Requirement	Description
NFR-1	Usability	Even a non- technical person should be able to understanding working of the application and use it
NFR-2	Security	Patient medical data is very sensitive and therefore must be secured so that the data is not misused
NFR-3	Reliability	Application should be fault tolerant. Any changes made need to be committed and backup must be present in case of system crash.
NFR-4	Performance	Application needs to be lightweight and efficient in terms of memory and resources used. Different users have different systems so that must be taken into account.
NFR-5	Availability	Data should be available to users at all times. Data integrity needs to be maintained.
NFR-6	Scalability	Application needs to be able to handle multiple or large amounts of data and produce advanced visualizations without affecting performance.

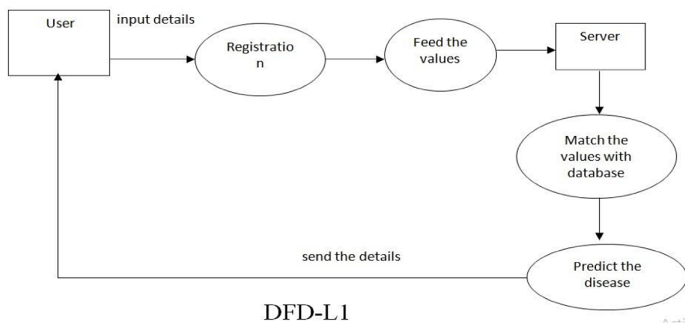
5. PROJECT DESIGN

5.1 Data Flow Diagrams

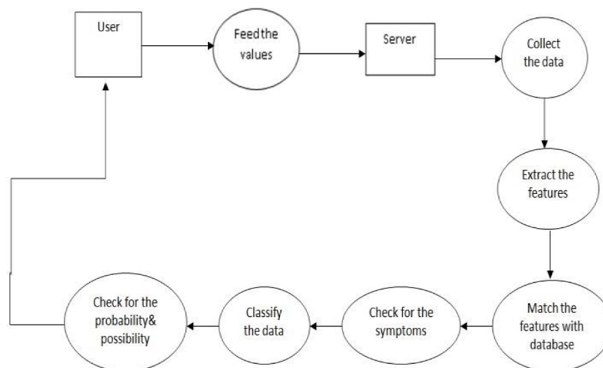
DFD-L0



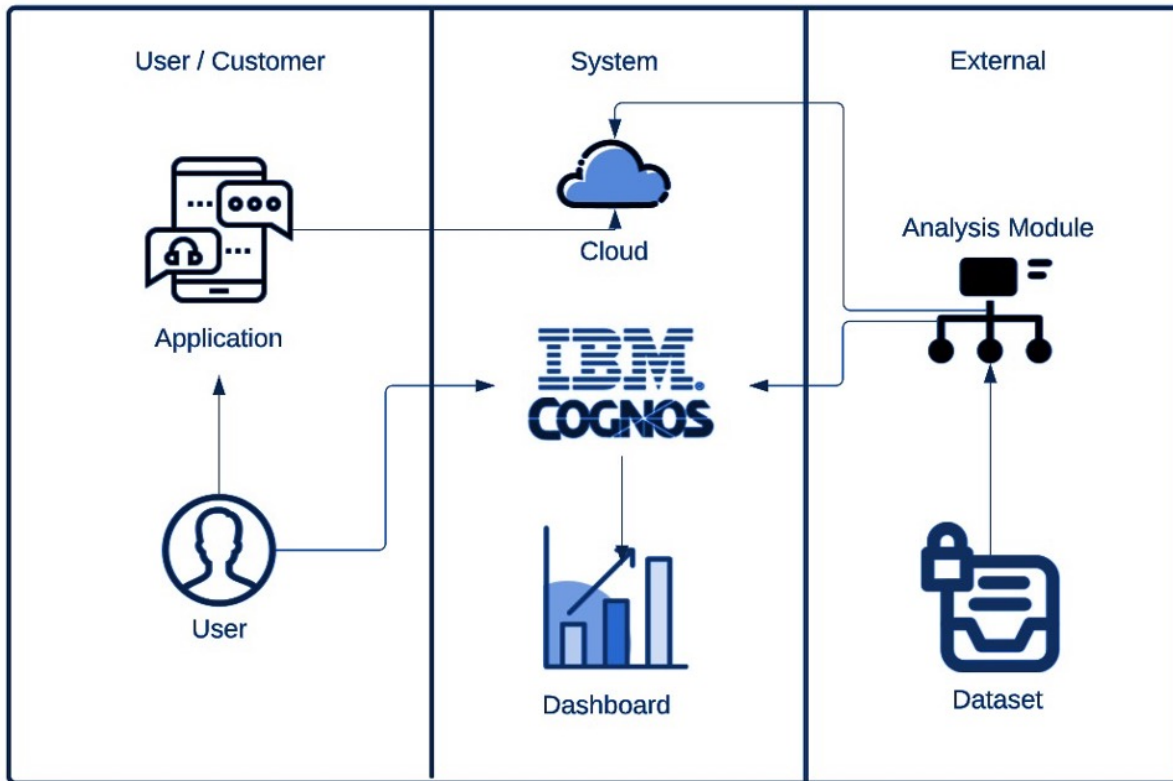
DFD-L1



DFD-L2



5.2 Solution & Technical Architecture



Components and Technologies

S.No	Component	Description	Technology
1	User interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	IBM Cognos
2	Storage infrastructure	Medical data related to heart is uploaded in cloud through cloud	IBM Cloud
3	Working with dataset	Uploading, cleaning and pre-processing dataset	IBM Cognos + Cloud
4	Data exploration	Uploaded data is explored to identify trends	IBM Cognos
5	Data visualization	Multiple types of graphs are shown according to patient medical data and requirements	IBM Cognos Dashboard
6	Cloud database	Database Service on Cloud	IBM Cognos, IBM Cloud etc.
7	Viewing Data	User logs in to application to view visualizations for uploaded data	IBM Cognos Dashboard

Application characteristics

S.No	Characteristics	Description	Technology
1	Open-Source Frameworks	List the open-source frameworks used	IBM Cognos, IBM Cloud, IBM Watson
2	Security Implementations	Secure user information and data	Active Directory
3	Scalable Architecture	Supports various data sizes	Web 3.0 IBM Cloud
4	Availability	Multi page layout providing various visualizations of data and provide full support irrespective of platform and device specifications	Cognos Business Intelligence Server
5	Performance	Withstand huge data and process them without crashing	IBM Cognos, Performance Management Hub

5.3 User Stories

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 High email and click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register and access the dashboard with Facebook Login	Low	Sprint-2

		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
		USN-5	As a user, I can log into the application by entering email and password		High	Sprint-1
		USN-6	I can access the dashboard and access the analytics reports based on data uploaded.		High	Sprint-3

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

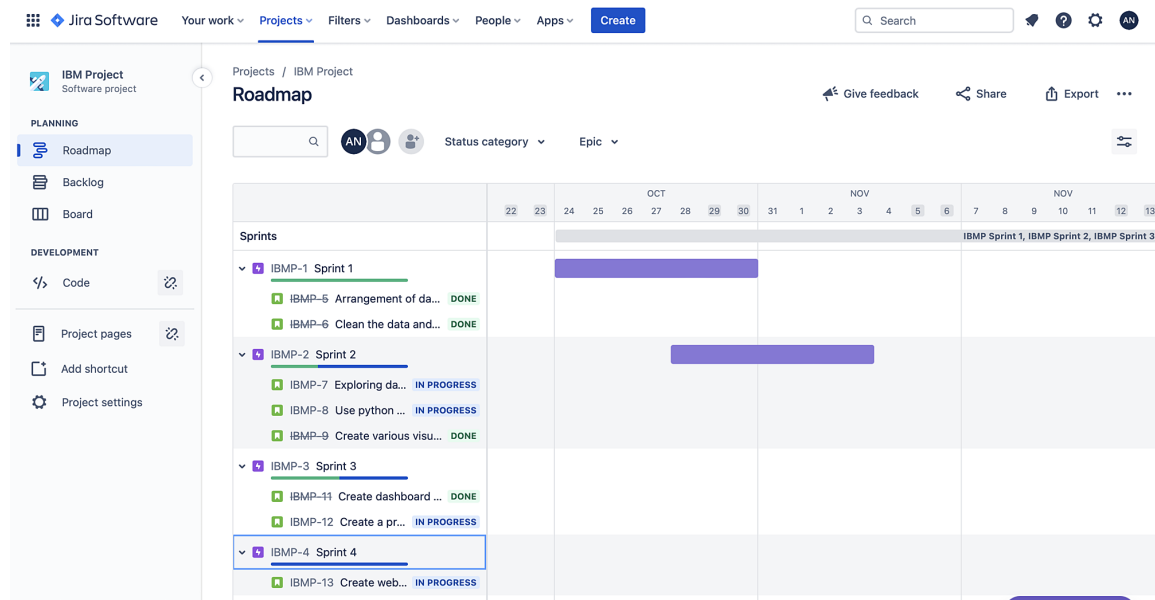
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority Team	Members
Sprint 1	Arrangement of data set	USN -1	Upload the dataset in IBM Cognos platform and create data module	5	High	Sai Krishna
Sprint 1		USN -2	Clean the data and create simple visualizations using python libraries	3	High	Adhetya Narayan, Kavın B
Sprint 2	Exploring data and creating mode	USN -3	As an analyst, I would like to find relationships between attributes to understand its importance	2	Low	Paavendhan K.S
Sprint 2		USN -4	Use python to analyse correlation between variables. Visualised in the form of correlation matrix and use classifier algorithms like decision tree	3	Medium	Adhetya Narayan
Sprint 2		USN -5	Create various visualizations using IBM	3	High	Sai Krishna,

			Cognos			Paavendhan K.S
Sprint 3	Dashboard	USN -6	Create dashboard in IBM Cognos to get a clear understanding of visualizations	3	Medium	Kavin B
Sprint 3	Story	USN -7	As an analyst, I will use IBM Cognos to create a story to understand the animated presentation of dataset	3	Medium	Paavendhan K S
Sprint 4	Creation of web page	USN -8	Create webpage so that users can easily access the dashboard and story created in IBM Cognos	3	High	Adhetya Narayan

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start date	Sprint End Date	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint 1	20	6 Days	24 Oct 2022	29 Oct 2022	5	29 Oct 2022
Sprint 2	20	6 Days	31 Oct 2022	05 Nov 2022	5	05 Nov 2022
Sprint 3	20	6 Days	07 Nov 2022	12 Nov 2022	5	14 Nov 2022
Sprint 4	20	6 Days	14 Nov 2022	19 Nov 2022	5	18 Nov 2022

6.3 Reports from JIRA



7. CODING & SOLUTIONING

7.1 Machine learning models

Using classification algorithms

We will test some classification algorithms: Logistic regression, svm, stochastic gradient descent , decision tree, random forest.

```
In [8]: from sklearn.linear_model import LogisticRegression
        from sklearn import svm
        from sklearn.linear_model import SGDClassifier
        from sklearn import tree
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.metrics import accuracy_score
        from sklearn.model_selection import train_test_split

        y = data['Heart Disease']

        X = data.drop('Heart Disease',axis=1)
        X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state=0)
        LR_classifier = LogisticRegression(random_state=0)
        clf = svm.SVC()
        sgd=SGDClassifier()
        forest=RandomForestClassifier(n_estimators=20, random_state=12,max_depth=6)
        treee = tree.DecisionTreeClassifier(criterion = 'entropy',random_state=0,max_depth = 6)
        LR_classifier.fit(X_train, y_train)
        clf.fit(X_train, y_train)
        sgd.fit(X_train, y_train)
        treee.fit(X_train, y_train)
        forest.fit(X_train, y_train)
```

```
C:\Users\adhet\anaconda3\envs\project\lib\site-packages\sklearn\linear_model\_logistic.py:444: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

```
Out[8]: RandomForestClassifier
        RandomForestClassifier(max_depth=6, n_estimators=20, random_state=12)
```

Training model

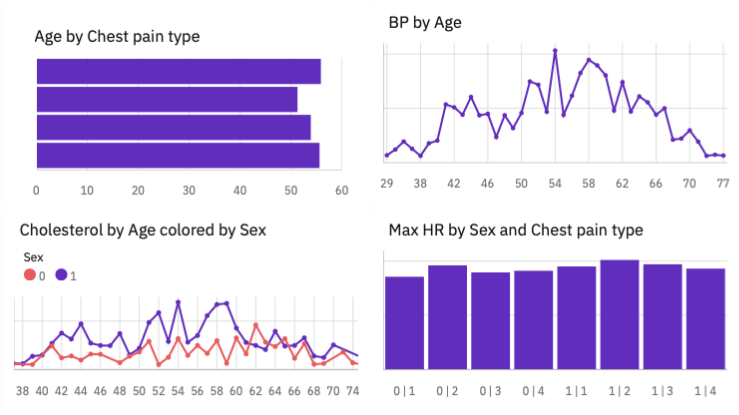
```
In [9]: y_pred=LR_classifier.predict(X_train)
        y_predsvm=clf.predict(X_train)
        y_predsgd=sgd.predict(X_train)
        y_predtree=treee.predict(X_train)
        y_predforest=forest.predict(X_train)
```

```
In [10]: print(accuracy_score(y_train, y_pred))
         print(accuracy_score(y_train, y_predsvm))
         print(accuracy_score(y_train, y_predsgd))
         print(accuracy_score(y_train, y_predtree))
         print(accuracy_score(y_train, y_predforest))
```

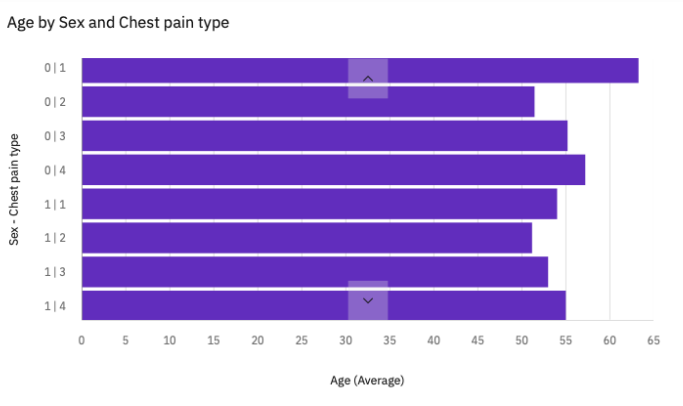
```
0.8677248677248677
0.6984126984126984
0.7248677248677249
0.9841269841269841
0.9735449735449735
0.8677248677248677
0.6984126984126984
0.7248677248677249
0.9841269841269841
0.9735449735449735
```

7.2 Dashboard

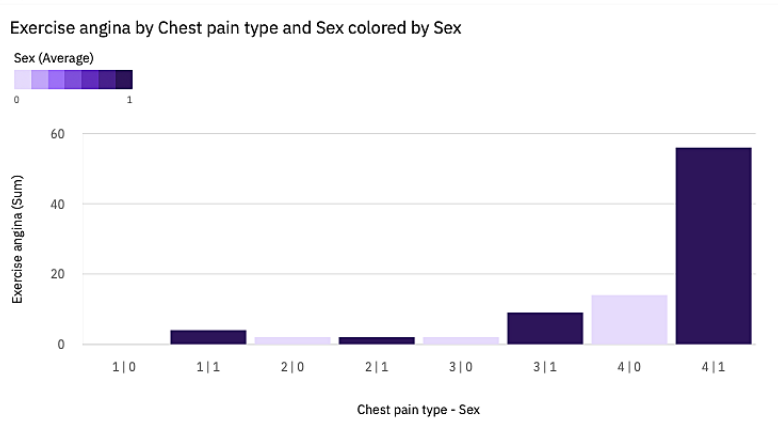
Tab 1



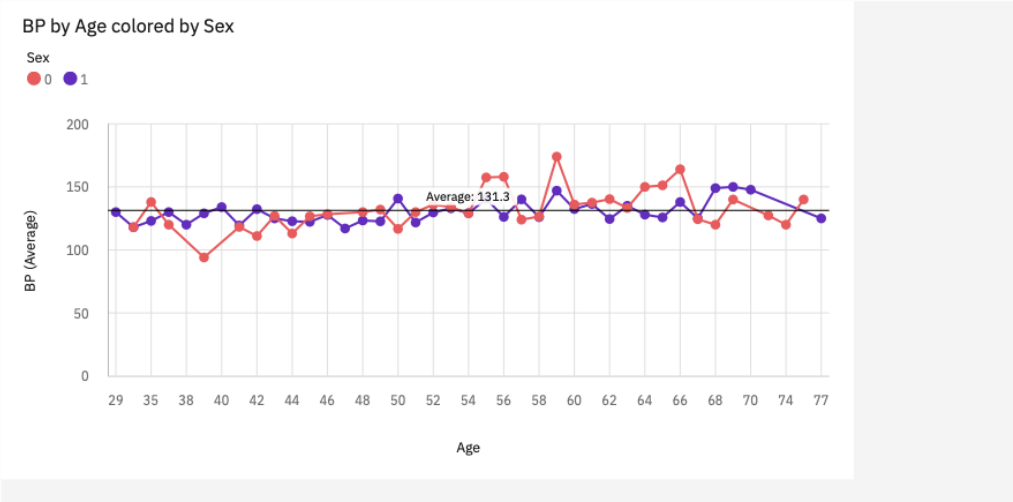
Tab 2



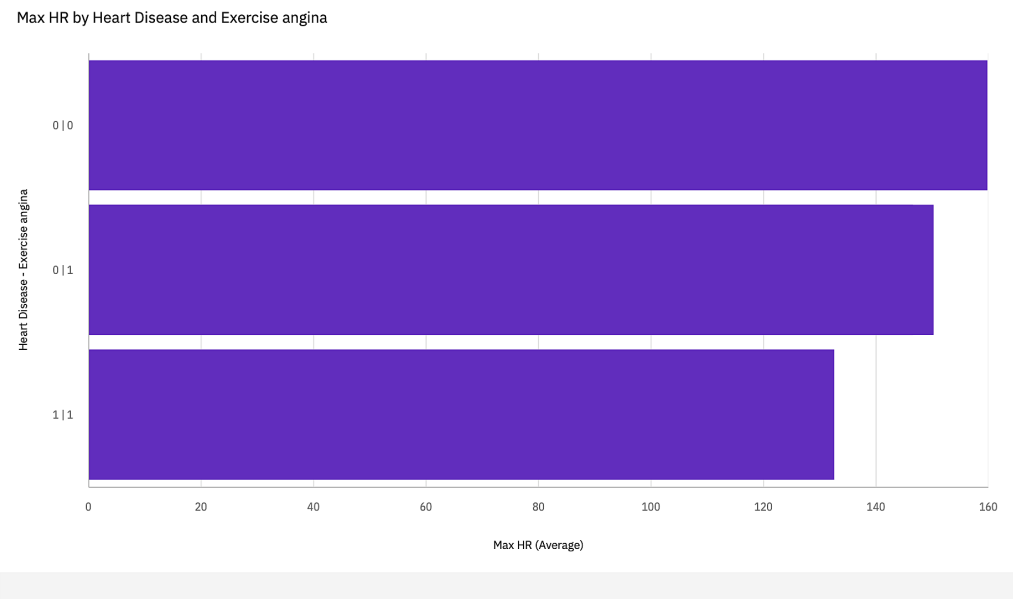
Tab 3



Tab 4



Tab 5



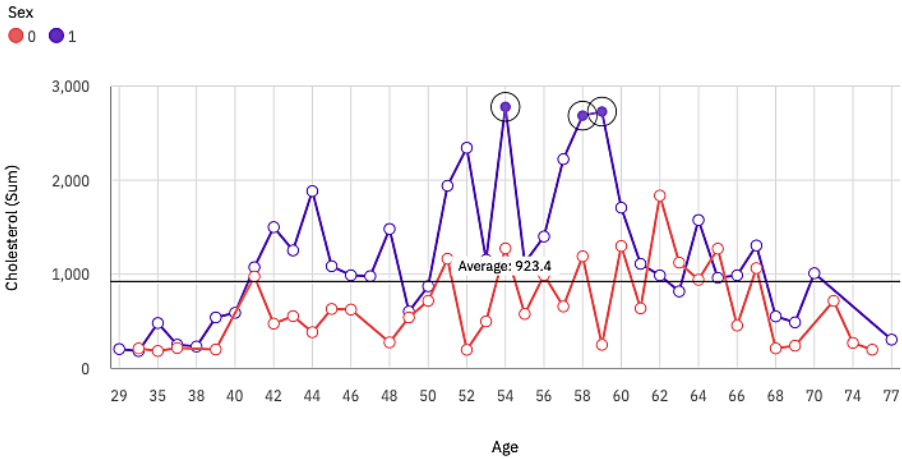
Tab 6

Heart Disease for Chest pain type and Sex

Heart Disease		2	3	4	Summary
0	4	16	32	35	87
1	16	26	47	94	183
Summary	20	42	79	129	270

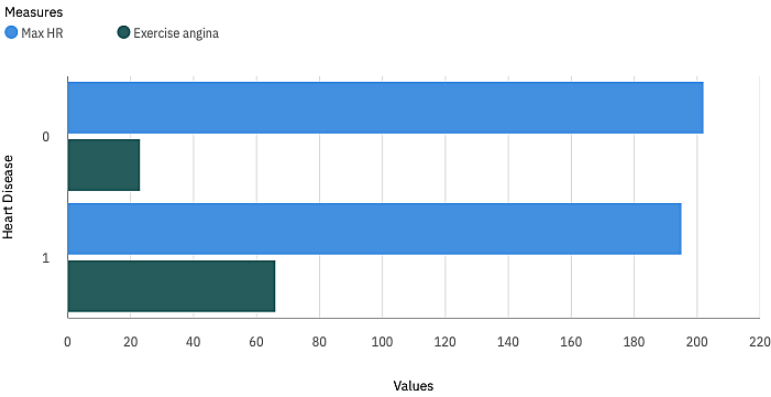
Tab 7

Cholesterol by Age colored by Sex



Tab 8

Max HR and Exercise angina by Heart Disease



8. TESTING

8.1 Test Cases

```
In [11]: y_pred=LR_classifier.predict(X_test)
y_predsvm=clf.predict(X_test)
y_predsgd=sgd.predict(X_test)
y_predtree=tree.predict(X_test)
y_predforest=forest.predict(X_test)

In [12]: print(accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_predsvm))
print(accuracy_score(y_test, y_predsgd))
print(accuracy_score(y_test, y_predtree))
print(accuracy_score(y_test, y_predforest))

0.8518518518518519
0.7407407407407407
0.5802469135802469
0.6419753086419753
0.7777777777777778
0.8518518518518519
0.7407407407407407
0.5802469135802469
0.6419753086419753
0.7777777777777778

In [13]: import pickle
pickle.dump(forest, open('model.pkl', 'wb'))
```

Logistic Regression is the best model

8.2 User Acceptance Testing

User Input Features

Enter your age:
55.00

Sex
1

Chest pain type
1

BP:
125.00

Serum cholestoral in mg/dl:
234.00

Fasting blood sugar over 120
0

Heart disease Prediction App

This app predicts If a patient has a heart disease. Data obtained from
Kaggle:<https://www.kaggle.com/datasets/rishidamrta/heart-disease-prediction>

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	
0	55.0000	1	1	125.0000	34.0000	0	1	125.0000	0	0.2000	1.0000	2	1

Prediction Probability

	0	1
0	0.5502	0.4498

9. RESULTS

9.1 Performance Metrics

```
In [11]: y_pred=LR_classifier.predict(X_test)
          y_predsvm=clf.predict(X_test)
          y_predsgd=sgd.predict(X_test)
          y_predtree=tree.predict(X_test)
          y_predforest=forest.predict(X_test)

In [12]: print(accuracy_score(y_test, y_pred))
          print(accuracy_score(y_test, y_predsvm))
          print(accuracy_score(y_test, y_predsgd))
          print(accuracy_score(y_test, y_predtree))
          print(accuracy_score(y_test, y_predforest))

0.8518518518518519
0.7407407407407407
0.5802469135802469
0.6419753086419753
0.7777777777777778
0.8518518518518519
0.7407407407407407
0.5802469135802469
0.6419753086419753
0.7777777777777778

In [13]: import pickle
          pickle.dump(forest, open('model.pkl', 'wb'))
```

Logistic Regression is the best model

10. ADVANTAGES & DISADVANTAGES

Advantages:

- Very easy to use and understand for the user.
- Secure
- Dashboard provides useful insight to the user
- Can be used to easily classify users who have heart disease to those who do not.

Disadvantages:

- User needs to know value of all parameters.
- Does not provide suggestions to user.
- Cannot enter null value so will be an issue if user does not know value of all parameters.
- Still needs more work to improve accuracy

11. CONCLUSION

If not detected early, people who get heart disease can get a heart attack or stroke which could be fatal. Therefore, it is better for users to adopt a healthy lifestyle to minimize risk of getting heart disease. This website can help users analyze their health based on certain parameters and hopefully help them get the required treatment early

on if they show symptoms of heart disease.

12. FUTURE SCOPE

For this website, a lot of parameters are required some of which the user may not be aware of. Therefore, it would be better if it is possible to simplify the model to improve usability.

13. APPENDIX

Source Code

<https://github.com/IBM-EPBL/IBM-Project-4903-1658742306>

Project Demo Link

https://drive.google.com/file/d/12LKDLLeQF46dItlRndCQydIWA-rDI-qPA/view?usp=share_link