VISUALIZING AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASH BOARD

Date	19 November 2022
Team ID	PNT2022TMID12915
Team Members Name	Sree Nanda R, Nithish V, Raja Keerthana N, Suriya Sankameswaran S

1. INTRODUCTION

1.1 PROJECT OVERVIEW

The overview of this project is that first suitable data set is collected. Then, the data is pre- processed, and the different types of data present in the dataset are visualized using the Pandas library. Then, in order to predict the possibility of heart diseases, a machine learning model is built using Random Forest Classifier. Various metrics like accuracy, confusion matrix, etc. are viewed. Finally, a descriptive report is generated for the data which can be viewed in a HTML page.

1.2 PURPOSE

The aim of this project is to visualize the existing data for heart diseases using Data Analytics and to predict the possibility of heart diseases for people with their medical records using Machine Learning Techniques. The ultimate purpose of this project is to develop a model that could save people's lives or at least incentivise people to lead a healthy lifestyle.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

The main problem that exists in this field, is having a high accurate model for predicting heart diseases. The other problem that exists, is a dashboard with all the required information and visualizations displayed in one place. Our project solves both these issues.

2.2 REFERENCES

S.NO	TITLE	AUTHORS	YEAR	JOURNAL	DESCRIPTION
1	Visualization and Prediction of Heart Diseases Using Data Science Framework	Vaibhav Gupta; Vaibhav Aggarwal; Shagun Gupta; Neeti Sharma; Kiran Sharma; Neetu Sharma	2021 August	2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)	The main aim of this paper is to use various classification algorithms of data science framework to somehow detect the chances of having a heart disease. Also, the main aim of this research paper is to find out the most efficient classification algorithm that can help us to detect heart diseases at early stage. This algorithm can be used on heart records of the patient or by using it on classification reports.
2	Heart Disease Prediction using Exploratory Data Analysis	R.Indrakumari ; T.Poongodi; Soumya Ranjan Jena	2020 July	International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020	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, K-means 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.

S.NO	TITLE	AUTHORS	YEAR	JOURNAL	DESCRIPTION
3	Prediction of Cardio-vascular Disease Using Machine Learning Algorithms	Kumar G; Dinesh; K Arumugaraj; Kumar D; Santhosh; V Mareeswari	2018 March	2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)	In this paper, data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This paper proposes a prediction model to predict whether people have a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random Forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.
4	Heart Disease Classifi-cation and Risk Prediction by Using Convolution al Neural Network	V.Archana Reddy; K Venkatesh Sharma	2021 June	International Journal of Aquatic Science, Vol 12, Issue 02, 2021	The proposed work in this paper includes Machine Learning-based classifiers to assess classification accuracy and the use of Deep Learning techniques to enhance the accuracy of heart disease prediction. The classification accuracy of the KNN, SVM, and Naive Bayes Classifiers is also compared.

S.NO	TITLE	AUTHORS	YEAR	JOURNAL	DESCRIPTION
5	Using Data Visualization to Analyze the Correlation of Heart Disease Triggers and Using Machine Learning to Predict Heart Disease	Xinyu Zhang	2021 April	IMIP '21: 2021 3rd International Conference on Intelligent Medicine and Image Processing	In this paper an in-depth understanding of data analysis and machine learning-related knowledge, data analysis and data training are carried out on a dataset containing 14 columns of features. First, Python is used to visualize and analyze data. And then train_test_split is used to divide the data into the training set and the learning set. At last, three methods including logistic regression, decision tree classifier, and random forest classifier are used to train the data and observe which method gets the best effect. This article mainly uses numpy, matplotlib, pandas, seaborn and scikit-learn libraries in Python language for data analysis and processing.
6	Heart Disease Prediction Using Machine Learning	Chaimaa Boukhatem; Heba Yahia Youssef; Ali Bou Nassif	2022 Febru-ary	2022 Advances in Science and Engineering Technology International Conferences (ASET)	This paper presents several machine learning approaches for predicting heart diseases, using data of major health factors from patients. The paper demonstrated four classification methods: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), to build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models were evaluated based on the accuracy, precision, recall, and F1-score. The SVM model performed best with 91.67% accuracy.

S.NO	TITLE	AUTHORS	YEAR	JOURNAL	DESCRIPTION
7	Heart disease prediction using machine learning algorithms	Harshit Jindal; Sarthak Agrawal; Rishabh Khera; Rachna Jain; Preeti Nagrath	2020 October	IOP Conference Series: Materials Science and Engineering, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India	The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. It prepares a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. It uses different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease.
8	A novel approach for heart disease prediction using strength scores with significant predictors	Armin Yazdani; Kasturi Dewi Varathan; Yin Kia Chiam; Asad Waqar Malik; Wan Azman; Wan Ahmad	2021 June	BMC Medical Informatics and Decision Making	This paper is motivated by the gap in the literature, thus proposes an algorithm that measures the strength of the significant features that contribute to heart disease prediction. The study is aimed at predicting heart disease based on the scores of significant features using Weighted Associative Rule Mining.

2.3 PROBLEM STATEMENT DEFINITION

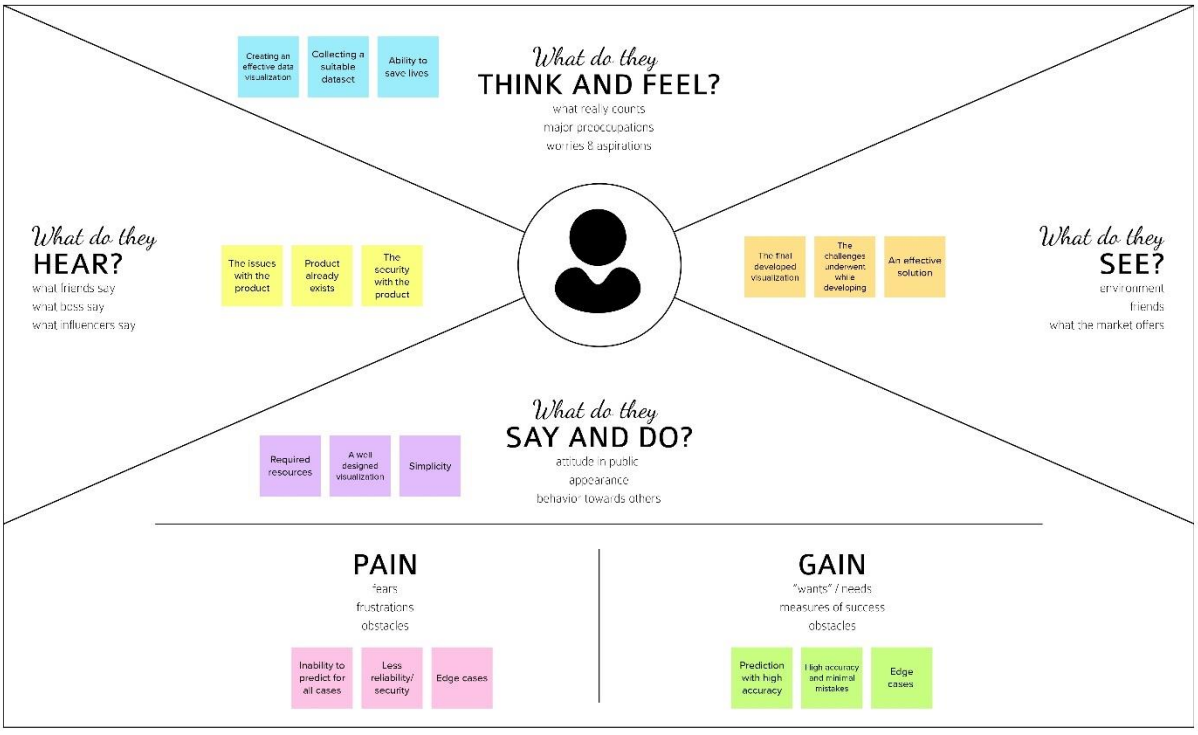
Create a problem statement to understand your customer's point of view. The Customer Problem Statement template helps you focus on what matters to create experiences people will love.

A well-articulated customer problem statement allows you and your team to find the ideal solution for the challenges your customers face. Throughout the process, you'll also be able to empathize with your customers, which helps you better understand how they perceive your product or service.

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	a patient	check the chances of me getting a heart diease	it takes a long time	the website is not responsive	Frustrated
PS-2	a patient	check the chances of me getting a heart diease	is less accurate	proper methods are not used	Disappointed
PS-3	a patient	check the chances of me getting a heart disease	is costly	overpriced	Frustrated

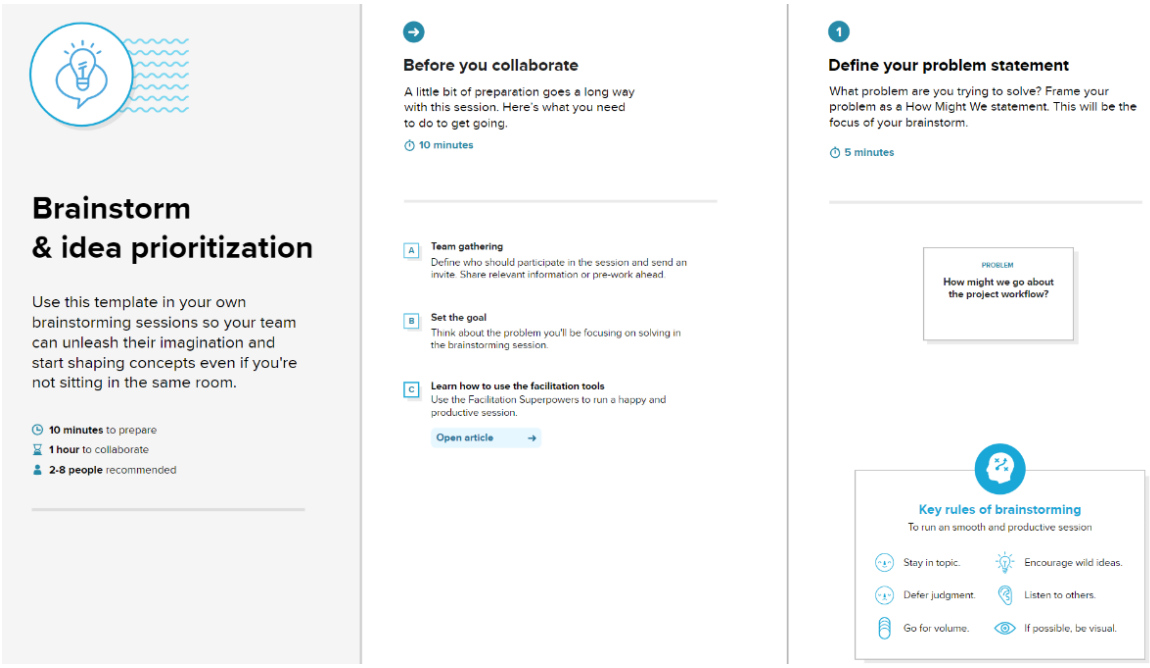
3. IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



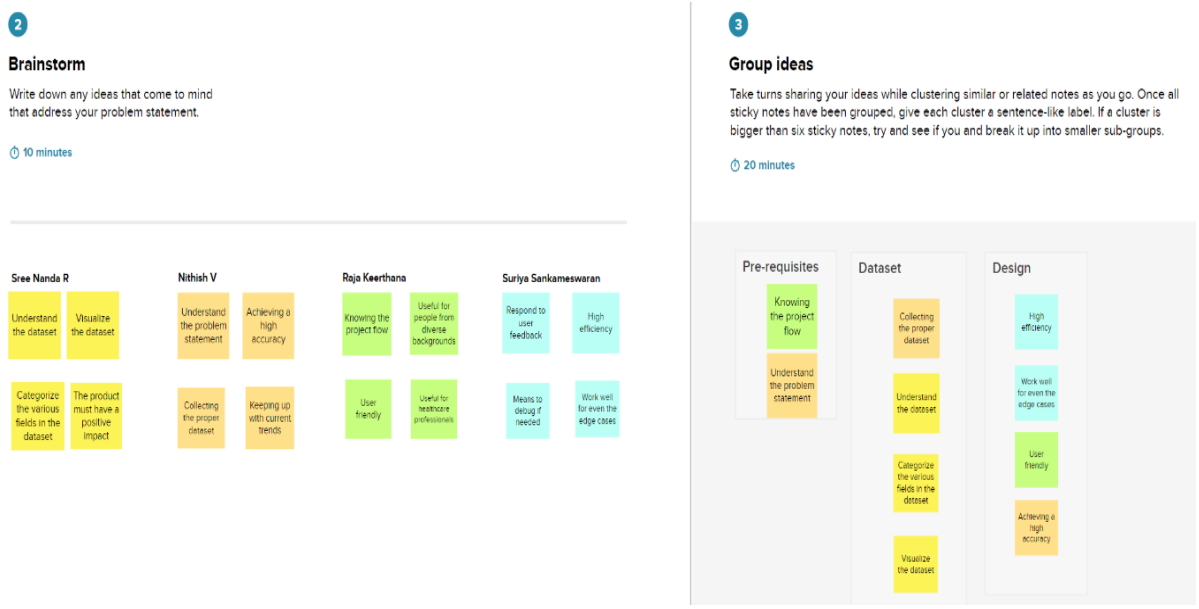
3.2 IDEATION AND BRAINSTORMING

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich number of creative solutions.

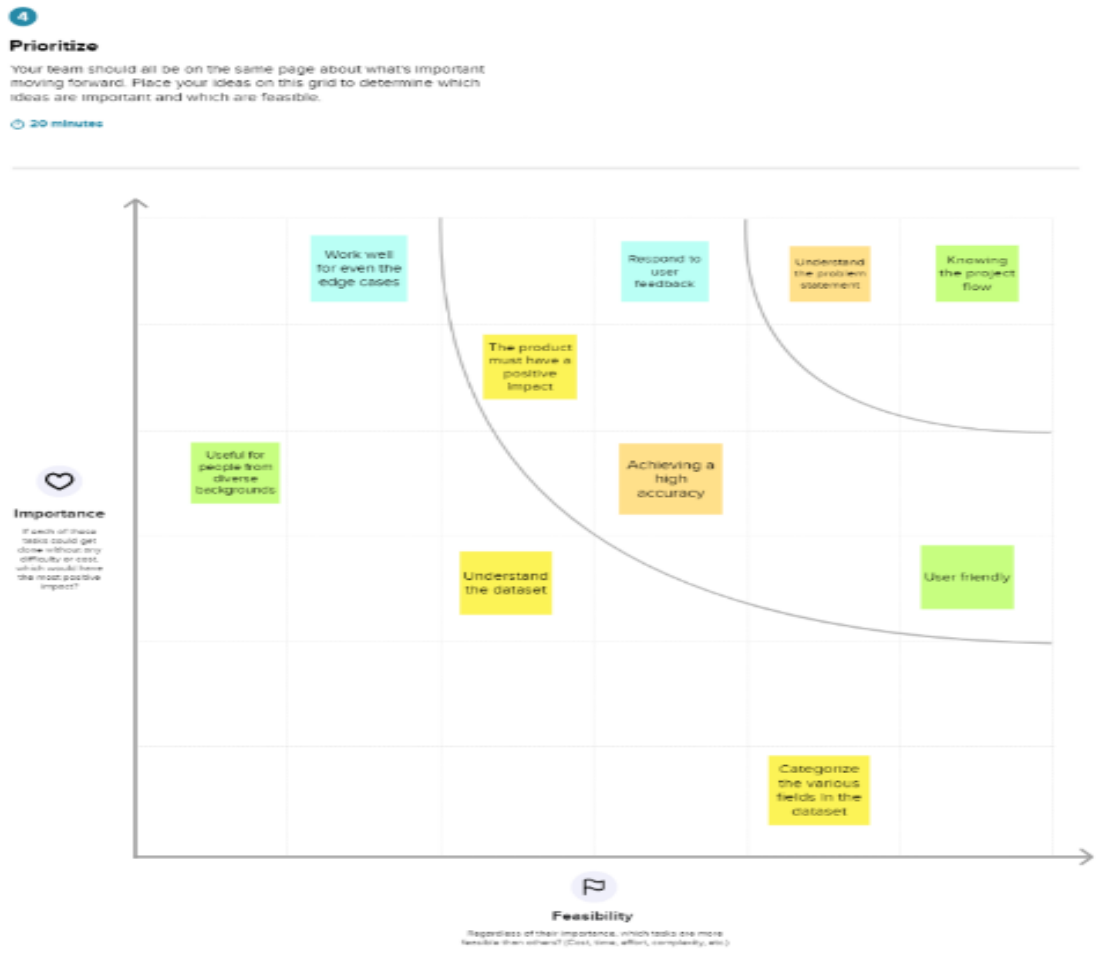


Step-1: Team Gathering, Collaboration and Select the Problem Statement

Step-2: Brainstorm, Idea Listing and Grouping



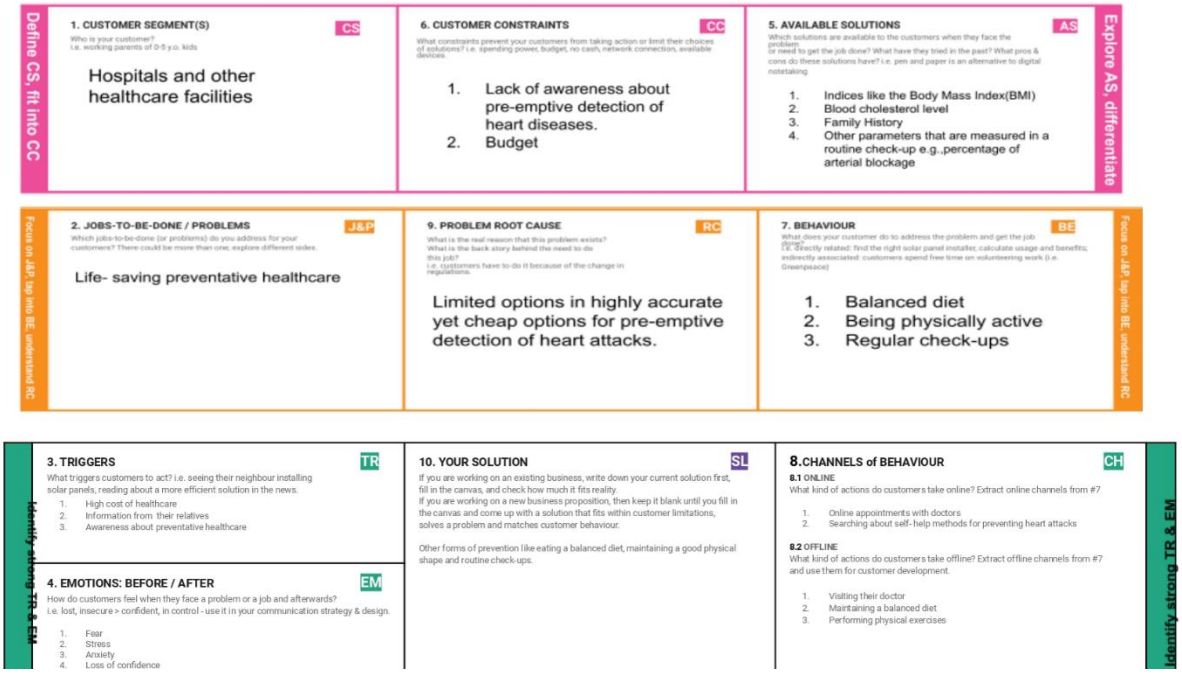
Step-3: Idea Prioritization



3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Visualizing And Predicting Heart Diseases with An Interactive Dash Board; using the dataset to predict which patients are most likely to suffer from a heart disease in the near future using the features given
2.	Idea / Solution description	The risk factors that cause heart disease is considered and predicted using K-means algorithm and the analysis is carried out using a publicly available data for heart disease. 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain are considered. To predict the heart disease, K-means clustering algorithm is used along with data analytics and visualization tool.
3.	Novelty / Uniqueness	K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The role of exploratory data using tableau provided a visual appealing and accurate clustering experience.
4.	Social Impact / Customer Satisfaction	Heart stroke and vascular disease are the major cause of disability and premature death. Chest pain is the key to recognize the heart disease. By predicting the possibility of heart diseases in at- risk patients, many lives can be saved.
5.	Business Model (Revenue Model)	This would be highly useful in hospitals and other related healthcare facilities.
6.	Scalability of the Solution	Increasing the size of dataset would significantly improve the accuracy. A better initialization technique in K- means clustering algorithm can also improve the accuracy.

3.4 PROBLEM SOLUTION FIT



4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL 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 through Gmail
FR-2	User Confirmation	Confirmation via Email
FR-3	User Input	The medical records of the user can be directly obtained from the user or can be obtained from their doctor.
FR-4	Reports and Visualizations	The final results will be generated in the form of an extensive report along with a creative visualization.

4.2 NON-FUNCTIONAL REQUIREMENTS

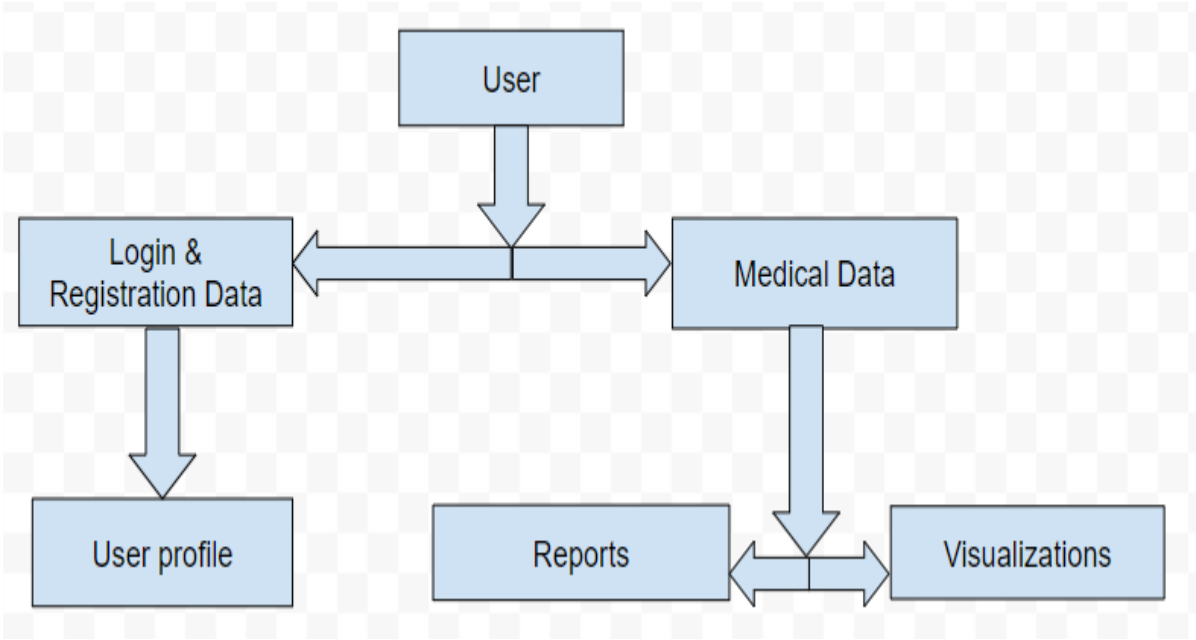
Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The user interface is easy to use for everyone and the entire process could be completed in a reasonable amount of time.
NFR-2	**Security**	Medical records are private, so authentication using passwords and OTP confirmation are used.
NFR-3	**Reliability**	The site undergoes regular maintenance to ensure that there are no issues.
NFR-4	**Performance**	The output will be generated without any delays.
NFR-5	**Availability**	The site will be available 24x7.
NFR-6	**Scalability**	By increasing the number of servers, more users can use the site at the same time.

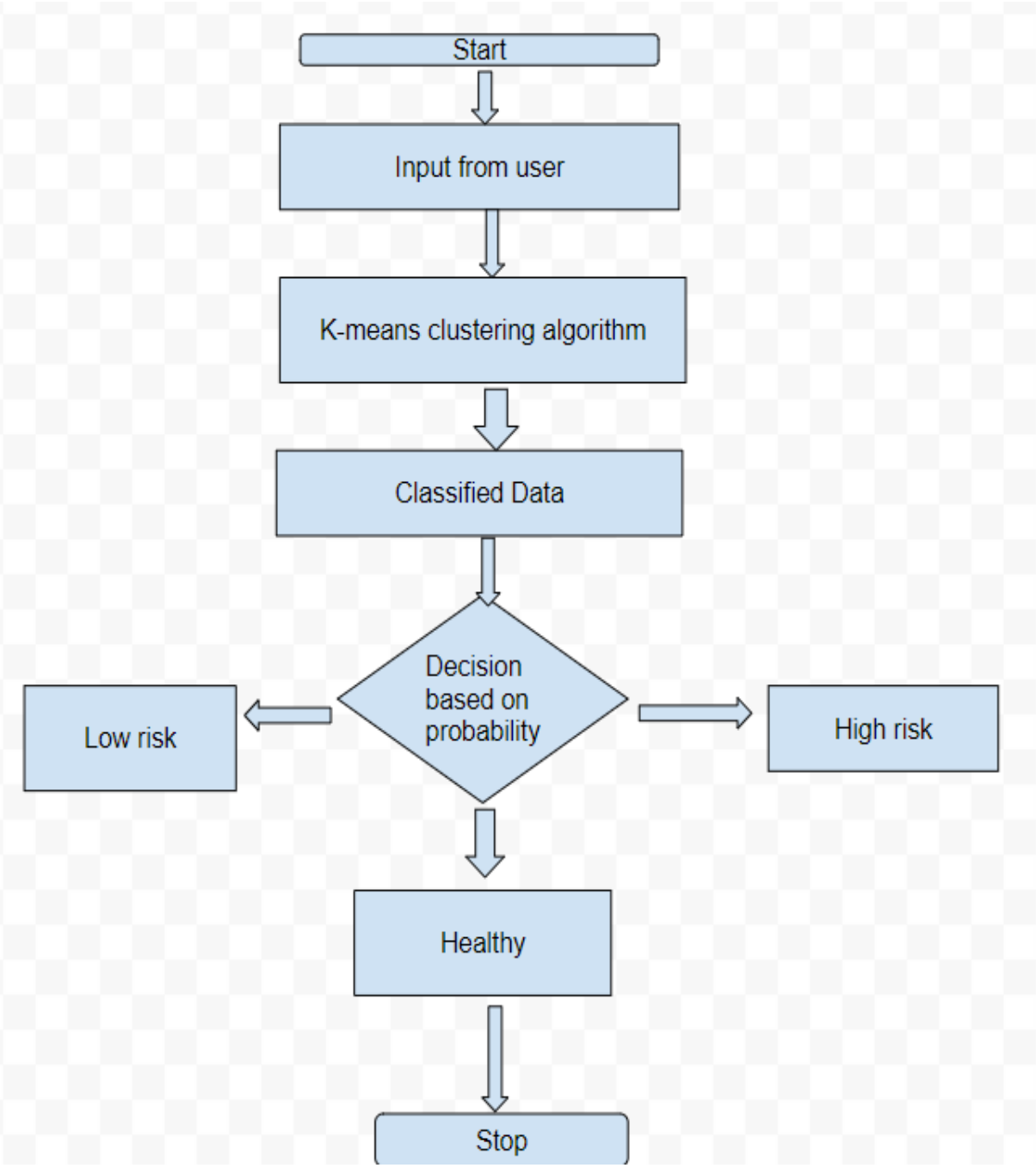
5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS

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.



5.2 SOLUTION & TECHNICAL ARCHITECTURE



5.3 USER STORIES

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web 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
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can receive confirmation email & click confirm	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can enter after providing the valid details	High	Sprint-1
	Dashboard	USN-6	As a user, I can view visualizations and a detailed report after entering my details	I can view the correct results if the details entered are accurate	High	Sprint-2
Customer Care Executive	Helpdesk	USN-7	I help the users with any queries and direct any reported issues to the administrator	I can view the feedback from the users	Medium	Sprint-3
Administr--ator	Maintenance	USN-8	I make sure all the servers are working properly	I can check the status pf the servers	High	Sprint-3
	Relaying queries	USN-9	I make sure all the reported queries are addressed by delivering the queries to the concerned people	I can make sure there are no issues in the site	High	Sprint-4
	Security	USN-10	I make sure no unwanted users get access to the personal medical records	I use various network security tools to have a secure site	High	Sprint-4

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	Dataset Collection and Pre-Processing	USN-1	Users can enter their data and that data is pre-processed for further visualization	1	High	All 4
Sprint-2	Model Training	USN-2	A machine learning model is generated for the inputs using Random Forest Classifier	2	High	All 4
Sprint-3	Performance Testing	USN-3	Various performance metrics are generated for the trained model	1	High	All 4
Sprint-4	Dashboard	USN-4	A dashboard with various visualizations is produced	2	High	All 4

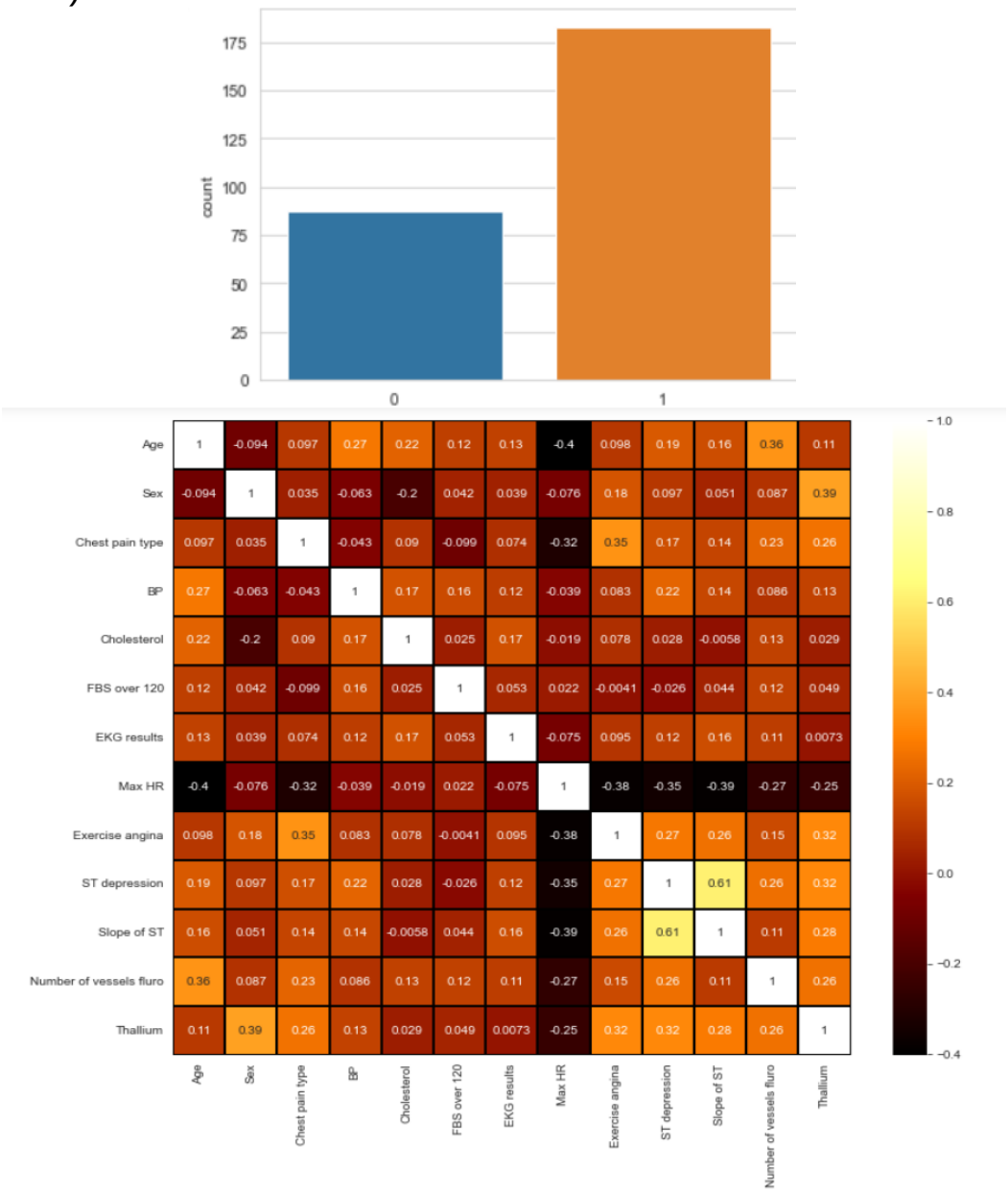
6.2 SPRINT DELIVERY SCHEDULE

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	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

7. CODING & SOLUTIONING

7.1 FEATURE 1

```
colm = ['Sex', 'Chest pain type','FBS over 120','EKG results','Exercise angina','Slope of ST', 'Number of vessels fluro', 'Thallium', 'Heart Disease']
for col in colm:
  sns.countplot(data[col])
  plt.show()
plt.figure(figsize=(12,10))
corr = data.corr()
sns.heatmap(corr, annot = True, linewidths= 0.2, linecolor= 'black', cmap = 'afmhot')
```



7.2 FEATURE 2

```
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
pred = rf.predict(X_test)
print("Accuracy Score for Testing =
{}".format(round(accuracy_score(y_test,pred),5)))
cm = confusion_matrix(y_test,pred)
print(classification_report(y_test,pred))
```

```
Accuracy Score for Testing = 0.81481

              precision    recall  f1-score   support

           0       0.82      0.82      0.82        56
           1       0.81      0.81      0.81        52

    accuracy                           0.81       108
   macro avg       0.81      0.81      0.81       108
weighted avg       0.81      0.81      0.81       108
```

8. TESTING

8.1 TEST CASES

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	13	0	0	13
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

8.2 USER ACCEPTANCE TESTING

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

9. RESULTS

9.1 PERFORMANCE TESTING

S.No.	Parameter	Screenshot / Values
1.	Dashboard design	No of Visulizations / Graphs - 10
2.	Data Responsiveness	High
3.	Amount Data to Rendered (DB2 Metrics)	270 entries
4.	Utilization of Data Filters	Null values are removed
5.	Effective User Story	No of Scene Added - 1
6.	Descriptive Reports	No of Visulizations / Graphs - 1

10. ADVANTAGES & DISADVANTAGES

10.1 ADVANTAGES

- This model considers various factors that could cause heart disease like Age, Sex, BP, Cholesterol, Maximum Heart Rate, etc.
- Libraries like Seaborn and Pandas are used to visualize the data in the form of tables and other creative visualizations like a bar graph and a heatmap.
- An overall accuracy of 81% is obtained.
- A descriptive report containing various visualizations, and the correlation between various parameters is created in the form of a HTML webpage.

10.2 DISADVANTAGES

- The accuracy could be further improved.
- Other factors that could cause heart diseases like stress, family history, etc. are not considered.
- Prediction of any other disease that accompanies heart disease is not predicted.

11. CONCLUSION

Thus, this project helps in visualizing various factors that influence heart diseases and in predicting the risk of heart disease given the history of a patient. This can be used in hospitals and also can be directly used by the patient as the interface is easy to use and easy to visualize. Random Forest Classifier is a good machine learning model that helps in getting a high accuracy even for such a large dataset. The accuracy obtained is around 81% which helps in getting a highly accurate result.

12. FUTURE SCOPE

Using the back- end developed in this project, a front- end application can be built using Flask, HTML, JavaScript and CSS. Further, within the back- end, the size and scope of dataset can be increased by adding more parameters and more medical records. Even the accuracy could be improved by using a hybrid of 2 or more machine learning algorithms.

13. APPENDIX

13.1 SOURCE CODE

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from pandas_profiling import ProfileReport
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
sns.set_style('whitegrid')
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
data = pd.read_csv("H:/Heart/Heart_Disease_Prediction.csv")
data.head()
data.info()
data.describe(include = 'all')
data.isnull().sum()
data.nunique()
data.columns
colm = ['Sex', 'Chest pain type','FBS over 120','EKG results','Exercise
angina','Slope of ST', 'Number of vessels fluro', 'Thallium', 'Heart
Disease']
for col in colm:
  sns.countplot(data[col])
  plt.show()
```

```
plt.figure(figsize=(12,10))
corr = data.corr()
sns.heatmap(corr, annot = True, linewidths= 0.2, linecolor= 'black', cmap
= 'afmhot')
data.columns
X = data[['Age', 'Sex', 'Chest pain type', 'BP', 'Cholesterol', 'FBS over
120',
       'EKG results', 'Max HR', 'Exercise angina', 'ST depression',
       'Slope of ST', 'Number of vessels fluro', 'Thallium']]
y = data['Heart Disease']
print(X.shape,y.shape)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
random_state=42529)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
train_convert = {"Absence":0,"Presence":1}
y_train = y_train.replace(train_convert)
test_convert = {"Absence":0,"Presence":1}
y_test = y_test.replace(test_convert)
mms = MinMaxScaler()
X_train = mms.fit_transform(X_train)
X_test = mms.fit_transform(X_test)
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
rf.get_params()
pred = rf.predict(X_test)
print("Accuracy Score for Testing =
{}".format(round(accuracy_score(y_test,pred),5)))
cm = confusion_matrix(y_test,pred)
print(classification_report(y_test,pred))
sns.heatmap(cm, annot = True, fmt = 'g', cbar = False, cmap = 'icefire',
linewidths= 0.5, linecolor= 'grey')
plt.title('Confusion Matrix')
plt.ylabel('Actal Values')
plt.xlabel('Predicted Values')
y_train_pred = rf.predict(X_train)
print("Accuracy Score for Training =
{}".format(round(accuracy_score(y_train, y_train_pred))))
report = ProfileReport(data)
report.to_notebook_iframe()
report.to_file('H:/Heart_Report')
```

```
from sklearn.model_selection import ShuffleSplit,cross_val_score
from sklearn.datasets import load_iris
shuffle_split=ShuffleSplit(test_size=0.3,train_size=0.5,n_splits=10)
iris = load_iris()
scores=cross_val_score(rf,iris.data,iris.target,cv=shuffle_split)
print("cross Validation scores:n {}".format(scores))
print("Average Cross Validation score :{}".format(scores.mean()))
```

13.2 GITHUB & PROJECT DEMO LINK

GitHub: https://github.com/IBM-EPBL/IBM-Project-37740-1660320250
Project Demo Link: https://drive.google.com/file/d/1udd-R4SDPg_4Vh2xd24pI1JSwzAMdWRr/view?usp=drivesdk