

PROJECT REPORT

Date	11 November 2022
Team ID	PNT2022TMID16827
Project Name	Project – Visualizing and Predicting Heart Diseases with an Interactive Dashboard

TEAM MEMBERS:

JYOTSANA K B

KAVITHAZRI G

KEERTHIKA A

GAYATHRI R

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

INTRODUCTION

1.1 Project Overview

The so-called big data produced by the healthcare sector contains vast amounts of information that can be used to make decisions. In order to create decisions that are more accurate than intuition, a vast amount of data is used. Exploratory Data Analysis (EDA) identifies errors, locates pertinent data, verifies presumptions, and establishes the correlation between explanatory elements. In this context, EDA is understood to be data analysis without statistical modelling or inferences. Any profession needs analytics since it can predict the future and reveal hidden patterns. In the recent past, data analytics has been regarded as a cost-effective technology and it now plays a crucial role in healthcare, including new study discoveries, emergency circumstances, and disease outbreaks. Analytics' application in healthcare advances. A crucial stage in data analysis is to promote preventive care and EDA. The risk variables that lead to heart disease are taken into account and forecasted in this study utilising the Random Forest method, and the analysis is conducted using publically available heart disease data. The dataset contains 209 records with eight parameters, including age, the type of chest pain, blood pressure, blood sugar level, resting ECG, heart rate, and four different types of chest pain. K-means clustering method, together with data analytics and visualisation tools, are utilised to forecast cardiac disease. The pre-processing techniques, classifier performances, and assessment criteria are covered in the study. The visual data in the outcome section demonstrates that the forecast was correct.

1.2 Purpose

Due to their lifestyle choices and the state of the environment today, individuals are susceptible to many diseases. To prevent such

diseases from becoming severe, early detection and prediction of these disorders are crucial. Most of the time, it is challenging for doctors to appropriately diagnose ailments by hand. This study aims to identify and forecast patients who have more prevalent chronic illnesses. This might be achieved by making sure that this category accurately identifies people with chronic conditions by employing a cutting-edge machine learning technique. Another difficult task is illness forecasting. Data mining is therefore essential for disease prediction. Based on the patient's symptoms, the proposed system provides a comprehensive disease prognosis. A crucial stage in data analysis is to promote preventive care and EDA. The risk variables that lead to heart disease are taken into account and forecasted in this study utilising the Random Forest method, and the analysis is conducted using publically available heart disease data. The dataset contains 209 records with eight parameters, including age, the type of chest pain, blood pressure, blood sugar level, resting ECG, heart rate, and four different types of chest pain. K-means clustering method, together with data analytics and visualisation tools, are utilised to forecast cardiac disease. The pre-processing techniques, classifier performances, and assessment criteria are covered in the study. The visual data in the outcome section demonstrates that the forecast was correct.

2. LITERATURE SURVEY

2.1 Existing problem

As per the recent study by WHO, heart related diseases are increasing. 17.9 million people die every-year due to this. With growing population, it gets further difficult to diagnose and start treatment at early stage. But due to the recent advancement in technology, Machine Learning techniques have accelerated the health sector by multiple researches. Thus, the objective of this paper is to build a ML model for heart disease prediction based on

the related parameters. We have used a benchmark dataset of UCI Heart disease prediction for this research work, which consist of 14 different parameters related to Heart Disease. Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree have been used for the development of model. In our research we have also tried to find the correlations between the different attributes available in the dataset with the help of standard Machine Learning methods and then using them efficiently in the prediction of chances of Heart disease. Result shows that compared to other ML techniques, Random Forest gives more accuracy in less time for the prediction. This model can be helpful to the medical practitioners at their clinic as decision support system.

S.NO	TITLE	AUTHOR & PUBLISHED YEAR	KEYWORDS	PROPOSED WORK
1.	Prediction of Cardiovascular Disease Using Machine Learning Algorithms	<u>Kumar G Dinesh; K Arumugam; Kumar D Santhosh; V Mareeswari</u> <u>2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)</u>	Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset.	This project proposes a prediction model to predict whether a people have a heart disease or not and to provide an awareness or diagnosis on that.
2.	Data mining and visualization for prediction of multiple diseases in healthcare	<u>Alinkya Kunjir; Harshal Sawant; Nuzhat F. Shaikh</u> <u>2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)</u>	Data mining methods namely, Naive Bayes and J48 algorithms are compared for testing their accuracy and performance on the training medical datasets.	The main aim of this project is to build a basic decision support system which can determine and extract previously unseen patterns, relations and concepts related with multiple disease from a historical database records of specified multiple diseases. The proposed system can solve difficult queries for detecting a particular disease and also can assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. The decisions taken by medical practitioners with the help of technology can result in effective and low cost treatments. There is an insufficiency of technology and analysis system and methods to discover connections, concepts and patterns in the medical data. Data mining is an engineering study of extracting previously undiscovered patterns from a selected set of data.
3.	A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease	<u>Sayedamin Pouriyeh; Sara Vahid; Giovanna Sannino; Giuseppe De Pietro; Hamid Arabnia; Juan Gutierrez</u> <u>2017 IEEE Symposium on Computers and Communications (ISCC)</u>	Different classifiers, namely Decision Tree (DT), Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM)	This paper aims to investigate and compare the accuracy of different data mining classification schemes, employing Ensemble Machine Learning Techniques, for the prediction of heart disease. The Cleveland data set for heart diseases, containing 303 instances, has been used as the main database for the training and testing of the developed system. 10-Fold Cross-Validation has been applied in order to increase the amount of data, which would otherwise have been limited.

4.	Prediction of cardiovascular disease	<u>Faisal Perva; Harun Tucaković; Muhammed Mušanović; Emine Yaman</u> <u>2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT)</u>	Using decision trees (C4.5), k-NN, and Naïve Bayes, in combination with cross-validation and holdout methods.	Nowadays, cardiovascular diseases are one of the leading causes of death. Earlier and better detection of such diseases would lead to earlier treatment and eventually to better chances of patients being able to overcome those diseases. Machine learning algorithms have been proven useful in detecting several medical conditions based on patients' characteristics. In this paper, we are trying to predict whether a patient has a cardiovascular disease based on their characteristics.
5.	Early Prediction of Cardiovascular Diseases Using Feature Selection and Machine Learning Techniques	<u>Tamanna Yesmin Rashme; Linta Islam; Sohely Jahan; Ayesha Aziz Prova</u> <u>2021 6th International Conference on Communication and Electronics Systems (ICCES)</u>	Random Forest algorithm is used to select suitable attributes for the prediction process. The proposed model is assessed based on evaluation metrics; accuracy, precision, recall (sensitivity), f1-score, and specificity. In this exploration of predicting cardiovascular disease, the XGBoost machine learning classifier accomplished a higher rate of accuracy 75.10%.	cardiovascular disease is one of the most important diseases that affects the heart and blood vessels. The loss of lives is mostly linked to a lack of early disease detection, and a preemptive prediction of cardiovascular disease risk will greatly alleviate the situation. Due to the increasing amount of data growth in the health care industry, therefore Machine Learning techniques predict the disease depends on the severity of the patient's side effect. This research work proposes a model to perform early prediction of cardiovascular disease by using different machine learning algorithms, which are used for different prediction purposes.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

An empathy map is a widely-used visualization tool within the field of UX and HCI practice. In relation to empathetic design, the primary purpose of an empathy map is to bridge the understanding of the end user. Within context of its application, this tool is used to build a shared understanding of the user's needs and provide context to a user-centered solution

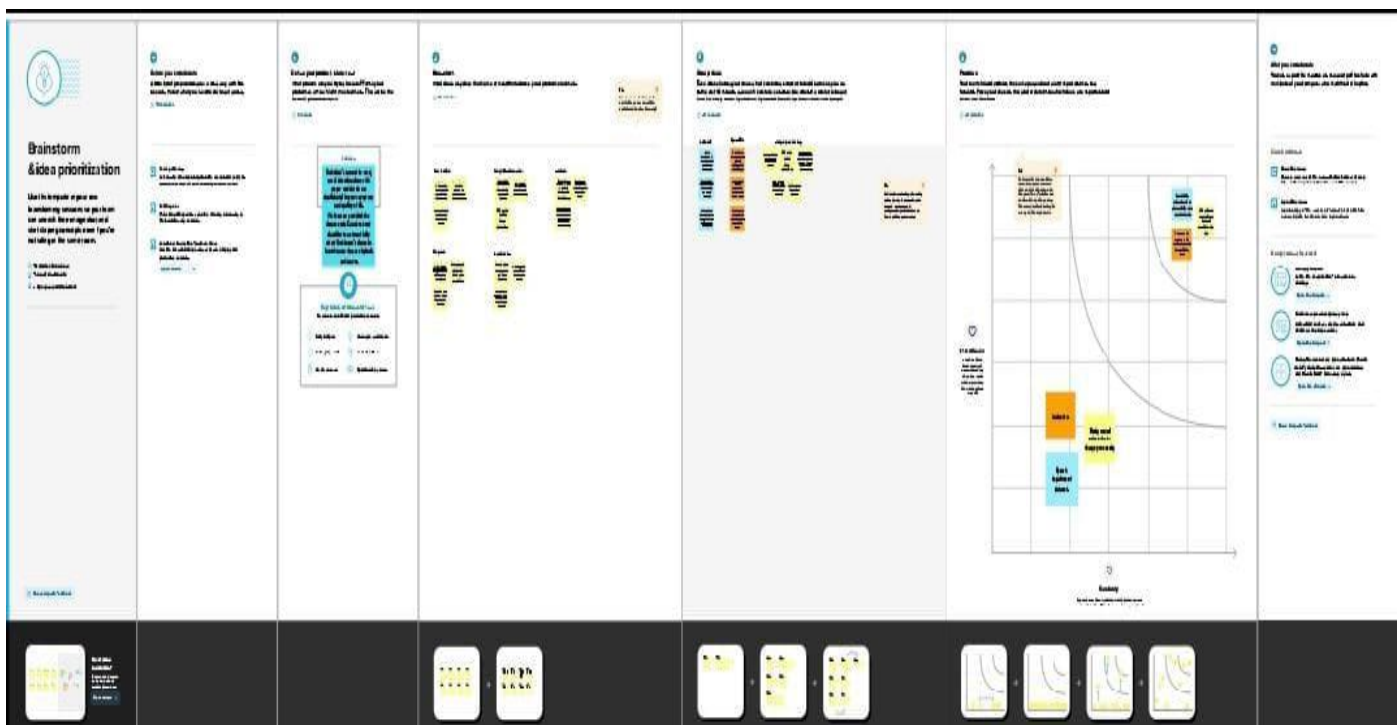
An empathy map in Visualizing and Predicting Heart Diseases with an Interactive Dash Board is used to describe and understand about the feeling of the customer or patient when they got attacked by the heart disease

Empathy Map Canvas

[illegible]

3.2 Ideation & Brainstorming

Ideation is often closely related to the practice of brainstorming, a specific technique that is utilized to generate new ideas. A principal difference between ideation and brainstorming is that ideation is commonly more thought of as being an individual pursuit, while brainstorming is almost always a group activity. Brainstorming is usually conducted by getting a group of people together to come up with either general new ideas or ideas for solving a specific problem or dealing with a specific situation



3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	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.
2.	Idea / Solution description	Predicting using rain forest algorithm.
3.	Novelty / Uniqueness	Higher accuracy and for complex problem
4.	Social Impact / Customer Satisfaction	The efforts have been aimed toward the identification, modification and treatment of individual-level risk factors
5.	Business Model (Revenue Model)	We propose a model that allows users to get instant guidance on their heart disease
6.	Scalability of the Solution	Our prediction solution uses random forest 98% accuracy is achieved.

3.4 Problem Solution fit

Define CS, fit into CC	<div>1. CUSTOMER SEGMENT(S)<div>CS</div></div> <div>Who is your customer? i.e. working parents of 0-5 y.o. kids</div> <div>People who suffer from heart disease.</div>	<div>6. CUSTOMER CONSTRAINTS<div>CC</div></div> <div>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> <div>Insufficient money for medical checkups. Unaware about regular checkup.</div>	<div>5. AVAILABLE SOLUTIONS<div>AS</div></div> <div>Which solutions are available to the customers when they face the problem 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</div> <div>Customers can go to the doctor for a medical checkup Based on the test results doctors will advise them. The patient can do manual prediction.</div>	Explore AS, differentiate
	<div>2. JOBS-TO-BE-DONE / PROBLEMS<div>J&P</div></div> <div>Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.</div> <div>Visualizations give doctors very good insights on the potential chances for a patient to get heart disease. Visualizing and predicting heart disease.</div>	<div>9. PROBLEM ROOT CAUSE<div>RC</div></div> <div>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> <div>The main reason of getting chdr diabetics, high cholesterol and blood pressure, smoking, mental depression, eating an unhealthy diet and any family history of heart disease.</div>	<div>7. BEHAVIOUR<div>BE</div></div> <div>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> <div>First of all they(Customer or patients) should report what problem they are undergoing according to their health condition. After that they are instructed to follow the steps that the solution provider given(that is jobs to be done for curing their illness).</div>	
Focus on J&P, tap into BE, understand RC	<div>3. TRIGGERS<div>TR</div></div> <div>What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.</div> <div>By seeing the advanced technology providing a solution for their problem with low cost and getting benefit from where they are, so this makes customers to act.</div>	<div>10. YOUR SOLUTION<div>SL</div></div> <div>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 behaviour.</div> <div>To clean data and provide visualizations to help doctors in their diagnosis of patient as well as make customers more aware of this issue. Develop an application to predict heart disease with machine learning.</div>	<div>8. CHANNELS of BEHAVIOUR<div>CH</div></div> <div>8.1 ONLINE What kind of actions do customers take online? Extract online channels from #7</div> <div>8.2 OFFLINE What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.</div> <div>ONLINE Searching about heart disease symptoms in internet.</div> <div>OFFLINE Asking other people if they feels the same?</div>	Identify strong TR & EM
	<div>4. EMOTIONS: BEFORE / AFTER<div>EM</div></div> <div>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> <div>When they are facing problem of health illness, they feel lonenly depressed of them and their family, feel insecure etc. After knowing their illness can be treated they have hope confidence to tackle their problem.</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 Form Registration through GMail Registration through LinkedIn
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Personal details for account	They have to enter their personal details.
FR-4	Regular medical condition	Enter the medical details and reports required.
FR-5	Doctor consultation	Get the doctor's consultation.

4.2 Non-Functional requirement

Following are the non-functional requirements of the proposed solution

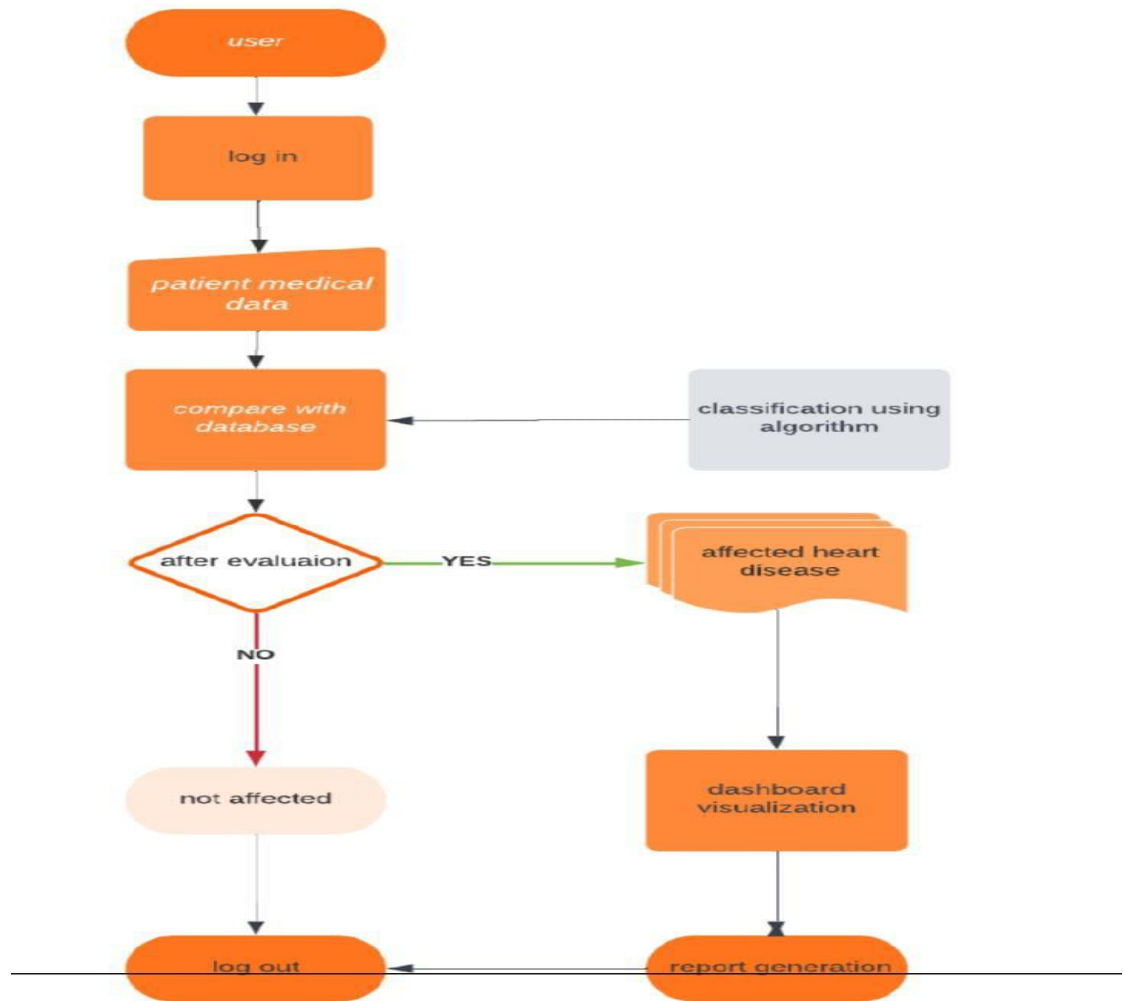
FR No	Non-Functional Requirement	Description
NFR-1	Usability	As usability is a prerequisite for success of health and wellness mobile apps, our proposed solution aims to provide insights and suggestions for improving usability and experience of the mobile health app.
NFR-2	Security	A proposed solution can empower patient, streamline communication, and provide real time monitoring and self ,management of medical condition by building a secure app that puts

		security, privacy and authentication.
NFR-3	Reliability	Measuring reliability can improve the quality and value of health care apps. Our proposed solution will provide accurate prediction of disease with a lower risk of errors that cause harm to user and reduces the death rate. Our solution provides safety to user's data with lot of benefits simply in home which is efficient without wasting equipments, supplies, ideas and energy
NFR-4	Performance	The performance of this project is to reduce heart disease death rate by earlier accurate disease prediction. Our solution offers services such as disease prevention, diagnosis and treatment and rehabilitation.
NFR-5	Availability	Availability is important because while there are often shortages in human resources , deployed providers are frequently inappropriately absent or , when present, are not actively delivering health care because they are engaged in other duties.
NFR-6	Scalability	It can be integrated with smart watch and apps for further advancement which is very helpful for earlier prediction.

5. PROJECT DESIGN

5.1 Data Flow Diagrams

The flow of data of a system or a process is represented by DFD. It also gives insight into the inputs and outputs of each entity and the process itself. DFD does not have control flow and no loops or decision rules are present. Specific operations depending on the type of data can be explained by a flowchart. Data Flow Diagram can be represented in several ways. The DFD belongs to structured-analysis modeling tools. Data Flow diagrams are very popular because they help us to visualize the major steps and data involved in software system processes.



5.2 Solution & Technical Architecture

Technical architecture—which is also often referred to as application architecture, IT architecture, business architecture, etc.—refers to creating a structured software solution that will meet the business needs and expectations while providing a strong technical plan for the growth of the software application through its lifetime. IT architecture is equally important to the business team and the information technology team.

Technical architecture includes the major components of the system, their relationships, and the contracts that define the interactions between the components. The goal of technical architects is to achieve all the business needs with an application that is optimized for both performance and security.

5.3 User Stories

A user story is defined incrementally, in three stages:

- *The brief description of the need
- *The conversations that happen during backlog grooming and iteration planning to solidify the details
- *The tests that confirm the story's satisfactory completion.

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
		USN-3	As a user, I can register for the application through Gmail	I can receive confirmation email	Medium	Sprint-1
Customer (Web user)	Login	USN-4	As a user, I can log into the application by entering email & password	I can access my account using my details	High	Sprint-1
	Dashboard	USN-5	User can view his/her complete medical analysis and accuracy of disease prediction	I can view my medical analysis and accuracy	High	Sprint-2
	Dashboard	USN-6	User can view the accuracy of occurrence of heart disease through report generation	I can view the accuracy of heart disease in the dashboard	high	Sprint-2
Customer Care Executive	Helpdesk	USN-7	As a customer care executive, he/she can view the customer queries.	I can post my queries in the dashboard	Medium	Sprint-3
		USN-8	As a customer care executive, he/she can answer the customer queries	I can get support from helpdesk	High	Sprint-3
Administrator	User profile	USN-9	As an admin, he/she can update the health details of users.	I can view my updated health details	High	Sprint-4
		USN-10	As an admin, he/she can add or delete users.	I can access my account / Dashboard when logged in	High	Sprint-4
		USN-11	As an admin, he/she can manage the user details.	I can view the organized data of myself.	High	Sprint-4

6. PROJECT PLANNING & SCHEDULING

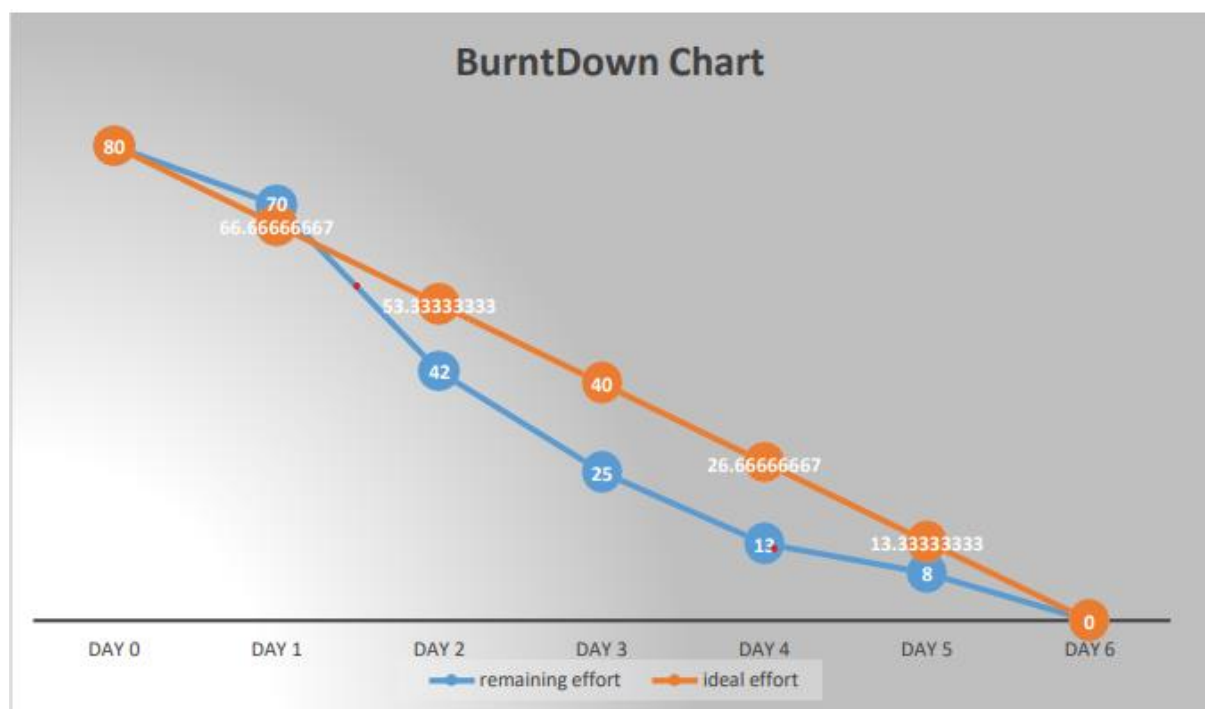
6.1 Sprint Planning & Estimation

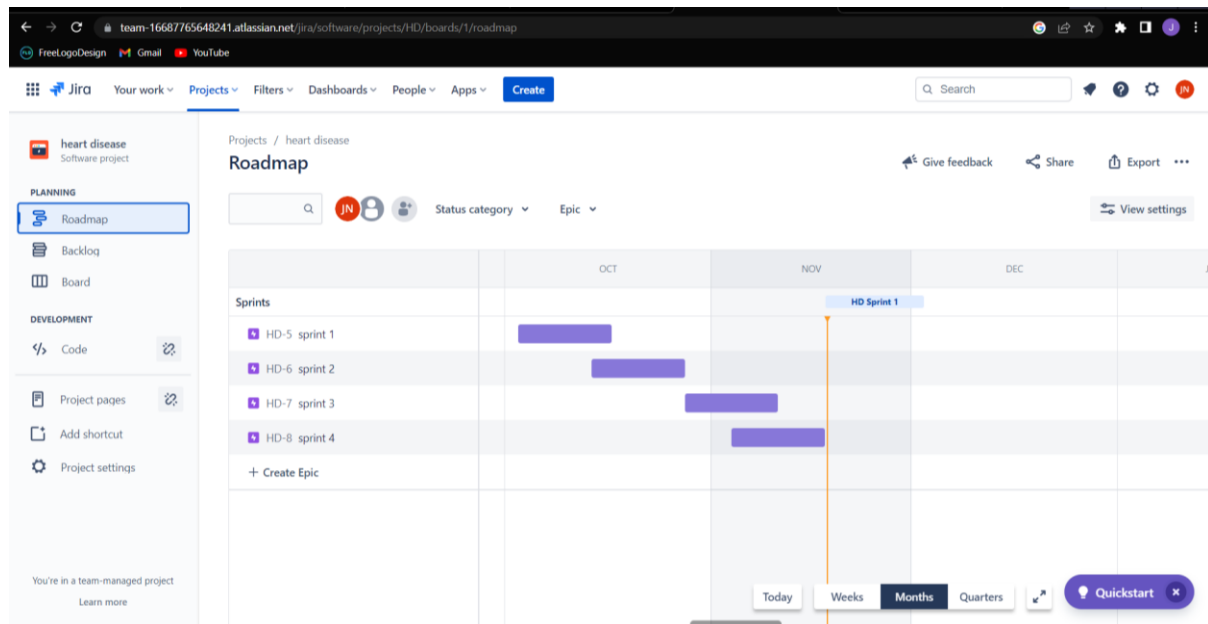
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story points	Priority	Team Members
Sprint1	Preparing Data for Analysis and Exploratory Data Analysis (EDA)	USN-1	To check for any null values or outliers, data cleansing is performed	10	Medium	Jyotsana KB Kavithazri G Keerthika A Gayathri R
		USN-2	Training and Testing Google Colab is used to implement the data model.	10	High	Jyotsana KB Kavithazri G Keerthika A Gayathri R
Sprint2	Working with dataset	USN-3	Working with the Dataset. Understand Dataset Load the Dataset Explore the Data Visualize the Data.	20	Medium	Jyotsana KB Kavithazri G Keerthika A Gayathri R
Sprint3	Data Visualization	USN-4	We plan to create a range of graphs and charts to display the insights and visualisations using the predetermined criteria.	20	High	Jyotsana KB Kavithazri G Keerthika A Gayathri R
Sprint4	Dashboard	USN-5	Dashboard Displaying Various Images	15	High	Jyotsana KB Kavithazri G Keerthika A Gayathri R
		USN-6	User can create reports and stories	5	Medium	Jyotsana KB Kavithazri G Keerthika A Gayathri R

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	7 Days	24 Oct 2022	31 Oct 2022	20	26 Oct 2022
Sprint-2	20	6 Days	01 Nov 2022	06 Nov 2022	20	02 Nov 2022
Sprint-3	20	7 Days	06 Nov 2022	13 Nov 2022	20	09 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	16 Nov 2022

6.3 Reports from JIRA





7. CODING & SOLUTIONING

Pre-processing and EDA

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rcParams
from matplotlib.cm import rainbow
import seaborn as sns
%matplotlib inline

from sklearn.model_selection import
train_test_split
from sklearn.preprocessing import
StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn import tree
from warnings import filterwarnings
filterwarnings("ignore")
```

```

#model validation
from sklearn.metrics import
log_loss,roc_auc_score,precision_score,f1_score
,recall_score,roc_curve,auc,plot_roc_curve
from sklearn.metrics import
classification_report,
confusion_matrix,accuracy_score,fbeta_score,mat
thews_corrcoef
from sklearn import metrics
from mlxtend.plotting import
plot_confusion_matrix

#extra
from sklearn.pipeline import make_pipeline,
make_union
from sklearn.preprocessing import
PolynomialFeatures
from sklearn.feature_selection import
SelectFwe, f_regression

from sklearn.ensemble import
RandomForestClassifier

```

Import dataset

```

import pandas as pd
df=pd.read_csv("https://raw.githubusercontent.com/IBM-EPBL/IBM-Project-6870-1658841462/main/Project%20Development%20Phase/Sprint%201/understandeing%20dataset/Heart_Disease_Prediction.csv")

```

```
df.head()
```

Out[2]:

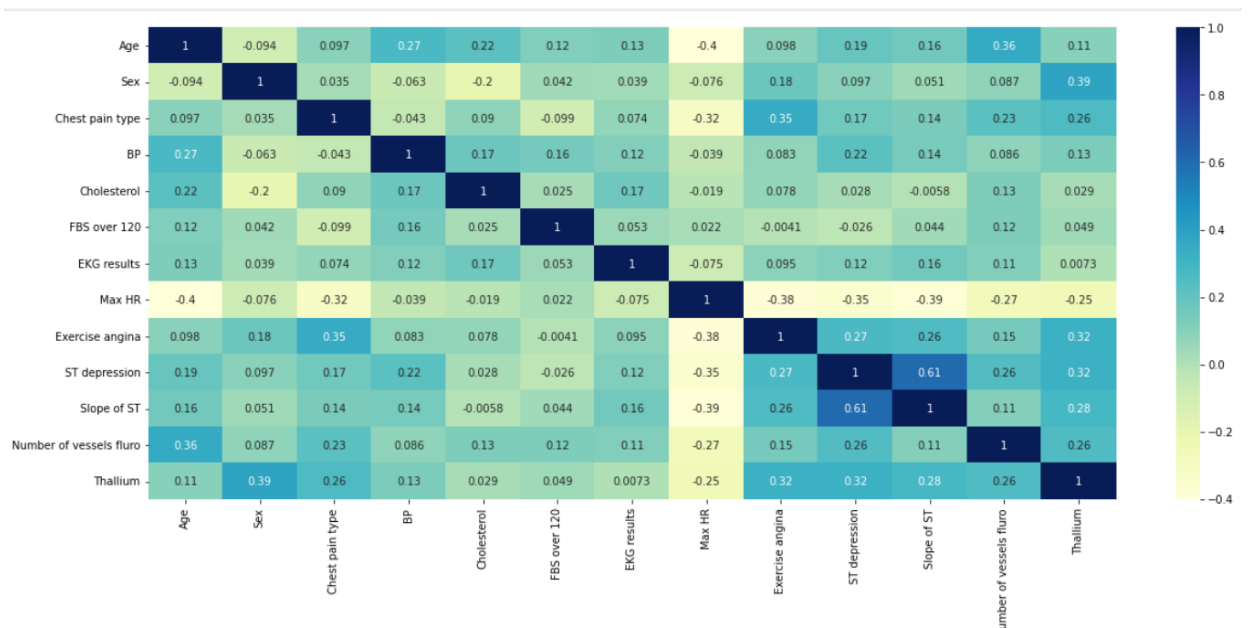
	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluoro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence

```
df.columns
```

```
df.isnull().sum()
```

```
df.apply(lambda x:len(x.unique()))
```

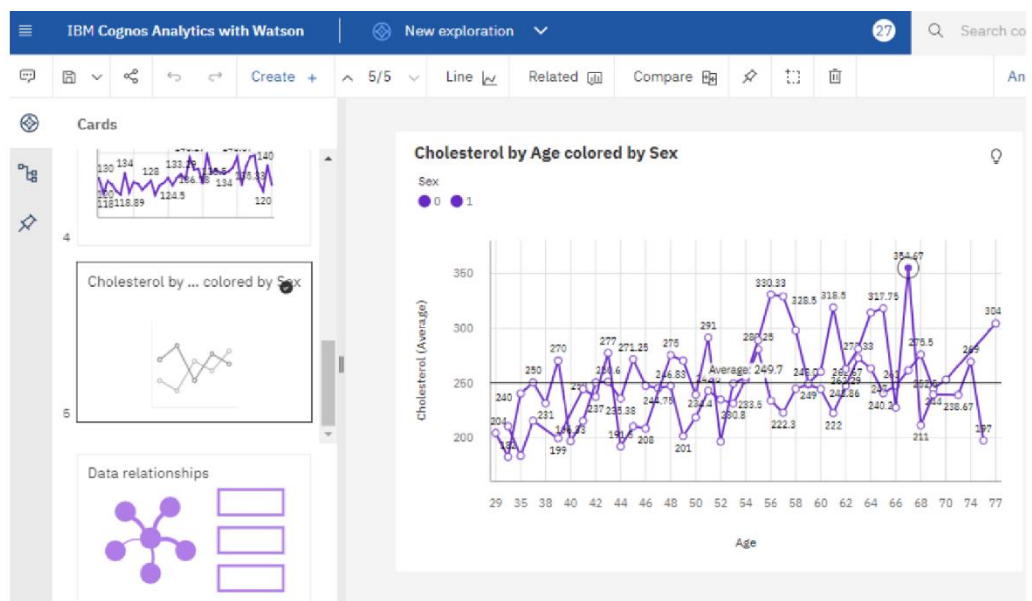
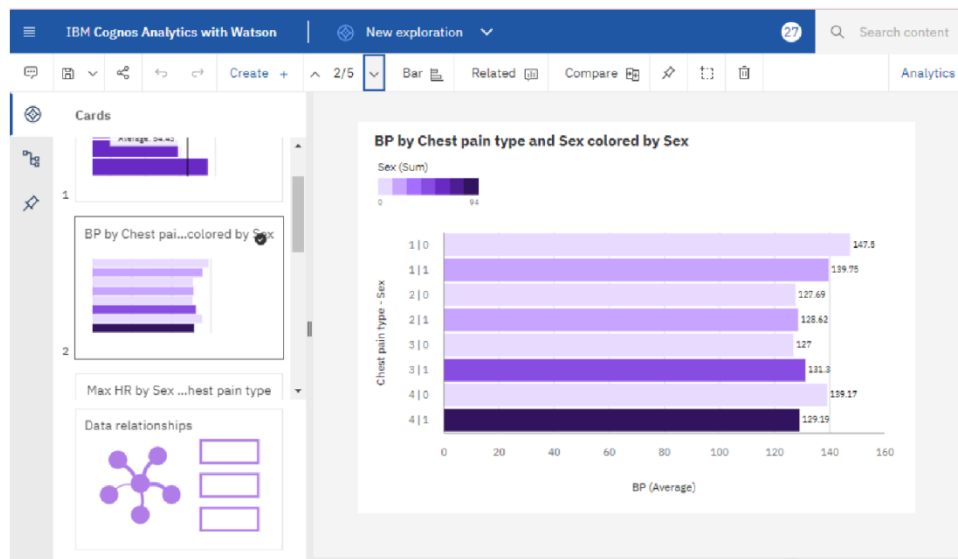
```
plt.figure(figsize=(20,8))
sns.heatmap(df.corr(), cmap="YlGnBu",
annot=True)
plt.show()
```

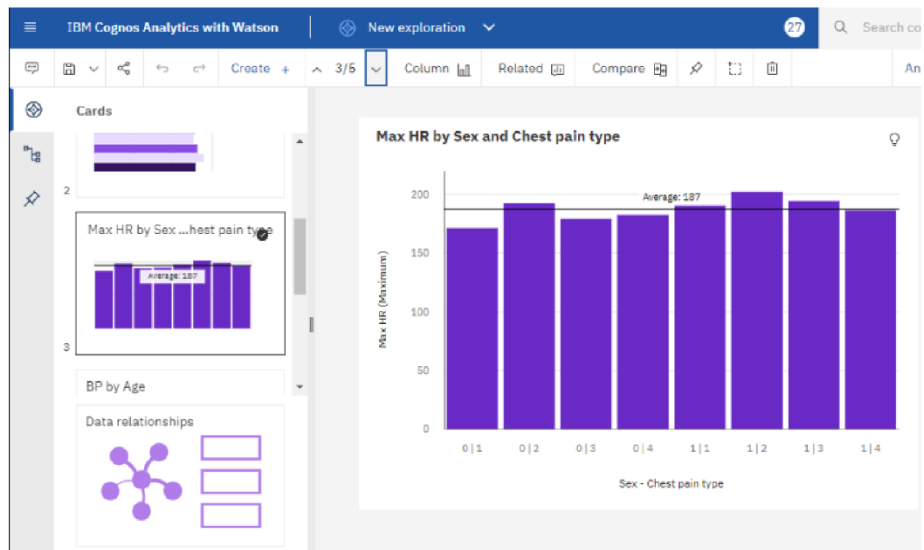
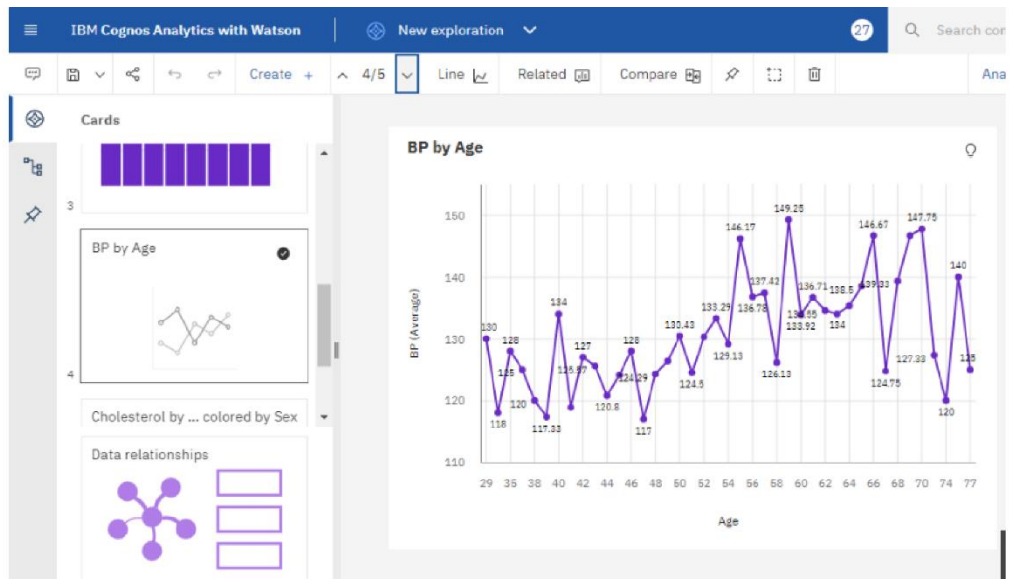


8. TESTING

DATASET:

<https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction>





9. RESULTS

9.1 PERFORMANCE METRIC

1. Hours worked: 50 hours
2. Stick to Timelines: 100%
3. Stay within budget: 100%
4. Consistency of the product: 85%
5. Efficiency of the product: 85%
6. Quality of the product: 85%

10. ADVANTAGES & DISADVANTAGES

Advantages

- *User can search for doctor's help at any point of time.
- *User can talk about their Heart Disease and get instant diagnosis.
- *Doctors get more clients online.
- *Very useful in case of emergency.

Disadvantages

- *Accuracy Issues: A computerized system alone does not ensure accuracy, and the warehouse data is only as good as the data entry that created it.

*The system is not fully automated, it needs data from user for full diagnosis.

11. CONCLUSION

Our study mainly focused on the use of data mining techniques in healthcare especially in detection of heart disease. Heart disease is a fatal disease which may cause death. Data mining techniques were implemented using the following algorithm, KNN, Neural Networks, Decision Tree, and Naive Bayes and Random Forest. We measured performance on the basis of Accuracy, TN, FP, FN and TP rate and in some algorithm. We conducted five experiments with the same data set to predict heart disease. The result of all the implemented algorithm are shown in tabular form for better understanding and comparison. The experiment shows that Naive Bayes gives the highest accuracy which is 88% followed by ANN and KNN with accuracy of 87% . Our findings indicate that data mining can be used and applied in the healthcare industry to predict and diagnose the disease at early stages.

12. FUTURE SCOPE

The Future Scope of the visualizing and Predicting Heart Diseases with an Interactive Dash Board is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc

13. APPENDIX

Source Code

Pre-processing and EDA

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rcParams
from matplotlib.cm import rainbow
import seaborn as sns
%matplotlib inline

from sklearn.model_selection import
train_test_split
from sklearn.preprocessing import
StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn import tree
from warnings import filterwarnings
filterwarnings("ignore")

#model validation
from sklearn.metrics import
log_loss,roc_auc_score,precision_score,f1_score
,recall_score,roc_curve,auc,plot_roc_curve
from sklearn.metrics import
classification_report,
confusion_matrix,accuracy_score,fbeta_score,mat
thews_corrcoef
from sklearn import metrics
```

```

from mlxtend.plotting import
plot_confusion_matrix

#extra
from sklearn.pipeline import make_pipeline,
make_union
from sklearn.preprocessing import
PolynomialFeatures
from sklearn.feature_selection import
SelectFwe, f_regression

from sklearn.ensemble import
RandomForestClassifier

```

Import dataset

```

import pandas as pd
df=pd.read_csv("https://raw.githubusercontent.com/IBM-EPBL/IBM-Project-6870-1658841462/main/Project%20Development%20Phase/Sprint%201/understandeing%20dataset/Heart_Disease_Prediction.csv")

```

```
df.head()
```

Out[2]:

	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	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence

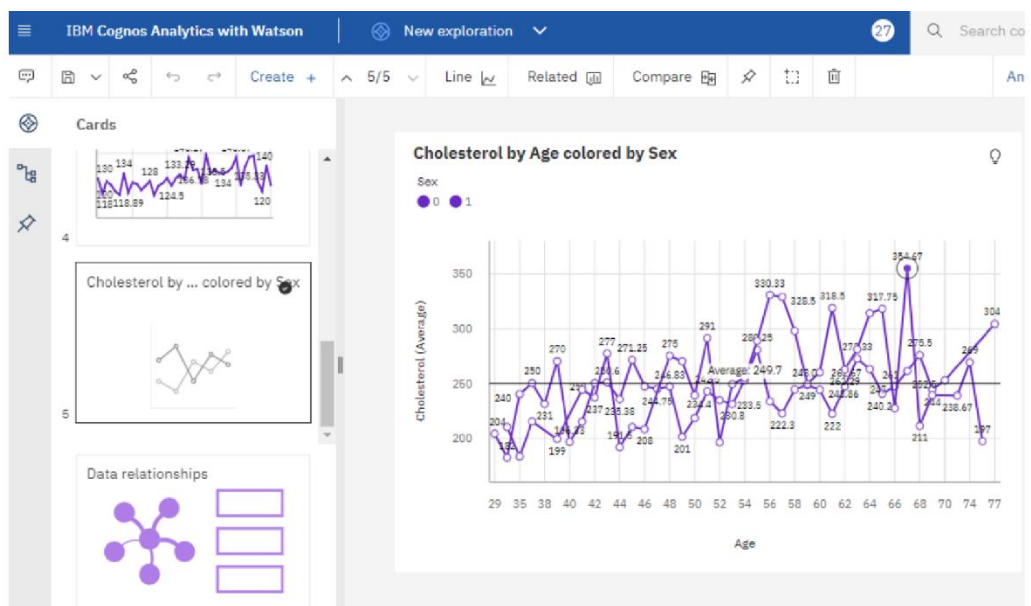
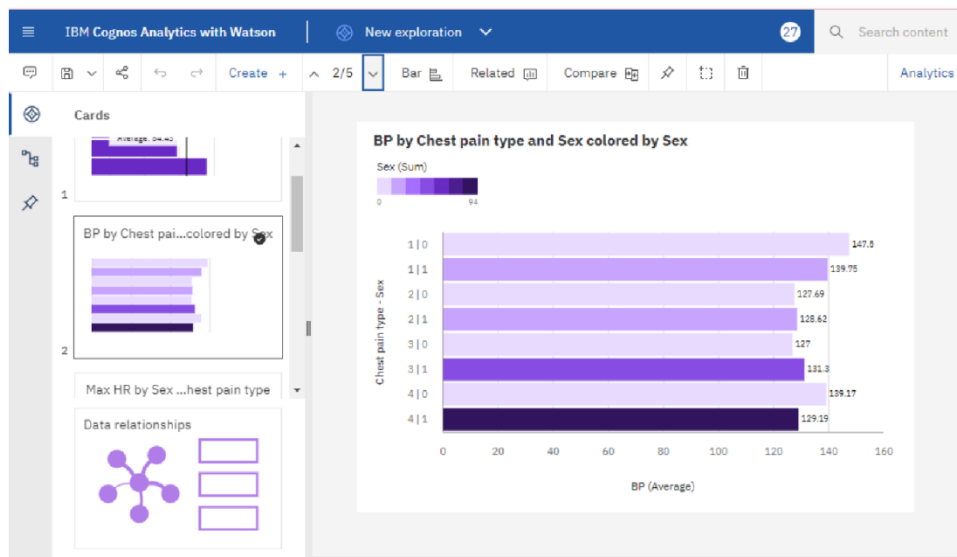
```
df.columns
```

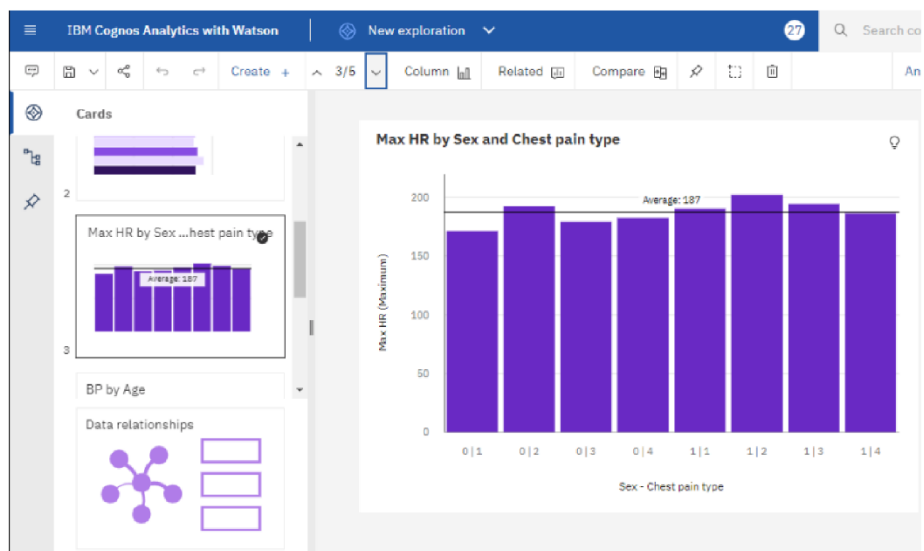
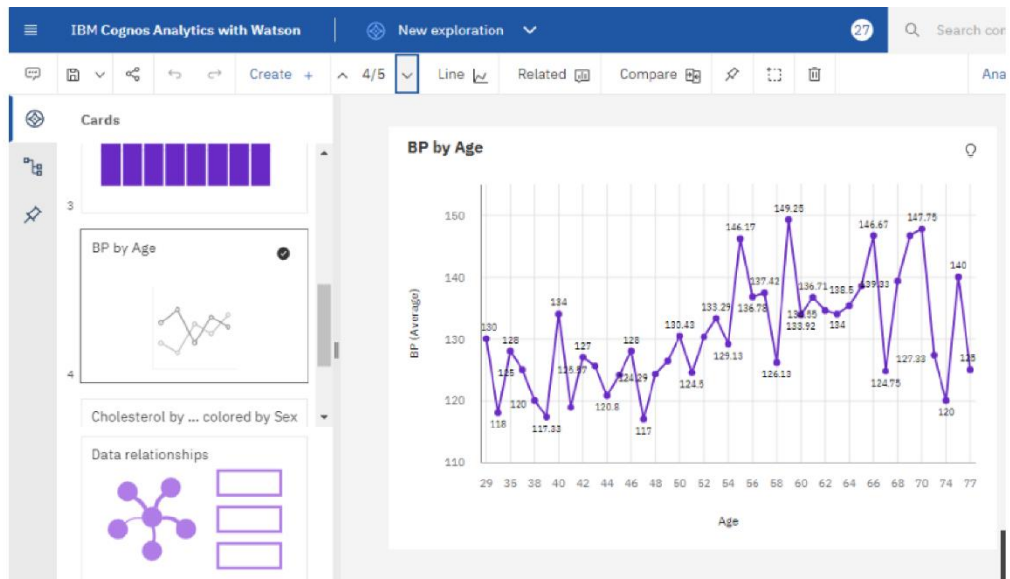
```
df.isnull().sum()
```

```
df.apply(lambda x:len(x.unique()))
```

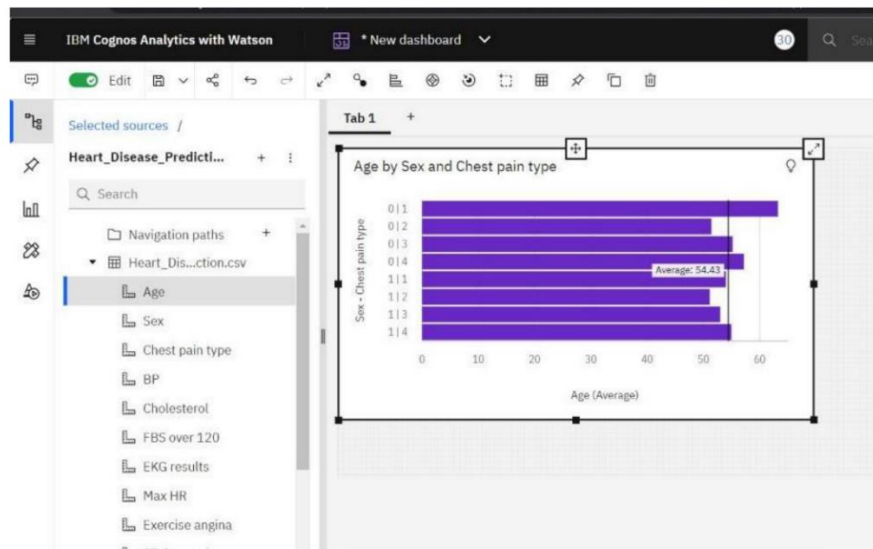
```
plt.figure(figsize=(20,8))  
sns.heatmap(df.corr(), cmap="YlGnBu",  
annot=True)  
plt.show()
```

Data Visualisation





AVERAGE AGE FOR DIFFERENT TYPE OF CHEST PAIN



Average Age For Different Types Of Chest Pain In Existing Heart Disease

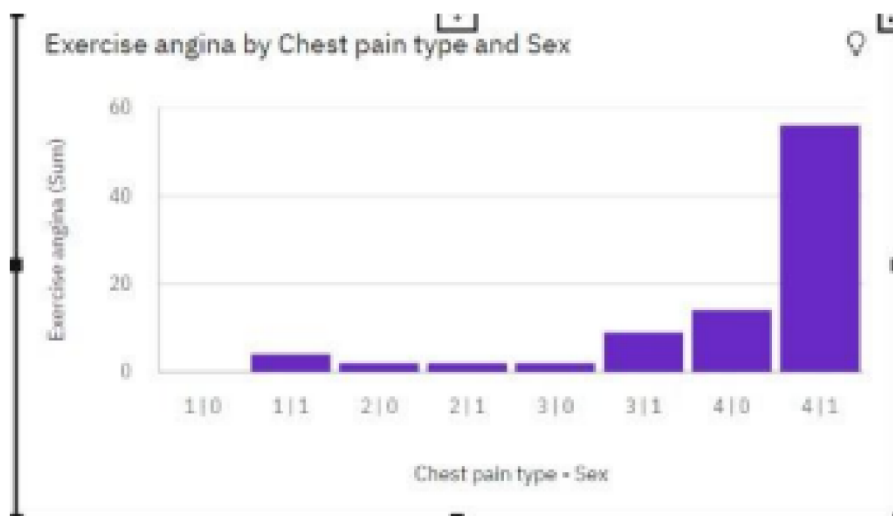
IBM Cognos Analytics with Watson | New dashboard: heart disease | 27 | Search content

Tab 1 Tab 2 Tab 3 Tab 4 **Tab 5** Tab 6 Tab 7 Tab 8

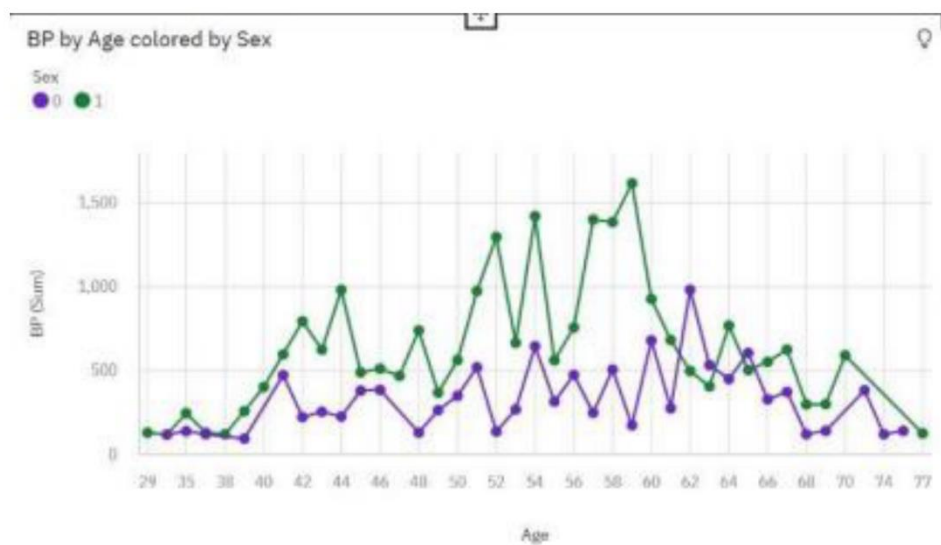
Heart Disease for Chest pain type and Sex

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

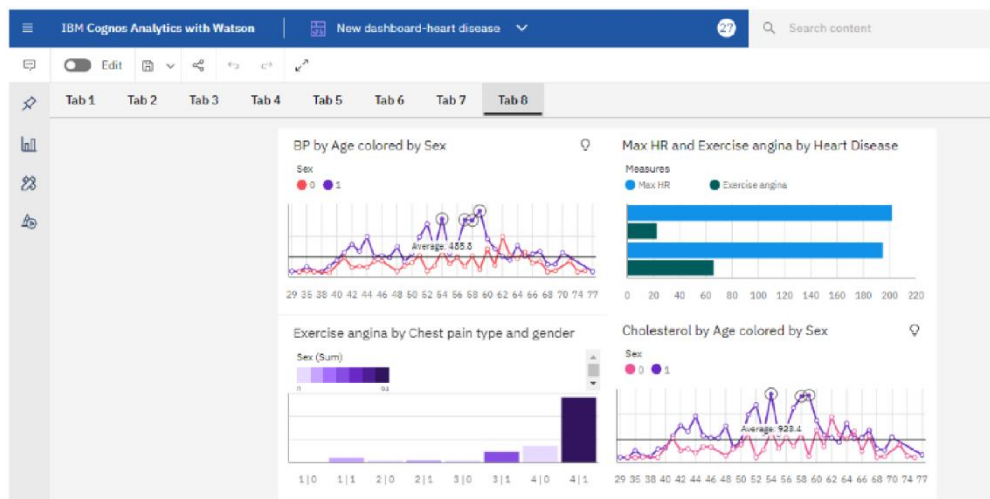
Average Exercise Angina During Chest Pain



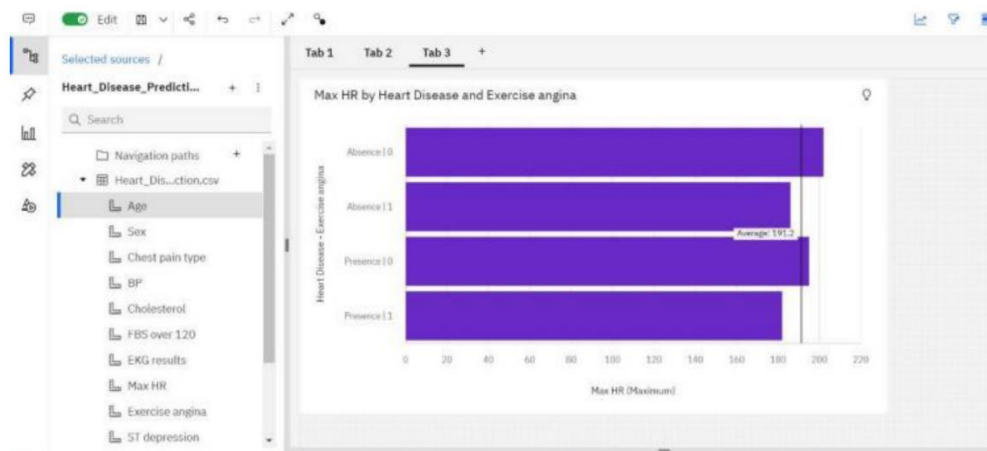
BP Variation with Respect To Age



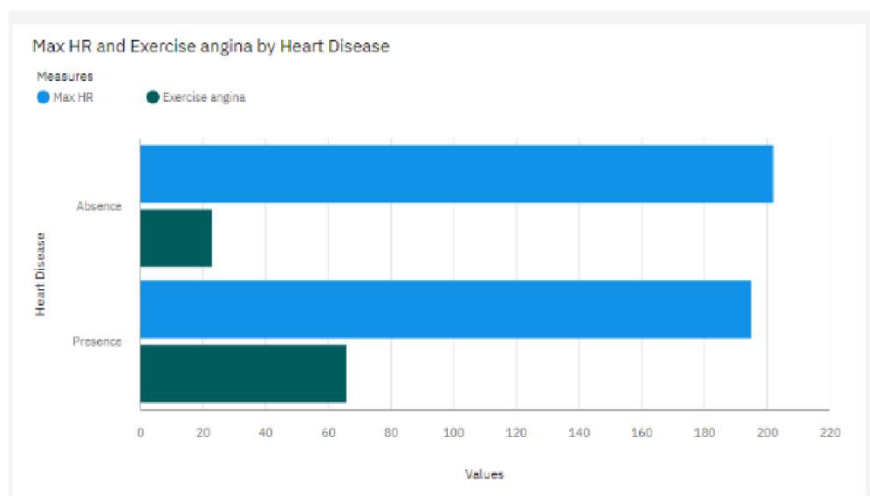
Dashboard Showing Different Types of Visuals



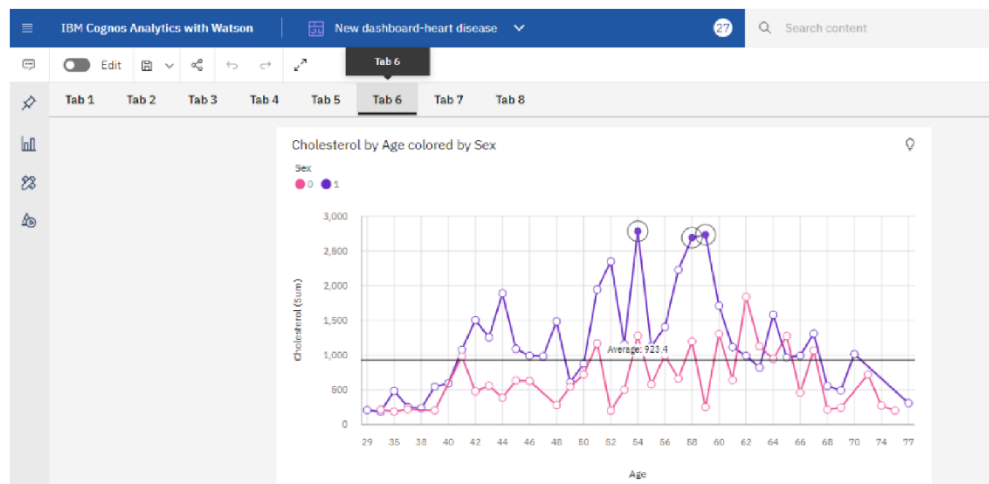
Effect Of Existing Heart Disease on Average Of Exercise Angina



Maximum Heart Rate In Existing Heart Disease By Exercise Angina



Serum Cholesterol Levels Vs Age



GitHub & Project Demo Link:

<https://github.com/IBM-EPBL/IBM-Project-6870-1658841462>