DATA ANALYTICS

VISUALIZING AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASHBOARD

TEAM ID: PNT2022TMID49830

TEAM MEMBERS:

1. R.SARAVANA PRIYA
2. B.ANITHA
3. P.SWETHA
4. S.VANASUNDARI

INDEX

S.NO.	TOPIC	PAGE NO.
1	INTRODUCTION 1.1 Project Overview 1.2 Purpose	**1**
2	LITERATURE SURVEY 2.1 Existing problem 2.2 References 2.3 Problem Statement Definition	**2**
3	IDEATION & PROPOSED SOLUTION 3.1 Empathy Map Canvas 3.2 Ideation & Brainstorming 3.3 Proposed Solution 3.4 Problem Solution fit	**4**
4	REQUIREMENT ANALYSIS 4.1 Functional requirement 4.2 Non-Functional requirements	**11**
5	PROJECT DESIGN 5.1 Data Flow Diagrams 5.2 Solution & Technical Architecture 5.3 User Stories	**13**
6	PROJECT PLANNING & SCHEDULING 6.1 Sprint Planning & Estimation 6.2 Sprint Delivery Plan	**17**
7	RESULTS 7.1 Performance Metrics	**20**
8	ADVANTAGES & DISADVANTAGES	**22**
9	CONCLUSION	**22**
10	FUTURE SCOPE	**23**
11	APPENDIX Source Code GitHub & Project Demo Link	**23**

1. INTRODUCTION

1.1 Project Overview

In many nations, cardiovascular disease is the main cause of death. Cardiovascular illness is frequently identified by doctors based on the results of recent clinical testing and their prior experience treating patients who presented with comparable symptoms. Analytics is an essential technique for any profession because it predicts the future and uncovers hidden patterns. In recent years, data analytics has been regarded as a cost-effective technology, and it plays an important role in healthcare, including new research findings, emergency situations, and disease outbreaks. The use of analytics in healthcare improves care by facilitating preventive care, and EDA is a critical step when analysing data. The risk factors that cause heart disease are considered and predicted using the K-means algorithm, and the analysis is carried out using publicly available heart disease data. The K-means clustering algorithm, in conjunction with data analytics and visualization tools, is used to predict heart disease. Pre-processing methods, classifier performance, and evaluation metrics are all covered. The visualized data in the result section shows that the prediction is correct.

1.2 Purpose

The diagnosis of heart illness is a difficult undertaking, but it can provide an automated prognosis of the patient's heart status to help with subsequent treatment. Typically, the patient's physical examination, signs, and symptoms of heart disease serve as the foundation for the diagnosis. Due to their lifestyle choices and the state of the environment today, individuals are susceptible to many diseases. To prevent the severity of these disorders, early detection and prediction of their occurrence are crucial. Predictive analytics in healthcare can raise the standard of care, gather more clinical data for individualized treatment, and correctly identify each patient's medical condition. For the purpose of making wise decisions, healthcare businesses gather enormous amounts of data that may contain some hidden information. Some sophisticated data mining techniques are utilized to deliver accurate results and make data-driven judgments. In this study, we established a project for estimating the degree of heart disease risk using a neural network. For prediction, the algorithm makes use of 12 medical variables, including age, sex, blood pressure, cholesterol, and obesity.

2. LITERATURE SURVEY

2.1 Existing problem

"Predicting the Risk of Heart Failure With EHR Sequential Data Modelling," suggested a neural network-based approach. This essay utilized real-world datasets derived from electronic health record (EHR) data connected to congestive heart failure to complete the experiment and foretell cardiac disease in advance. We tend to utilize word vectors and one-hot encryption when modelling the cardiac failure events predicted by the diagnosis using a long memory's fundamental tenets against its network theory. Analyzing the outcomes, we frequently find the significance of honouring clinical procedures' logical sequence documents. In the literature, methods for identifying cardiac disease include waveform analysis, time-frequency analysis, Neuro-Fuzzy RBF ANN, and Total Least Square-based Prony modelling algorithms. However,
Marshall et al investigation's found (Marshall et al 1991), This method's classification accuracy (up to 79%) was poor, and choosing the best model required considering a range of modifications was still adequate.

Heart Attack Prediction and Visualization of Contributing Factors Using Machine Learning" It associates many risk factors in heart disease and a need for time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data analysis and machine learning are the most commonly used techniques for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes like age, gender, chest pain, cholesterol, etc which are used to predict heart attack, and the model is trained using 4 machine learning algorithms namely- Logistic Regression, Gaussian Naïve Bayes, Decision tree, and Random Forest algorithm. It uses the existing dataset from the UCI Heart Disease Data set of heart disease patients. The dataset comprises 303 instances and 76 attributes. This research paper aims to envision the probability of developing heart attacks in patients. The results portray that the highest accuracy score is achieved with Logistic Regression.

This research paper "A novel approach for heart disease prediction using strength scores with significant predictors" talks about CVDs are disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, and other conditions. Heart attacks and strokes are the main causes of mortality in cardiovascular disease with a rate near one out of three. With the high rate of mortality, diagnosis and prevention measures need to be performed effectively and efficiently. Many data mining techniques have been used to help address these issues. Most of the past Heart Attack Prediction and Visualization of Contributing Factors Using Machine Learning research looked into identifying features that contribute to better heart prediction accuracy. However, very little research looked into the relationships that exist between these features. The association between each feature that contributes to heart disease prediction can be obtained using the Associative Rule Mining (ARM) technique.

The ARM technique is popular in transactional and relational datasets. The hidden knowledge in large datasets such as business transactions developed the interest of many

business owners to understand the patterns that can help them to improve their business decisions (Agarwal and Mithal). For instance, discovering the frequently bought items by customers in market basket analysis.

2.2 References

[1]" Predicting the Risk of Heart Failure With EHR Sequential Data Modelling," Bo Jin, Chao Che.

[2]"Heart Attack Prediction and Visualization of Contributing Factors Using Machine Learning" by Megha Banerjee, Reetodeep Hazra, Suvranil Saha, Megha Bhushan, Subhankar Bhattacharjee.

[3]"A novel approach for heart disease prediction using strength scores with significant Predictors" by Armin Yazdani, Kasturi Dewi Varathan, Yin Kia Chiam, Asad Waqar Malik, and Wan Azman Wan Ahmad.

[4]"Heart Disease Risk Prediction Using Machine Learning Classifiers with

Attribute Evaluators" by Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam, Hui Na Chua, and S. Pranavanand

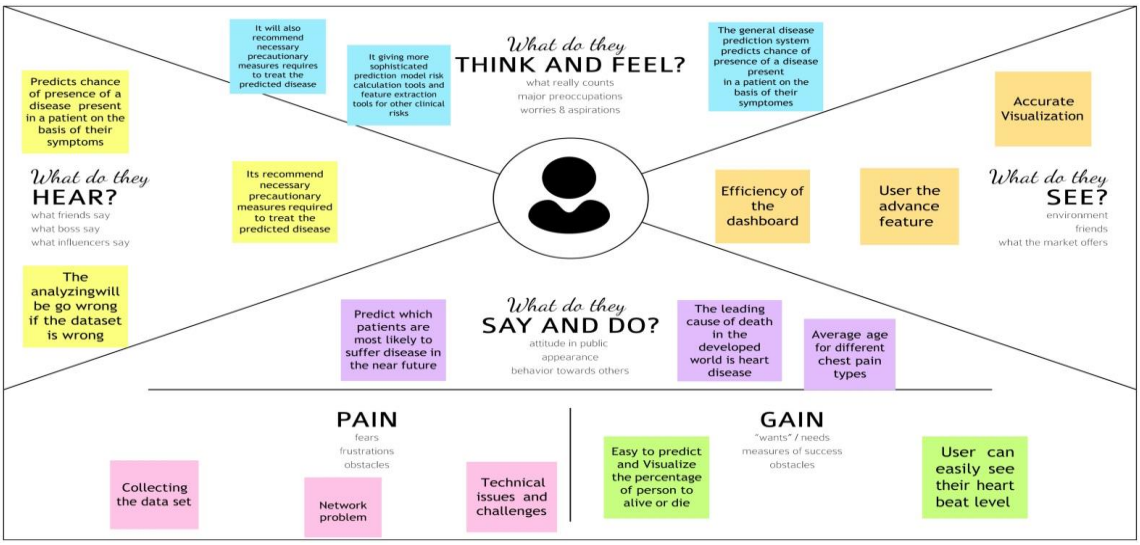
2.3 Problem Statement Definition

Heart acts a major role in the corporeal organisms. The diseases of the heart want more perfection and exactness for diagnosis and analyses. Heart disease is a dangerous disease. This disease occurs due to various problems such as overpressure, blood sugar, high blood pressure, Cholesterol, etc. in the human body By using Python and machine learning, this paper is analyzed and predicted heart disease. We can predict this disease by using various attributes in the data set. We have collected a data set consisting of 13 elements and 383 individual values to analyze the patient's performance. The main aim of the paper is to get better accuracy to detect heart disease using the ML algorithm.

Problem Statement(PS)	**I am (Customer)**	**I'm trying to**	**But**	**Because**	**Which makes me feel**
PS-1	Cust1	Check if I have any heart disease	I have Diabetes	Age factor	Stressed
PS-2	Cust2	Check if I have any heart disease	I am completely healthy	Healthy Diet	Relieved

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



A psychological aid to the general practitioner creates an empathy map for a patient with a lifestyle issue during a consultation.

3.2 Ideation & Brainstorming

In order to make the dataset we are working with easier to understand, we will first take a look at data wrangling. A timely diagnosis of heart disease (HD), one of the most prevalent diseases today, is essential for many healthcare professionals to protect their patients from the condition and save lives. To accurately classify and/or predict HD cases with a few variables, a comparison analysis of various classifiers can be done for the classification of the heart disease dataset.

Accurate decision-making and ideal therapy are needed to address cardiac risk. Five machine learning models were utilized in a Canadian study to examine 1-month mortality among hospitalized patients with congestive heart failure. Several tests, including auscultation, blood pressure, cholesterol, ECG, and blood sugar, are carried out prior to the diagnosis of a condition. These tests assist in identifying the patient's medication requirements. In this work, the predictive accuracy of various machine learning methods is investigated to calculate cardiovascular risk. The performance comparison of the most recent REP Tree and Random Tree machine learning algorithms in terms of cardiovascular disease prediction is innovative.

Pre-processing of Data:

You can eliminate the missing numbers or use the mean value in their place. Therefore, employing a filtering strategy, the data obtained must be slightly adjusted in order to conduct a good study. Here, the multifilter method is applied.

Extraction of Features:

Reduce the number of input attributes prior to data processing. Not all characteristics affect prediction success in the same way. Multiple attributes lead to increased complexity and worse performance. It is necessary to carefully extract features without sacrificing system performance as a result.

Data Collection and Processing Data Collection:

The dataset was compiled from multiple patient records that included 14 attributes such as restagc, fbs, thal, ca, cp, sex, age, thalach, chol, trestbps, slope, oldpeak, exangand target. The table below contains a description of the attributes used for analysis. The visualization shows that age does not play a significant role in predicting heart disease because the same age groups have an equal number of people with and without heart disease. Outlier detection and data preprocessing. For each attribute or feature, we calculate the Z-score of each individual value of that attribute in relation to the column mean and standard deviation. After that, take the magnitude or absolute value of the obtained z score. If the z score is less than a certain threshold, a particular row or record is an outlier and is removed.

3.3 Proposed Solution

3.3.1. 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.

3.3.2. Idea / Solution description

An interactive dashboard that visualizes and predicts using Machine Learning algorithms.

3.3.3. Novelty / Uniqueness

We are employing the Naive Bayes algorithm in this system to develop an efficient heart attack prediction system. The system can receive input from a CSV file or manually. After receiving input, the Nave Bayes algorithm is used on that input. After gaining access to the data set, the process is carried out, and a useful heart attack level is generated. The suggested approach will include additional heart attack risk factors such as weight, age, and priority levels after speaking with doctors and other medical professionals.

3.3.4. Social Impact / Customer Satisfaction

It helps in time of emergencies. If any heart problem is predicted in advance, we can provide valuable insights to clinicians, allowing them to tailor their diagnosis and treatment to each individual patient.

3.3.5. Business Model (Revenue Model)

For prediction, the algorithm makes use of 15 medical variables, including age, sex, blood pressure, cholesterol, and obesity. The likelihood that a patient may develop heart disease is predicted by the EHDPS. It makes it possible to build linkages between important knowledge, including patterns between medical parameters associated with heart disease.

3.3.6. Scalability of the Solution

Our key contribution to this work is to predict heart disease diagnosis using a modest number of parameters. Our prediction system employs random forest on Apache Spark, allowing healthcare analysts to deploy this solution on a constantly changing, scalable big data landscape for informed decision-making. We demonstrate that this method achieves up to 98% accuracy. We also compare our classifier to the Nave-Bayes classifier.

ALGORITHMS USED:

1. KNN Algorithm:

One of the simplest machine learning algorithms, based on the supervised learning method, is K Nearest Neighbour.

The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.

A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilising the K-NN method, fresh data can be quickly and accurately sorted into a suitable category.

Although the K-NN approach is most frequently employed for classification problems, it can also be Utilised for regression.

Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

```
knn = KNeighborsClassifier(n_neighbors = 20)
knn.fit(KX_train, KY_train)
print(knn.score(KX_test, KY_test))
```

```
0.6111111111111112
```

Training the data set using KNN algorithm

2.Random Forest Classifier:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

```
from sklearn.metrics import accuracy_score
max_accuracy = 0

for x in range(500):
    rf_classifier = RandomForestClassifier(random_state=x)
    rf_classifier.fit(X_train,y_train)
    Y_pred_rf = rf_classifier.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

print(max_accuracy)
print(best_x)
```

Overall Comparison of all 2 Algorithm:

```
scores = [score_rf,score_knn]
Models = ["Random Forest Classifier"," K-Nearest Neighbors Classifier"]

for i in range(len(Models)):
    print("The accuracy score achieved using "+Models[i]+" is: "+str(scores[i])+" %")
```

The accuracy score achieved using Random Forest Classifier is: 85.19 %
The accuracy score achieved using K-Nearest Neighbors Classifier is: 64.81 %

```
sns.set(style="darkgrid",rc={'figure.figsize':(20,10)})
plt.xlabel("Models")
plt.ylabel("Accuracy score")

sns.barplot(Models,scores)
plt.savefig("AccuracyScores.png")
```

3.4 Problem Solution fit

3.4.1. Customer Segment(s)

All adults. Especially people who are elder than 40 years and those who are on the verge of getting heart disease due to various factors such as age, obesity, diabetes, stress, etc.

3.4.2. Jobs-To-Be-Done / Problems J&P

To predict and identify the heart disease patient. It is a very useful strategy that was used to control how the model can be utilized to increase the accuracy of the prediction of Heart Attack in each.

3.4.3. Triggers TR

The generation currently living now leads an extremely unhealthy lifestyle. People worry about the sharp rise in mortality from heat -related illnesses. They, therefore, desire to adopt a better lifestyle.

3.4.4. Emotions: Before / After EM

People frequently worry that their health will decline. They suffer unneeded tension and emotional breakdowns as a result of this. Our prediction system would enable them to keep track of their health independently and assist them in overcoming their erroneous concerns.

3.4.5. Available Solutions

EDA: Exploratory data analysis is the key step for getting meaningful results.

Pros: Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc. Discover errors, outliers, and missing values in the data. Identify patterns by visualizing data in graphs such as box plots, scatter plots, and histograms.

Cons: Exploratory research comes with disadvantages that include offering inconclusive results, lack of standardized analysis, a small sample population, and outdated information that can adversely affect the authenticity of the information.

3.4.6. Customer Constraints

The patient's medical status must be continuously monitored. Unpredictability could lead to inaccurate results. The patient must be genuine about the periodic readings they record. The process could consume the internet and could be slightly expensive.

3.4.7. Behaviour

To solve their problem, a suitable application must be available. To effectively diagnose the situation for their present health status, appropriate information such as age, weight, current symptoms, and cholesterol should be provided.

3.4.8. Channels of Behaviour

3.4.8.1 ONLINE • Data Collected from offline devices is used in this application in order to visualize and predict heart diseases

3.4.8.2 OFFLINE • ECG • Blood Sugar Level • Blood Pressure • Cholesterol

3.4.9. Problem Root Cause

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.

3.4.10. Your Solution

We classified heart disease using python and pandas operations for the data taken from the repository after analyzing the outcomes from the existing approaches. It offers a simple-to-understand visual depiction of the dataset, the working environment, and the process of developing predictive analytics.

The machine learning (ML) process begins with a data pre processing phase, which is followed by feature selection based on data cleaning, classification, and modeling performance evaluation. The accuracy of the outcome is increased using the Naive Bayes approach. Unsupervised machine learning is a branch of data analytics that is used for the prediction of heart disease. Unsupervised machine learning techniques include K-means clustering. Unsupervised algorithms often forecast the intended result without reference to any values. The K-means clustering technique clusters the data so that there is a high intraclass similarity and a low inter-class similarity. The sum of squares' distance from the cluster centroid is reduced by this algorithm. The program creates k clusters with a centroid out of the data. K-means iteratively locates the center which minimizes the separation between the cluster's individual points and its center.The k-means clustering technique is demonstrated in the following flow chart.

This section presents the findings of the data analysis conducted to find the necessary hidden patterns for forecasting cardiac illnesses. Age, the type of chest pain, blood pressure, blood sugar level, resting ECG, heart rate, the four different types of chest pain, and exercise-induced angina are the variables here taken into account to predict heart disease. Pre-processing the heart disease dataset efficiently involves removing irrelevant records and assigning values to tuples that are missing. The K-means technique is then used to put together the pre-processed heart disease data set. In this article, four different types of heart diseases-asymptomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain-are explored. Visualization of Data visualization is a crucial phase in the data science process that enables teams and individuals to communicate data to co-workers and decision-makers more effectively. Teams that oversee reporting systems frequently use predefined template views to keep an eye on efficiency. However, performance dashboards aren't the only applications for data visualization. For instance, while text mining unstructured data, an analyst might employ a word cloud to identify important ideas, patterns, and undiscovered connections. As an alternative, they can show the connections between things in a knowledge graph using a graph structure. It's crucial to keep in mind that there are numerous methods to represent various sorts of data, and that this is a set of abilities that should go beyond your core analytics team. In order to make the dataset we are working with easier to understand, we will first take a look at data Wrangling.It would enable us to make better use of the data. We have to import pandas, matplotlib and seaborn. We can now conduct exploratory data analysis after finishing data

wrangling. A timely diagnosis of heart disease (HD), one of the most prevalent diseases today, is essential for many healthcare professionals to protect their patients from the condition and save lives. To accurately classify and/or predict HD cases with few variables, a comparison analysis of various classifiers can be done for the classification of the heart disease dataset. Accurate decision-making and ideal therapy are needed to address cardiac risk. Five machine learning models were utilized in a Canadian study to examine 1-month mortality among hospitalised patients with congestive heart failure. Several tests, including auscultation, blood pressure, cholesterol, ECG, and blood sugar, are carried out prior to the diagnosis of a condition. These tests assist in identifying the patient's medication requirements. In this work, the predictive accuracy of various machine learning methods is investigated to calculate cardiovascular risk.

4. REQUIREMENT ANALYSIS:

4.1 Functional Requirements:

Functional Requirements: These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

User Registration Registration through Form Registration through Gmail Registration through Google

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	Data Analysis	Obtained data is analysed and segregated based on set criteria (blood pressure, cholesterol levels etc.)
FR-4	Data Visualization and dashboard creation	User can visualise the trends on the heart diseases via graphs, charts etc. in the IBM Cognos dashboard that is created Reports are created based on the trends that the user can view

4.2.Non-Functional Requirements

These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to another.

They are also called non-behavioral requirements. The system should be described by non-functional requirements. behave in a practical manner and impose limitations. This sort is

known as the system's quality of requirements attributes. Features like performance, security, and usability, compatibility is not system feature; rather, they an essential quality.

We are unable to separate ourselves from the entire setup. To execute them, write a specific line of code. Any of the customer's desired characteristics are listed in the specification. We can only mention the requirements that are suitable for our project. They basically deal with issues like

- Usability
- Security
- Reliability
- Performance
- Availability
- Scalability

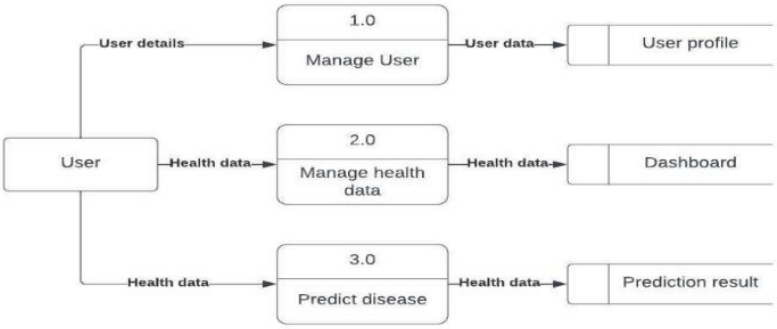
Following are the non-functional requirements of the proposed solution.

NFR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The application will have a simple and user friendly graphical interface. Users will be able to understand and use all the features of the application easily. Any action has to be performed with just a few clicks
NFR-2	**Security**	For security of the application the technique known as database replication should be used so that all the important data should be kept safe. In case of crash, the system should be able to backup and recover the data
NFR-3	**Reliability**	The application has to be consistent at every scenario and has to work without failure in any environment
NFR-4	**Performance**	Performance of the application depends on the response time and the speed of the data submission. The response time of the application is direct and faster which depends on the efficiency of implemented algorithm
NFR-5	**Availability**	The application has to be available 24 x 7 for users without any interruption
NFR-6	**Scalability**	The application can withstand the increase in the no. of users and has to be able to develop higher versions

4. 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**.**



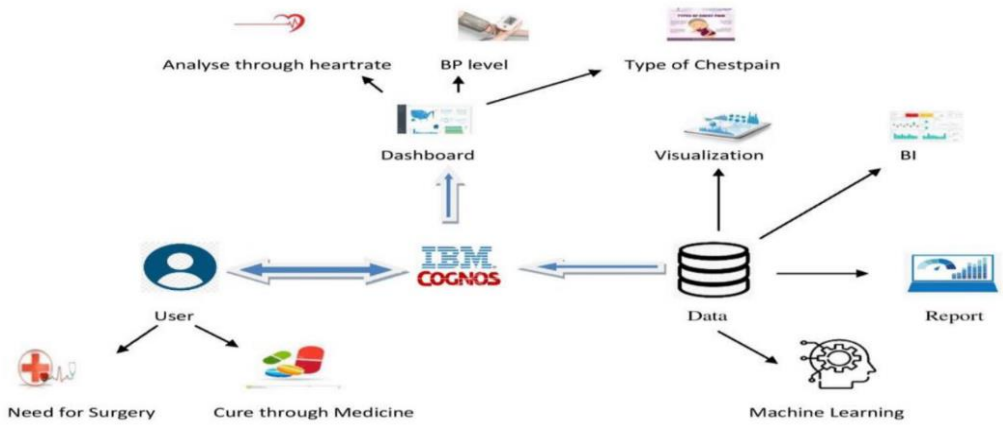
5.2 Solution & Technical Architecture

Solution Architecture:

The Software Architecture is followed by the following steps:

1. The patient first registers by providing certain parameters.
2. That registered data is collected in a database using machine learning techniques similar to data collection techniques, and when he goes to check on his health, the collected values or data that has been stored in the database is extracted.
3. Classification is done using some feature extraction methods when data is extracted, it goes through certain processes, and as a result, a disease is created.
4. The predicted data and a report are produced.

This is an overview of the machine-learning-based heart disease prediction system.



Technical Architecture:

The Deliverable shall include the architectural diagram as below and the information as per the table1 & table 2

Example: Order processing during pandemics for offline mode

Reference:https://developer.ibm.com/patterns/ai-powered-backend-system-for-order-processing-during-pandemics/

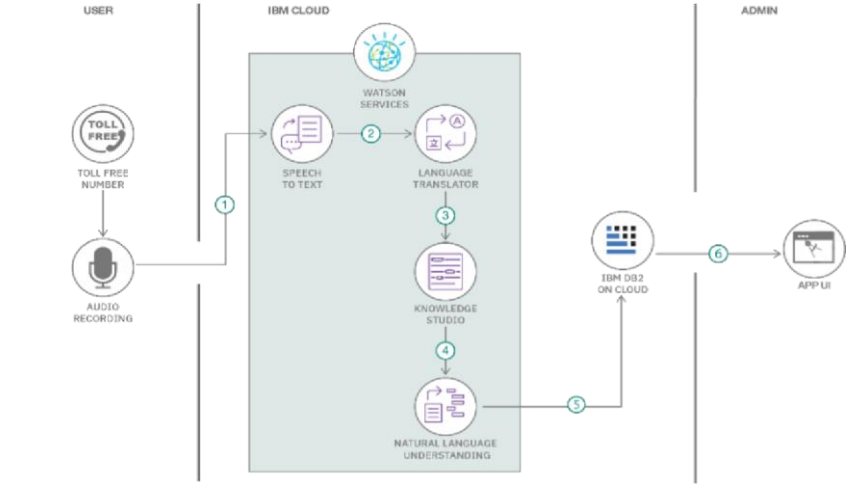


Table - 1: Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	HTML, CSS, JavaScript / Angular Js / React Js etc.

2.	Application Logic-1	Logic for a process in the application	Java / Python
3.	Application Logic-2	Logic for a process in the application	IBM Watson STT service
4.	Application Logic-3	Logic for a process in the application	IBM Watson Assistant
5.	Database	Data Type, Configurations etc.	MySQL, NoSQL, etc.
6.	Cloud Database	Database Service on Cloud	IBM DB2, IBM Cloudant etc.
7.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
8.	External API-1	Purpose of External API used in the application	IBM Weather API, etc.
9.	External API-2	Purpose of External API used in the application	Aadhar API, etc.
10.	Machine Learning Model	Purpose of Machine Learning Model	Object Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, Cloud Foundry, Kubernetes, etc.

Table - 2: Application Characteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Technology of Opensource framework
2.	Security Implementations	List all the security / access controls implemented, use of firewalls etc.	e.g. SHA-256, Encryptions, IAM Controls, OWASP etc.
3.	Scalable Architecture	Justify the scalability of architecture (3 – tier, Micro-services)	Technology used
4.	Availability	Justify the availability of application (e.g. use of loadbalancers, distributed servers etc.)	Technology used
5.	Performance	Design consideration for the performance of the application (number of requests per sec, use of Cache, use of CDN's) etc	Technology used

5.3 User Stories:

Story 1:

As an aging individual, I want an application so that I could predict my cardiac health. Let's break this down one step further, As the user is an aging individual, we are building a heart disease-predicting application that enables the user to predict their cardiac health immediately within a few seconds. The app has the user login and signup for the authentication of information, and it uses the Logistic Regression algorithm to predict the result. We have visualized the user's query for their requirement only concerning the Age and Cholesterol of the user. The prediction gives a result if the disease could be present or not. As a working person, I want an application so that I can predict my own cardiac health even while I am at work.

Story 2:

When we address this user issue, As the user is a working person, our heart disease predicting application would enable the user to predict his health just by simply typing the values into the boxes given. The process is very simple, it takes only a few seconds to view the results of the process. The app has the user login and signup for the authentication of information, and it uses the Logistic Regression algorithm to predict the result. The users can customize their range of values and can visualize the entire result in a graphical representation. It allows the user to have a clear knowledge of what must be done and what to be neglected.

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	Datasets	USN-1	As an analyst,I will develop code for data preparation and data description.	5	High	ANITHA B
Sprint-2	Cleaning, exploring data and creating model	USN-2	As an Analyst I will develop code for data exploration.	5	High	VANASUNDARI S
Sprint-3	Data visualization	USN-3	As an Analyst I can develop code for data visualization.	5	High	SWETHA P
Sprint-4	Data Prediction	USN-4	As a Data analyst, I will create code for different types of models in explored data	5	High	SARAVANAPRIYA R

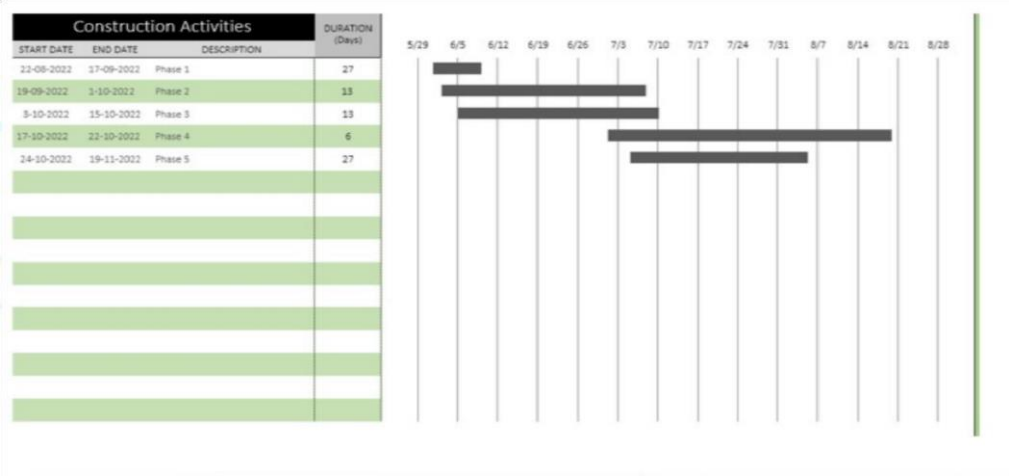
6.2.Sprint Delivery Plan

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	10	5 Days	24 Oct 2022	29 Oct 2022	10	29 Oct 2022

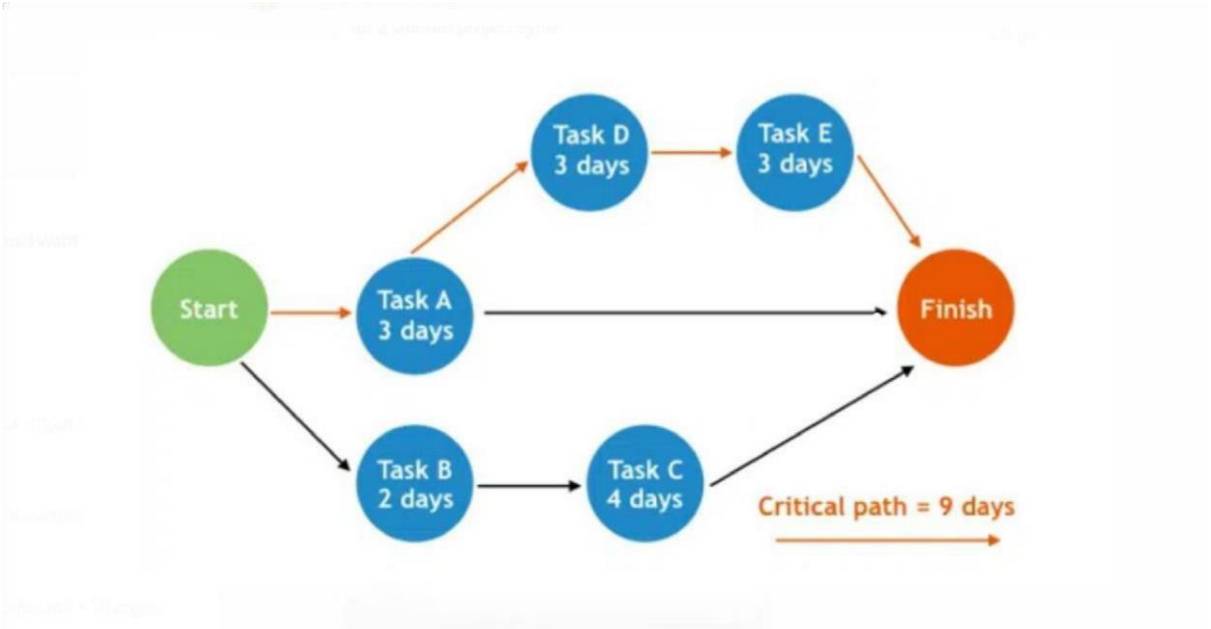
Sprint-2	10	5 Days	31 Oct 2022	05 Nov 2022	10	05 Nov 2022
Sprint-3	10	5 Days	07 Nov 2022	12 Nov 2022	10	12 Nov 2022
Sprint-4	10	5 Days	14 Nov 2022	19 Nov 2022	10	19 Nov 2022

Gantt Charts:

A Gantt chart is a diagram that shows all the jobs that are planned to be completed late for your project. Plans for projects of all sizes and kinds are made using them.



Project management is not a straightforward science, as we all know. It is an intricate synthesis of many different ideas, from strategy to people management, and IT communications to figure crunching. This article will teach us about the most crucial resources available to managers. They are used to plan or schedule tasks for projects, identify critical paths, keep track of progress, and do all other crucial duties required for a smooth project cycle.



Velocity:

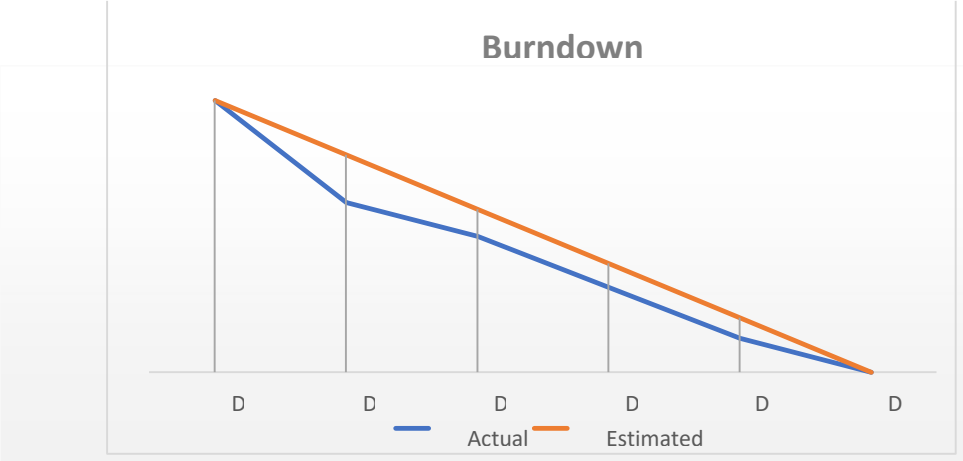
Imagine we have a 05-day sprint duration, and the velocity of the team is 10 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

AV=Sprint Duration/Velocity=10/5=2

Burn down Chart:

A burn down chart is a graphical representation of work left to do versus time. It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.

Goal: 60 hours in 5 days



7. RESULTS

7.1. Performance Metrics

Metrics are measurements and parameters obtained throughout the quality assurance procedure. They may make reference to several test types. As you might have guessed, performance testing data gives you the ability to evaluate the efficiency of performance testing. Alternatively said, these measurements demonstrate how well software reacts to user scenarios and manages user flow in real time.
The following two categories of data are appropriate:

1. Measurements are data that are kept track of when testing, such as how long it takes to react to a request.
2. Metrics, which include various types of percentages, average indicators, and other metrics, are computations performed with the aid of certain formulas.

Accuracy Percentages:

The number of wrong predictions on the test set as a whole divided by all of the test set predictions yields the error rate. Since accuracy and error rate are complementary quantities, we can always compute one from the other.

$$\text{Accuracy} = 1 - \text{Error Rate}$$
$$\text{Error rate} = 1 - \text{Accuracy}$$

Logistic Regression: 0.82
KNN: 0.61
Naïve Bayes: 0.77
Decision Tree: 0.79

Response Time:

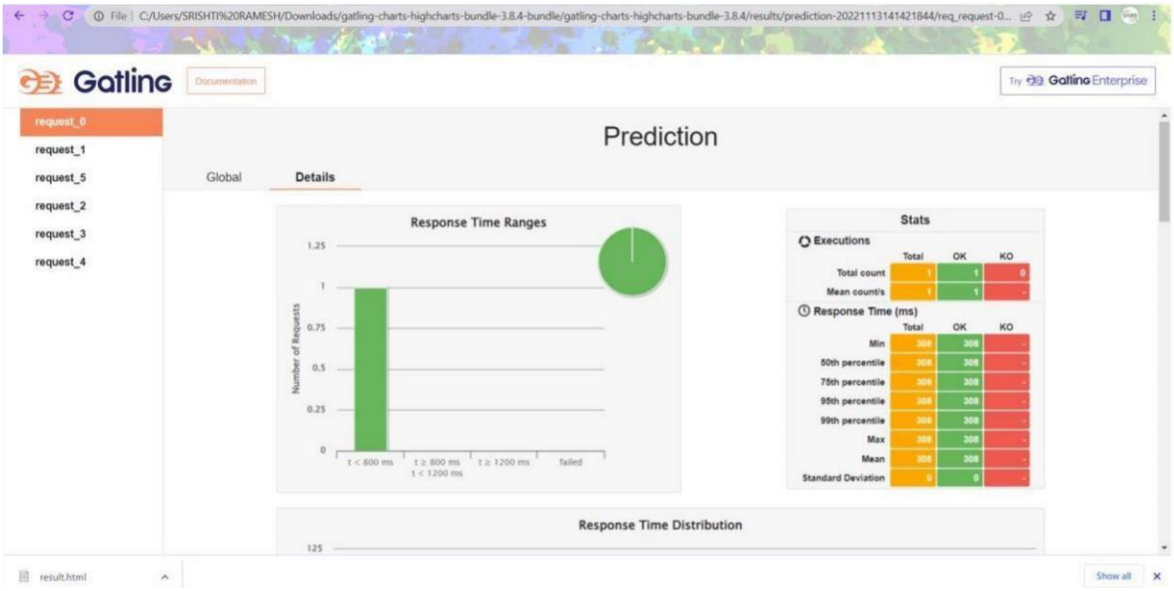
The response time is not too long as our project as we have used real time data analysis. So, once the user enters his/her data in the Heart Disease prediction phase then the data will immediately be displayed so the response time is very less.

Requests per second:

An HTTP request is created and sent to a server by a client application. The client receives a request from the server, which processes it, produces a response, and sends it back. The measure we are interested in is requests per second, which is the total number of consistent requests per second (RPS). These can be requests for any type of data source, including XML documents, HTML files, multimedia files, and JavaScript libraries.

User transactions:

A user transaction is a series of user operations performed through a software interface. You may determine how well the system survived the load testing by comparing the number of transactions per second or actual transaction time with the projected number of transactions per second.



Virtual users per unit of time:

This statistic also aids in determining whether the performance of a software product satisfies the stated requirements. A QA team finds it useful to predict average load as well as programme behavior under various load levels.

Error rate:

The ratio of incorrect to correct replies over time is used to calculate this measure. Percentages are used to calculate the findings. When software load surpasses its capacity, mistakes frequently happen.

8. ADVANTAGES & DISADVANTAGES

8.1. Advantages

a. The proposed work predicts the chances of Heart Disease and classifies patient's risk level

b. It is implementing different data mining techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest.

c. User friendly

18.2. Disadvantages

a. Data analytics techniques do not help to. provide effective decision making.

b. Cannot handle enormous datasets

c. Prediction of cardiovascular disease results is not accurate

9. CONCLUSION

The long-term preservation of human life and the early identification of irregularities in heart problems will benefit from the identification of the processing of raw healthcare data related to the heart. In this study, raw data was processed using machine learning techniques to produce a brand-new understanding of cardiac disease. In the medical field, heart disease prediction is difficult and crucial. The death rate, however, can be significantly reduced if the disease is identified in its early stages and preventative measures are put in place as soon as feasible. To move the investigations from simply theoretical frameworks and simulations to actual datasets, further elaboration of this study is extremely desirable. The model's ability to be employed to increase the precision of heart attack prediction in any individual was regulated using a very helpful technique. When compared to the previously employed classifiers, such as naive bayes, etc., the proposed model's strength was quite satisfying. It was able to predict signs of having a heart illness in a specific individual by applying KNN and Logistic Regression, which demonstrated good accuracy. Therefore, by utilizing the provided model to determine the likelihood that the classifier will correctly and reliably detect the heart illness, a large amount of pressure has been reduced. The Given heart disease prediction system improves and lowers the cost of medical care. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

10. FUTURE SCOPE

This study discusses the issue of constricting and summarizing various data mining strategies utilized in the field of medical forecasting. For intelligent and successful heart attack prediction via data mining, the emphasis is on combining various methods and combinations of numerous target attributes. Significantly, 15 attributes are specified for predicting heart attacks, and using simple data mining techniques, other approaches, including ANN, time series, clustering and association rules, soft computing approaches, etc., can also be included. The results of predictive data mining on the same dataset show that Decision Tree outperforms and, occasionally, Bayesian classification has accuracy levels comparable to those of decision tree, but other predictive methods, such as KNN, Neural Networks, and Classification based on clustering, are not performing well. The second finding is that using a genetic algorithm to lower the actual data quantity and obtain the ideal subset of attributes suitable for heart disease prediction increases the decision tree and Bayesian classification's accuracy. For the automation of heart disease prediction, the proposed work can be expanded and improved. Real data from healthcare institutions and agencies must be gathered, and all methods must be compared for the highest level of accuracy**.**

11. APPENDIX

Source code

```
Dataset=pd.read_csv('
Heart_Disease_Predict
ion.csv',sep=',',encodin
g="utf-8")

type(dataset)

dataset.shape

dataset.info()

dataset.columns

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

target = df['Heart Disease'].map({'Presence':1, 'Absence':0})
inputs = df.drop(['Heart Disease'], axis=1)

plt.suptitle("Correlation Map/Pearson correlation coefficient")
sns.heatmap(df.iloc[:,1:-1].corr())

plt.show()

data=df
sns.barplot(x=data.Age.value_counts()[:10].index,y=data.Age.value_counts()[:10].values)
plt.show()

plt.suptitle("Age")
sns.scatterplot(data=df, x='Age',
y=np.zeros(len(df['Age'])), hue=target)
plt.show()

minAge=min(data.Age)
maxAge=max(data.Age)
meanAge=data.Age.mean()
print('Min Age:',minAge)
print('Max Age:',maxAge)
print('Mean Age:',meanAge)

plt.suptitle('Age histogram', fontweight='heavy')
plt.title('The age average is around 54')
```

```
sns.histplot(data=df,
x='Age')
plt.show()

labels = ['Male',
'Female']
order =
df['Sex'].value_counts(
).index

plt.figure(figsize=(10,5
))
plt.suptitle("Sex
(Gender)")

plt.subplot(1,2,1)
plt.title('Pie chart')
plt.pie(df['Sex'].value_
counts(), labels=labels,
textprops={'fontsize':1
2})

plt.subplot(1,2,2)
plt.title('Count plot')
sns.countplot(x='Sex',
data=df, order=order)
plt.xticks([0, 1], labels)

plt.show()

print(df['Sex'].value_c
ounts())
print("It can be noticed
that predictor (Gender)
is imbalance")

labels = ["typical
angina", "atypical
angina", "non-anginal
pain", "asymptomatic"]
order = df['Chest pain
type'].value_counts().i
ndex

plt.figure(figsize=(10,5
))
```

```
plt.suptitle("Chest pain type")

plt.subplot(1,2,1)
plt.title('Pie chart')
plt.pie(df['Chest pain type'].value_counts(), textprops={'fontsize':12})
plt.subplots_adjust(left=0.125)

plt.subplot(1,2,2)
plt.title('Count plot')
sns.countplot(x='Chest pain type', data=df, order=order)
plt.xticks([0,1,2,3], labels, rotation=45)

plt.show()

df['Chest pain type'].value_counts()

knn_classifier= KNeighborsClassifier(n_neighbors=31,leaf_size=30)
knn_classifier.fit(X_train,y_train)
Y_pred_knn = knn_classifier.predict(X_test)
score_knn = round(accuracy_score(Y_pred_knn,y_test)*100,2)
score_knn

y_pred_knne = knn_classifier.predict(X_test)

plt.figure(figsize=(10, 8))
CM=confusion_matrix(y_test,y_pred_knne)
sns.heatmap(CM, annot=True)
```

```
TN = CM[0][0]
FN = CM[1][0]
TP = CM[1][1]
FP = CM[0][1]
specificity = TN/(TN+FP)
loss_log = log_loss(y_test, y_pred_knne)
acc= accuracy_score(y_test, y_pred_knne)
roc=roc_auc_score(y_test, y_pred_knne)
prec = precision_score(y_test, y_pred_knne)
rec = recall_score(y_test, y_pred_knne)
f1 = f1_score(y_test, y_pred_knne)

mathew = matthews_corrcoef(y_test, y_pred_knne)
model_results =pd.DataFrame([['K-Nearest Neighbors',acc, prec,rec,specificity, f1,roc, loss_log,mathew]], columns = ['Model', 'Accuracy','Precision', 'Sensitivity','Specificity', 'F1 Score','ROC','Log_Loss','mathew_corrcoef'])

model_results

from sklearn import metrics
```

```
Y_pred_knn = np.around(Y_pred_knn)
print(metrics.classification_report(y_test,Y_pred_knn))

from sklearn.metrics._plot.roc_curve import plot_roc_curve
plot_roc_curve(knn_classifier,X_test,y_test)
plot_roc_curve
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic Curve');
plt.savefig("KNN.png")
```

GitHub & Project Demo Link:

https://github.com/IBM-EPBL/IBM-Project-32980-1660213350