

**Project Report**

**Heart Attack Risk Prediction System**  
**Using Machine Learning**

**Group 12**

**Under the guidance of**  
**Prof. Safa Shubbar**



***Department of Computer Science***

Name	Cotribution
SHASHANK UPPALAMURTHY	Handled data preprocessing and machine learning model implementation, ensuring the algorithms were optimized for accuracy and deployed the application in streamlit cloud community.
VINAY KANCHARAM	Worked on front-end development, building the Streamlit interface, and ensuring the system's usability and responsiveness.
MOUNIKA CHENNUBOINA	Focused on integrating the chatbot and preparing the dataset for effective training and analysis.
MALREDDY VEERAREDDY	Led clustering implementation using K-means and created data visualizations to provide meaningful insights.

## Contents

1. Introduction .....	4
1.1 Motivation .....	5
1.2 Real Applications .....	5
2. Project Description .....	6
2.1 Heart Attack Risk Prediction .....	6
2.2 Data Preprocessing .....	6
2.3 Clustering with K-means .....	6
2.4 Data Visualization .....	6
2.5 Chatbot Integration .....	6
2.6 Challenges and Technical Contributions .....	7
2.7 Workload Distribution .....	7
3. Background .....	7
3.1 Machine Learning Models in Healthcare .....	7
3.2 Random Forest Classifier .....	7
3.3 Naive Bayes Classifier .....	7
3.4 K-means Clustering .....	8
3.5 Dataset Details .....	9
3.6 Related Papers and Surveys .....	9
3.7 Software Tools .....	10
3.8 Required Hardware .....	10
3.9 Related Programming Skills .....	10
4. Problem Definition .....	10
4.1 Heart Attack Risk Prediction .....	10
4.2 Clustering Patients .....	11
4.3 Data Visualization .....	11
4.4 Chatbot Integration .....	11
4.5 Formal Problem Definitions .....	11
4.6 Challenges in Addressing Problems .....	11
5. Proposed Techniques .....	
5.1 Machine Learning Techniques.....	12
5.1.1 Random Forest Classifier.....	12
5.1.2 Naive Bayes Classifier.....	12
5.1.3 K-means Clustering.....	12
5.2 Data Preprocessing.....	13

5.3 Data Visualization.....	13
5.4 Chatbot Integration.....	13
6. Visual Applications	
6.1 GUI Design.....	13.
6.2 Login, Signup, and Dashboard Page.....	14
6.3 Heart Attack Risk Prediction Page.....	14
6.4 Clustering Visualization.....	18
6.5 Data Visualization.....	18
6.6 Chatbot.....	21.
7. Experimental Evaluation	
7.1 Random Forest Classifier Evaluation.....	23
7.2 K-means Clustering Evaluation.....	23
7.3 Naive Bayes Classifier Evaluation.....	24
7.4 User Experience.....	24
7.5 Experimental Settings.....	24
8. Future Work	
8.1 Model Optimization.....	25
8.2 Incorporating More Features.....	25
8.3 Real-time Data Integration.....	25
8.4 Improved Chatbot.....	25
8.5 Integration with Electronic Health Records (EHR) .....	25
8.6 Mobile App Development.....	26
8.7 Integration of Social Determinants of Health (SDH) .....	26
8.8 Advanced Visualization and Analytics.....	26
8.9 Cross-Platform Deployment.....	26
9.Conclusion.....	27
10.References.....	28

## **1. Introduction**

Particularly, heart attacks are still one of the main causes of death worldwide. The World Health Organization (WHO) reports that heart attacks and other cardiovascular diseases claim millions of lives each year. This concerning figure highlights the need of early detection and prompt care in lowering heart disease-related death rates. Predictive models have become useful tools in healthcare with the rise of artificial intelligence and machine learning, allowing for early risk assessment and individualized treatment. The goal of this project is to create a Heart Attack Risk Prediction System with an easy-to-use Streamlit user interface using machine learning techniques, namely Random Forest classification models.

The program uses a dataset of medical indicators, including age, blood pressure, cholesterol, BMI, and others, to forecast an individual's risk of experiencing a heart attack. Users can enter their medical information to receive real-time forecasts using a smooth and user-friendly interface created with Streamlit, a Python framework for creating interactive apps. K-means clustering is also incorporated into the system to help with targeted interventions by grouping patients with comparable risk factors. In order to give users instant answers to commonly requested queries about heart health, food, and medicine, a chatbot is also implemented.

This study shows how technology may help prevent and treat heart disease by combining machine learning, data visualization, and a user-interactive platform, enabling users to make knowledgeable health decisions.

### **1.1 Motivation:**

This project is driven by the urgent need for practical, easily available tools that can assist in early risk assessment and mitigation of heart disease. Heart attacks in particular constitute a significant public health burden on a global scale. Predictive models are crucial for identifying high-risk individuals before a crisis occurs, and early intervention can greatly enhance results. This research uses machine learning techniques to close this gap by offering real-time evaluations that both patients and medical professionals may use.

### **1.2 Real Applications:**

There are several possible real-world uses for this heart attack risk prediction system. It can be incorporated into healthcare systems, for example, to help physicians identify high-risk patients and provide individualized treatment regimens. With the use of this application, healthcare facilities may keep an eye on sizable patient databases and analyze their risk profiles to determine which patients need urgent care. People can use the system to track their health measurements over time, get lifestyle

change advice, and track their progress in a more customized format. Furthermore, the system can be coupled with wearable technology, such as smartwatches or heart rate monitors, to give continuous monitoring and real-time updates, providing a proactive approach to the prevention of cardiac disease.

## **2. Project Description**

The objective of the Heart Attack Risk Prediction System project is to create an accessible and interactive platform that can predict the likelihood of a heart attack based on an individual's health data. The system is built using Python's Streamlit library to develop a front-end user interface, while machine learning algorithms, specifically the Random Forest classifier, are used for the prediction task. The project integrates several components, including risk prediction, data preprocessing, clustering, data visualization, and chatbot integration.

**2.1 Heart Attack Risk Prediction:** The heart of the system lies in the machine learning model that predicts the risk of a heart attack based on key health metrics. Using a labeled dataset, a Random Forest classifier is trained to identify patterns and relationships between medical features (e.g., cholesterol levels, BMI, age) and heart attack risk.

**2.2 Data Preprocessing:** Before training the machine learning model, data preprocessing is necessary to clean and transform the dataset into a suitable format for analysis. Missing values, outliers, and inconsistent data need to be handled to ensure that the model performs accurately. This step also involves normalizing numerical features and encoding categorical variables.

**2.3 Clustering with K-means:** K-means clustering is employed to segment the dataset into groups of patients with similar characteristics. By grouping patients based on their medical attributes, the system can identify specific high-risk clusters. This helps healthcare professionals understand common risk patterns and apply targeted treatment or preventive measures to those groups.

**2.4 Data Visualization:** Visualizations such as histograms, scatter plots, heatmaps, and bar charts are used to display health metrics, correlations between variables, and insights into heart attack risk. Data visualization enhances the understanding of how different factors (like cholesterol levels or BMI) influence the likelihood of a heart attack.

**2.5 Chatbot Integration:** The chatbot component of the system is designed to engage with users and answer queries about heart health. It can provide information about lifestyle changes, healthy diets, medication options, and preventive measures. The chatbot enhances the user experience by offering real-time, interactive assistance.

**2.6 Challenges and Technical Contributions:** During the project, several challenges were encountered, including handling an imbalanced dataset, ensuring model accuracy across diverse demographics, and designing an intuitive user interface. The technical contributions include creating a pipeline that integrates machine learning predictions with clustering algorithms for actionable insights and leveraging Streamlit for a highly interactive user experience. The incorporation of a chatbot tailored specifically for heart health FAQs further distinguishes this project as a comprehensive and user-centric tool.

**2.7 Workload Distribution:** The tasks were divided among the team to maximize efficiency and leverage individual expertise:

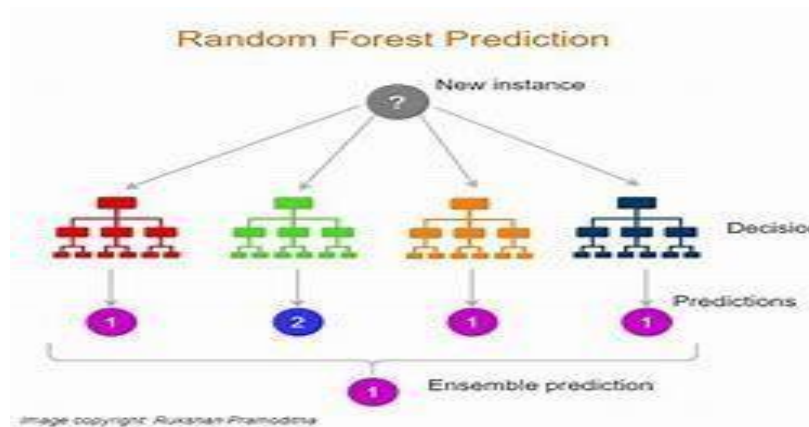
- **Shashank Uppalamurthy:** Handled data preprocessing and machine learning model implementation, ensuring the algorithms were optimized for accuracy and deployed the application in streamlit cloud community.
- **Mounika Chennuboina:** Focused on integrating the chatbot and preparing the dataset for effective training and analysis.
- **Vinay Kacharam:** Worked on front-end development, building the Streamlit interface, and ensuring the system's usability and responsiveness.
- **Veera Reddy Malreddy:** Led clustering implementation using K-means and created data visualizations to provide meaningful insights.

### 3. Background

Heart disease prediction has long been a key area of research in the field of healthcare informatics. The development of machine learning models to predict heart disease risk has gained significant traction, with algorithms capable of analyzing large datasets of medical records and identifying risk factors. Machine learning models, especially classification models like Random Forest, have shown to be particularly useful in medical diagnostics due to their high accuracy and ability to handle complex, non-linear relationships in data. These models can classify patients based on their risk profile, offering a tool for early intervention.

**3.1 Random Forest Classifier:** The Random Forest algorithm is an ensemble learning method that creates multiple decision trees, each trained on a random subset of the data. Each tree makes a prediction, and the final output is determined by aggregating the results of all the trees. Random Forest has been widely adopted in medical domains because of its robustness against overfitting, its ability to handle missing data, and its versatility with both numerical and categorical data. In heart attack prediction, Random Forest can process various patient features (e.g., age, cholesterol levels, and blood pressure) and provide an

accurate prediction of heart attack risk.



**Fig 1: Random Classification**

**3.2 Naive Bayes Classifier:** The Naive Bayes algorithm is a probabilistic machine learning method based on Bayes' Theorem, assuming independence among predictors. It is particularly well-suited for binary and multi-class classification problems due to its simplicity, speed, and ability to handle high-dimensional datasets effectively. For this project, Naive Bayes can be used as an alternative classification model to predict heart attack risk.

In the context of heart attack risk prediction, the algorithm calculates the probability of a person belonging to either the "high risk" or "low risk" category given their medical features such as age, cholesterol levels, blood pressure, and BMI. Mathematically, the classifier uses the formula:

$$P(Y|X) = P(X|Y) / P(Y)P(X)$$

Here:

- **$P(Y|X)$ :** The posterior probability of risk level Y (e.g., high or low) given input data X (e.g., medical attributes).
- **$P(X|Y)$ :** The likelihood of the input data given the risk level.
- **$P(Y)$ :** The prior probability of the risk level.

- **P(X):** The evidence, which is constant for all categories.

## Naive Bayes

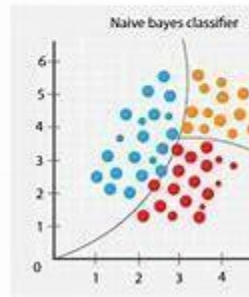
@thatware.co

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

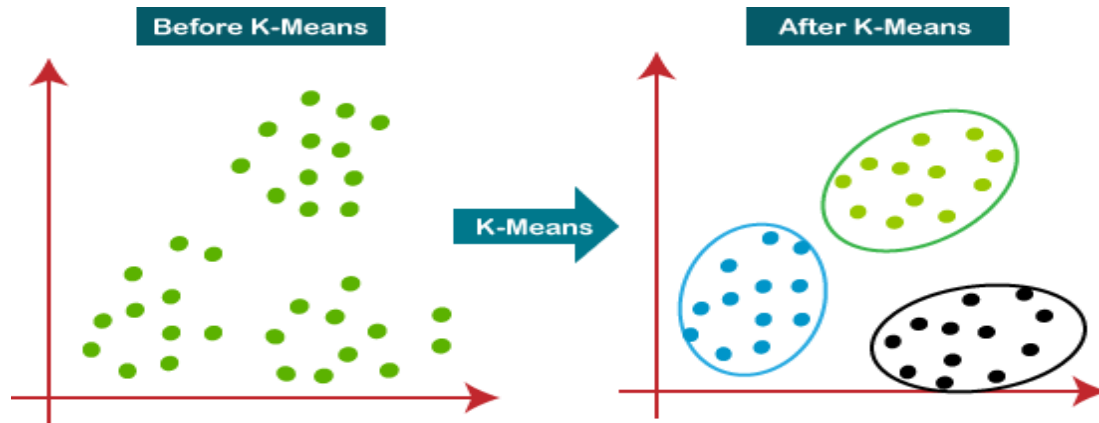


**Fig 2: Navie Bayes Classification**

The algorithm's strength lies in its ability to process large datasets quickly, making it ideal for real-time predictions. In this project, Naive Bayes can complement the Random Forest classifier by serving as a benchmark model to compare prediction performance and validate the dataset's structure. Additionally, its probabilistic outputs can enhance the interpretability of predictions, providing users with not only a classification result but also the associated likelihood of belonging to a specific risk category.



**3.3 K-means Clustering:** K-means is a widely used unsupervised learning algorithm for clustering data points into  $k$  distinct groups. In healthcare, clustering can help identify patients with similar medical profiles, which could point to common underlying health issues or similar risk factors for diseases like heart attacks. In this project, K-means clustering is employed to group patients based on their health metrics and risk levels. This segmentation can assist healthcare providers in identifying high-risk groups and implementing targeted interventions.



**Fig 3: K-Means Cluster Classification**

**3.4 Dataset Details:** The dataset used in this project is sourced from Kaggle's heart attack prediction dataset, which contains a variety of medical attributes linked to heart disease. These include age, sex, cholesterol, blood pressure, and electrocardiographic results, all of which are known to be significant predictors of heart attack risk. The ability to use machine learning models to accurately predict heart disease based on these features makes the approach highly relevant to the current healthcare needs.

**3.5 Related Papers and Surveys:** Numerous studies have demonstrated the importance of machine learning in predictive healthcare. For example, ensemble learning methods, such as Random Forest, are well-documented for their reliability and accuracy in classification tasks within the medical domain. Research into clustering techniques like K-means has also shown promise in patient segmentation for targeted interventions. Additionally, surveys on the application of AI in healthcare highlight the transformative potential of predictive analytics in reducing mortality rates associated with heart diseases. Recent studies also emphasize the role of chatbots in healthcare, showcasing their ability to provide personalized health advice and improve patient engagement. Another area of focus is the integration of user-friendly interfaces, like those built with frameworks such as Streamlit, which have been shown to enhance accessibility for non-technical users. Studies on data preprocessing techniques, such as normalization and handling missing values, underscore their critical role in improving model performance.

**3.6 Software Tools:** The project utilizes a combination of software tools to ensure efficiency and accuracy. Python libraries like Scikit-learn are employed for machine learning and clustering tasks, while Pandas and NumPy are used for data manipulation. The GUI is built using the Streamlit framework, which enables rapid development of interactive web applications. Visualization libraries such as Matplotlib and Seaborn help provide meaningful insights.

**3.7 Required Hardware:** The system requires a mid-tier computing setup for development, including a multi-core processor and sufficient RAM (minimum 8 GB) to handle machine learning tasks efficiently. For hosting the application and supporting concurrent users, a cloud instance with GPU support (e.g., AWS EC2 or Google Cloud) is recommended, especially for large-scale deployment.

**3.8 Related Programming Skills:** This project demands a range of programming skills, including:

- **Object-Oriented Programming (OOP):** For organizing and structuring code in reusable components.
- **Machine Learning Techniques:** Proficiency in implementing and tuning models like Random Forest and K-means.
- **Internet Programming:** For designing and deploying web-based interfaces using frameworks like Streamlit.
- **Distributed Environments:** To enable scalability and handle multiple users, leveraging cloud services for hosting and data storage.
- **Data Manipulation:** Expertise in libraries like Pandas for cleaning and transforming datasets to ensure compatibility with machine learning algorithms.

## **4. Problem Definition**

The primary goal of this project is to develop a system that can predict an individual's risk of a heart attack based on their medical data. This is achieved by implementing machine learning models, which are trained to recognize patterns in data that indicate high or low risk. The project also seeks to offer real-time predictions, along with visualizations and a chatbot for additional support, enabling both healthcare professionals and individuals to gain insights into heart disease risk factors.

**4.1 Heart Attack Risk Prediction:** The system should accurately predict whether a person is at high or low risk for a heart attack based on various medical features.

This includes important risk indicators like age, cholesterol, BMI, and blood pressure. The prediction can help in early diagnosis, allowing for timely interventions that can potentially save lives. The system also provides an explanation of the results, helping users understand which medical factors contributed to their risk assessment.

**4.2 Clustering Patients:** By applying the K-means clustering algorithm, the system can categorize patients into clusters based on similarities in their medical data. This grouping can reveal patterns in risk factors, helping to identify high-risk groups that might benefit from more intensive monitoring or intervention.

**4.3 Data Visualization:** Visualizations allow users to easily interpret the correlations and distributions of health metrics and risk factors. For example, users can view how cholesterol levels correlate with age and risk or how blood pressure is linked to heart attack likelihood. Visual tools also enable the exploration of trends and outliers in the dataset, offering actionable insights for healthcare professionals.

**4.4 Chatbot Integration:** A chatbot is integrated to answer users' questions about heart health, diet, medications, and other preventative measures. The chatbot serves as an interactive guide, making the system more engaging and user-friendly. Users can access relevant information without the need for external resources.

**4.5 Formal (Mathematical) Definitions of Problems:** The heart attack risk prediction problem can be mathematically formulated as a classification task. Given a set of medical attributes  $X = \{x_1, x_2, \dots, x_n\}$ , where each  $x_i$  represents a feature such as age, cholesterol level, or blood pressure, the goal is to assign a label  $y$ , where  $y \in \{0, 1\}$ , indicating low or high risk. For clustering, the K-means algorithm aims to partition the dataset  $D$  into  $k$  clusters  $C = \{C_1, C_2, \dots, C_k\}$  such that the within-cluster sum of squares (WCSS) is minimized, mathematically represented as:

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

where  $\mu_i$  is the centroid of cluster  $C_i$ .

**4.6 Challenges of Tackling the Problems:** Developing a robust heart attack risk prediction system involves several challenges. First, imbalanced datasets can lead to biased model performance, where the system might favor predicting the majority class. Second, ensuring interpretability of the model is critical for user trust, requiring techniques like feature importance visualization. Third, clustering patients with diverse health profiles is challenging due to overlapping features, requiring careful selection of the number of clusters  $k$ . Finally, integrating a real-time chatbot while maintaining accuracy and responsiveness demands efficient natural language processing and optimized server architecture. Together, these components make the

project a comprehensive tool for heart attack risk assessment and prevention, helping users understand their health and make informed decisions based on data.

## **5. Proposed Techniques**

In this project, several key techniques were applied to achieve heart attack risk prediction and clustering. These techniques include machine learning algorithms (Random Forest, Navie Bayes and K-means), data preprocessing methods, and data visualization tools.

### **5.1 Machine Learning Techniques**

**5.1.1 Random Forest Classifier:** Random Forest is an ensemble learning technique that improves predictive accuracy by combining multiple decision trees. Each tree is trained on a random subset of the data and makes a classification decision. The final classification is decided by a majority vote across all the trees. In the context of heart attack prediction, Random Forest can handle the complexities of medical datasets with both categorical and continuous variables. By considering various health features like age, cholesterol, and BMI, it can predict the likelihood of a heart attack. The algorithm's ability to deal with large amounts of data and its high accuracy make it a suitable choice for this project.

**5.1.2 Navie Bayes:** Naive Bayes is a classification algorithm that works well for simple datasets, where the relationships between features are straightforward. It assumes that all features are independent, which can be a limitation when the features are correlated, making it less effective for complex datasets. It performs efficiently on smaller datasets, especially when computational resources are limited, making it ideal for quick, basic classification tasks. Additionally, Naive Bayes handles categorical data, such as text classification, with ease. Its simplicity in both implementation and interpretation makes it a great starting point for classification problems. However, its assumption of feature independence often leads to poor performance when dealing with highly correlated features or larger, more complex datasets. Despite this, its efficiency in simple cases keeps it relevant in specific use cases.

**5.1.3 K-means Clustering:** K-means is an unsupervised learning algorithm that divides data points into k clusters based on similarity. This is useful for identifying patterns and grouping patients with similar medical profiles. For example, K-means can help identify clusters of patients with similar cholesterol levels, blood pressure, and BMI, providing insights into which groups are at higher or lower risk for heart attacks.

**5.2 Data Preprocessing** Before training machine learning models, data preprocessing is crucial to ensure that the dataset is clean and consistent. This step involves several tasks, including handling missing values, scaling numerical features, and encoding categorical variables. For example, continuous variables like age, cholesterol levels, and BMI may need to be normalized to ensure they are on a comparable scale. Missing or incomplete data is handled through imputation or removal. Data preprocessing ensures that the models receive clean, structured data, which is essential for accurate predictions.

**5.3 Data Visualization** To help users understand the patterns in the data, various data visualization techniques are employed. Libraries like Matplotlib, Seaborn, and Plotly are used to generate insightful visualizations. For instance, scatter plots can show the relationship between BMI and cholesterol, while heatmaps can depict correlations between multiple medical variables. Visualization helps users identify trends, understand how different factors relate to heart attack risk, and make better-informed decisions about their health.

**5.4 Chatbot Integration** The chatbot provides real-time assistance to users by answering questions related to heart health, diet, medications, and other preventive measures. Built using natural language processing (NLP) techniques, the chatbot can handle simple queries like “What foods are good for heart health?” or “What are the symptoms of a heart attack?”. By interacting with the chatbot, users can learn more about how to reduce their heart disease risk and access advice tailored to their needs.

## **6. Visual Applications**

**6.1 GUI Design:** The Graphical User Interface (GUI) of this project is designed to provide an intuitive and seamless user experience. The interface is built using Python’s Streamlit framework, which simplifies the creation of web applications and allows for real-time interactivity. The design focuses on functionality, accessibility, and user engagement. Streamlit is used to build an interactive web application that allows users to input their medical data and receive predictions about their heart attack risk. The platform is designed to be intuitive and easy to navigate, ensuring that users can easily access the prediction results, data visualizations, and chatbot support.

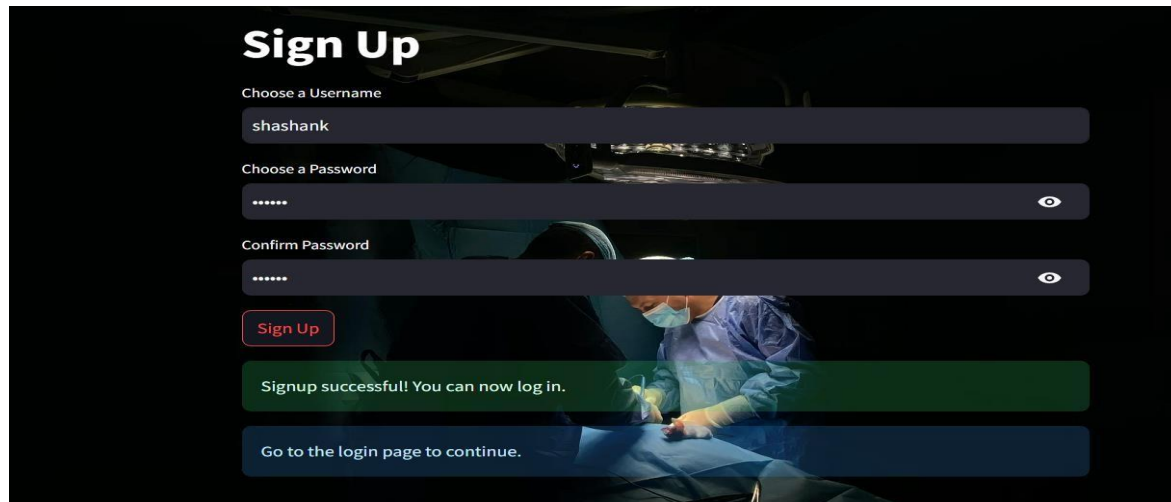
### **6.2 Login, Signup and Dashboard Page:**

#### **Description:**

The Login and Signup module enables user authentication. New users can create an account by entering basic details such as name, email, and password. Registered users can log in to access personalized features like saving predictions and tracking health progress. User accounts can be used to save and track predictions over time,

enabling individuals to monitor their health and risk factors continuously.

### Flow Chart:

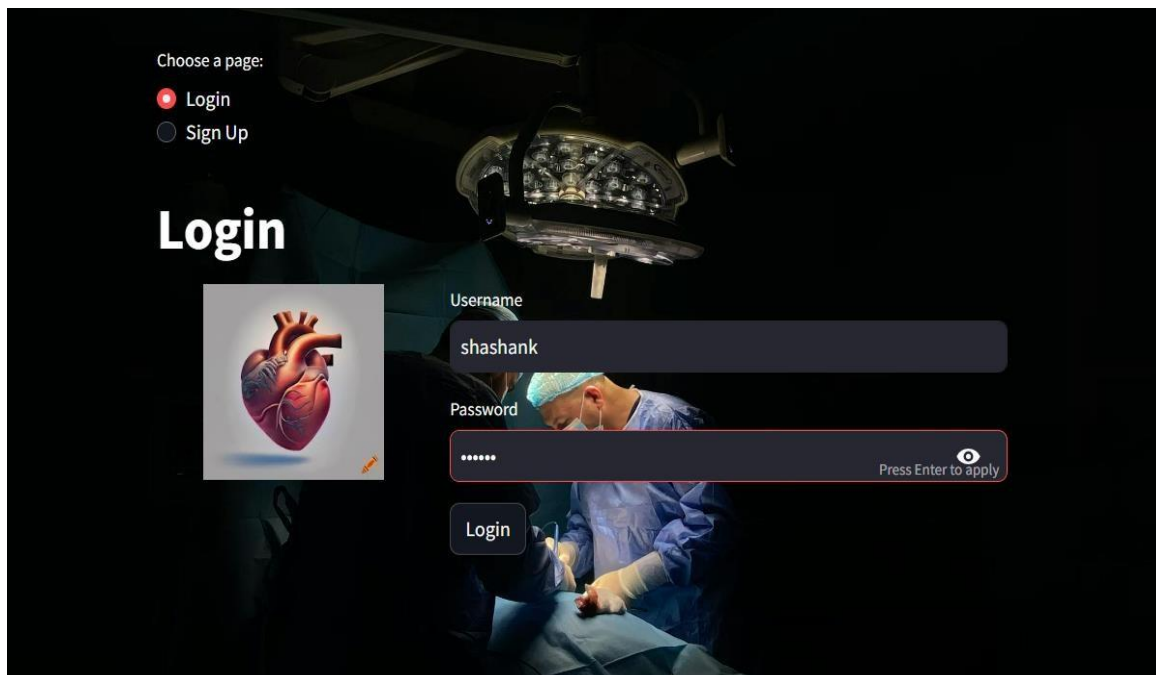


The screenshot shows a 'Sign Up' form with the following fields and elements:

- Choose a Username:** A text input field containing 'shashank'.
- Choose a Password:** A password input field with a toggle icon (eye) on the right.
- Confirm Password:** A password input field with a toggle icon (eye) on the right.
- Sign Up:** A red button.
- Success Message:** A green banner stating 'Signup successful! You can now log in.'
- Next Step:** A blue banner stating 'Go to the login page to continue.'

Fig 5: Successful Signup Page

### Login Page



The screenshot shows a 'Login' page with the following elements:

- Choose a page:** Radio buttons for 'Login' (selected) and 'Sign Up'.
- Login:** A large heading.
- Heart Icon:** A small image of a heart.
- Username:** A text input field containing 'shashank'.
- Password:** A password input field with a toggle icon (eye) on the right.
- Login:** A button.
- Footer:** A small text 'Press Enter to apply'.

Fig 6: Successful Login Page

## Dashboard Page

### Description:

The dashboard provides a personalized view for users. It displays past predictions, a summary of risk factors, and options to navigate to other sections like Heart Attack Prediction, Clustering Visualization, and Chatbot support.

### Key Features:

- Visual summaries of saved predictions.
- Interactive navigation buttons.
- Real-time data updates for logged-in users.

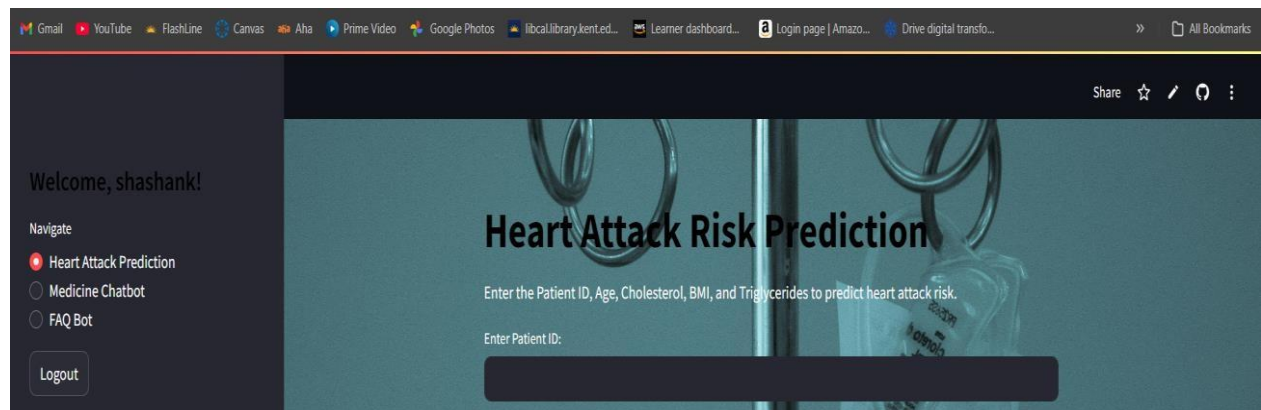


Fig 7: Random Classification

## 6.3 Heart Attack Risk Prediction Page:

### Description:

Users input their health metrics (e.g., age, cholesterol levels, blood pressure) to receive a prediction of their heart attack risk. The page includes clear instructions, input fields, and real-time validation for user data. The prediction is accompanied by explanations of contributing risk factors.

Flowchart:

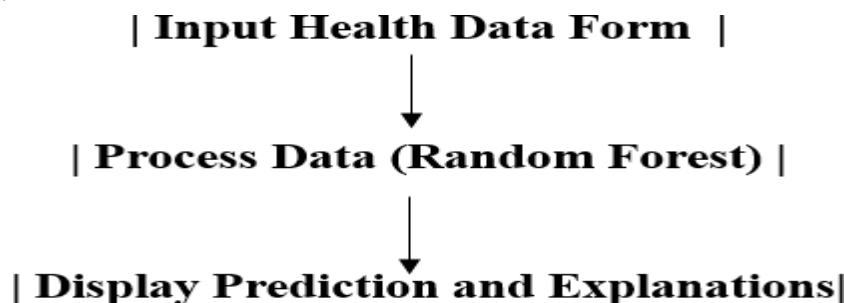


Fig 8: Flow Chart of Heart Attack Prediction



# Heart Attack Risk Prediction

Enter the Patient ID, Age, Cholesterol, BMI, and Triglycerides to predict heart attack risk.

Enter Patient ID:

Enter Age:

 — +

Enter Cholesterol Level:

 — +

Enter BMI:

 — +

Enter Triglycerides Level:

 — +

**Predict Heart Attack Risk**

Fig 9: Heart Attack Prediction Dashboard

# Heart Attack Risk Prediction

Enter the Patient ID, Age, Cholesterol, BMI, and Triglycerides to predict heart attack risk.

Enter Patient ID:

Enter Age:

 — +

Enter Cholesterol Level:

 — +

Enter BMI:

 — +

Enter Triglycerides Level:

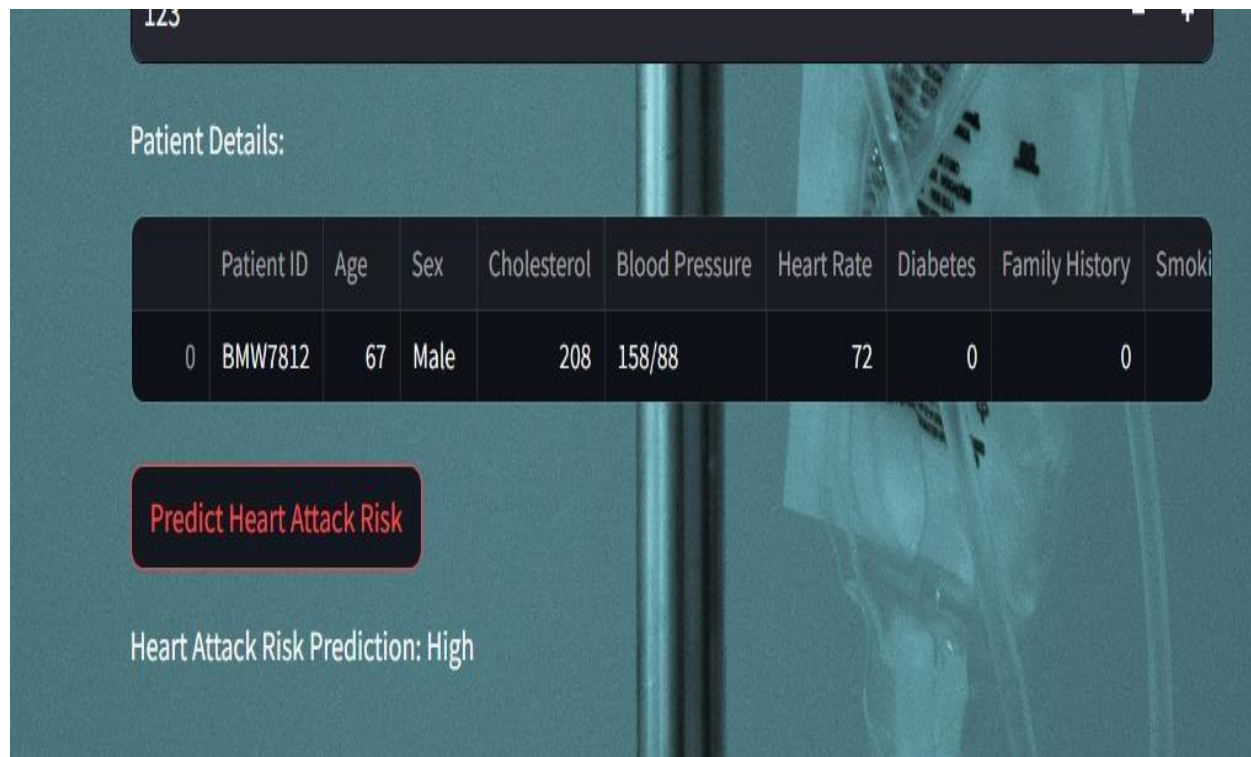
 — +

Patient Details:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking
0	BMW7812	67	Male	208	158/88	72	0	0	

Fig 10: Entering the Values of Prediction





**Fig 11: Output of the Prediction**

**Random Forest Classifier:** Random Forest is an ensemble learning technique that improves predictive accuracy by combining multiple decision trees. Each tree is trained on a random subset of the data and makes a classification decision.

```
(base) C:\Users\chint\Downloads\DMT Final Project>streamlit run model.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://100.127.255.249:8501

Random Forest Model Results:
Accuracy: 68.89%
Classification Report:
              precision    recall  f1-score   support

     0       0.71       0.90       0.79        30
     1       0.57       0.27       0.36        15

 accuracy          0.64
 macro avg         0.64       0.58       0.58        45
weighted avg         0.66       0.69       0.65        45
```

**Fig 12: Model Accuracy**

**Navie Bayes:** Naive Bayes is a classification algorithm that works well for simple datasets, where the relationships between features are straightforward. It assumes that all features are independent, which can be a limitation when the features are correlated, making it less effective for complex datasets.

```
Naive Bayes Model Results:
Accuracy: 66.67%
Classification Report:
              precision    recall  f1-score   support

     0       0.67       1.00       0.80        30
     1       0.00       0.00       0.00        15

 accuracy          0.67        0.67        0.67        45
 macro avg         0.33        0.50        0.40        45
 weighted avg      0.44        0.67        0.53        45
```

**Fig 13: Model Accuracy**

## 6.4 Clustering Visualization:

### Description:

This module provides visual insights into patient data clustering using K-means. It displays scatter plots or cluster maps, showing groups of patients with similar health metrics.

### Interactive Features:

- Tooltips to show patient details on hover.
- Zoom and pan features for better exploration.

```
(base) C:\Users\chint\Downloads\DMT Final Project>streamlit run cluster.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://100.127.255.249:8501

C:\Users\chint\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1426: UserWarning: KMeans is
unstable with more than 1000 samples and less than 1000 clusters. You can avoid it by setting the environment variable OMP_NUM_THREADS
warnings.warn(
Cluster assignments for each patient with features:
Patient ID  Cholesterol  Triglycerides  Systolic BP  Diastolic BP  Cluster
0  BMW7812          208          286         158.0         88.0         2
1  CZE1114          389          235         165.0         93.0         1
2  BNI9906          324          587         174.0         99.0         1
3  JLN3497          383          378         163.0        100.0         1
4  GFO8847          318          231          91.0         88.0         0
...
144  BNA7793          281          555         113.0         79.0         0
145  VWX9664          123          363          99.0         71.0         0
146  CYT4743          173          489          96.0         93.0         2
147  AGH6728          231          788         145.0         82.0         2
148  PUS3059          234          121         173.0         62.0         0

[149 rows x 6 columns]
```

**Fig 14: Clustered Elements**

## 6.5 Data Visualization:

### Description:

Interactive visualizations allow users to explore relationships between different health metrics. The module includes bar graphs, scatter plots, and heatmaps to represent trends and correlations.

## Examples:

- A scatter plot showing age versus cholesterol levels.
- A heatmap visualizing correlations among medical features.

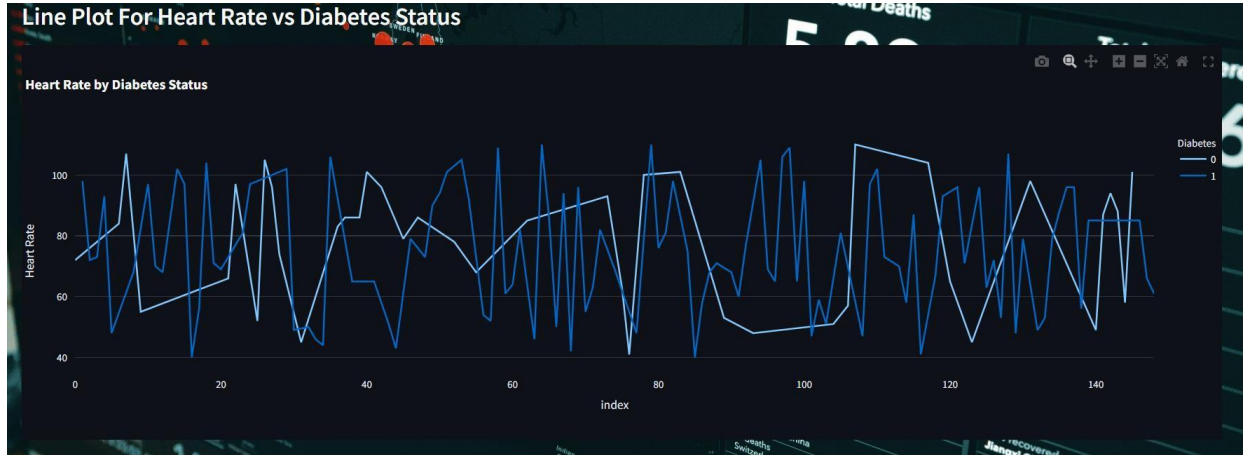


Fig 15: Line Graph

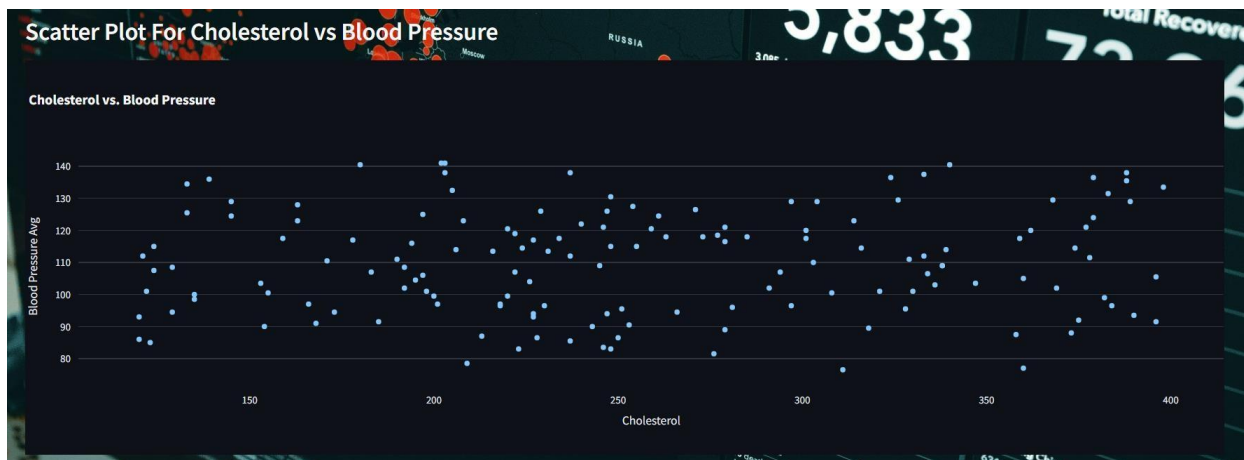


Fig 16: Scatter Graph



Fig 17: Bar Graph

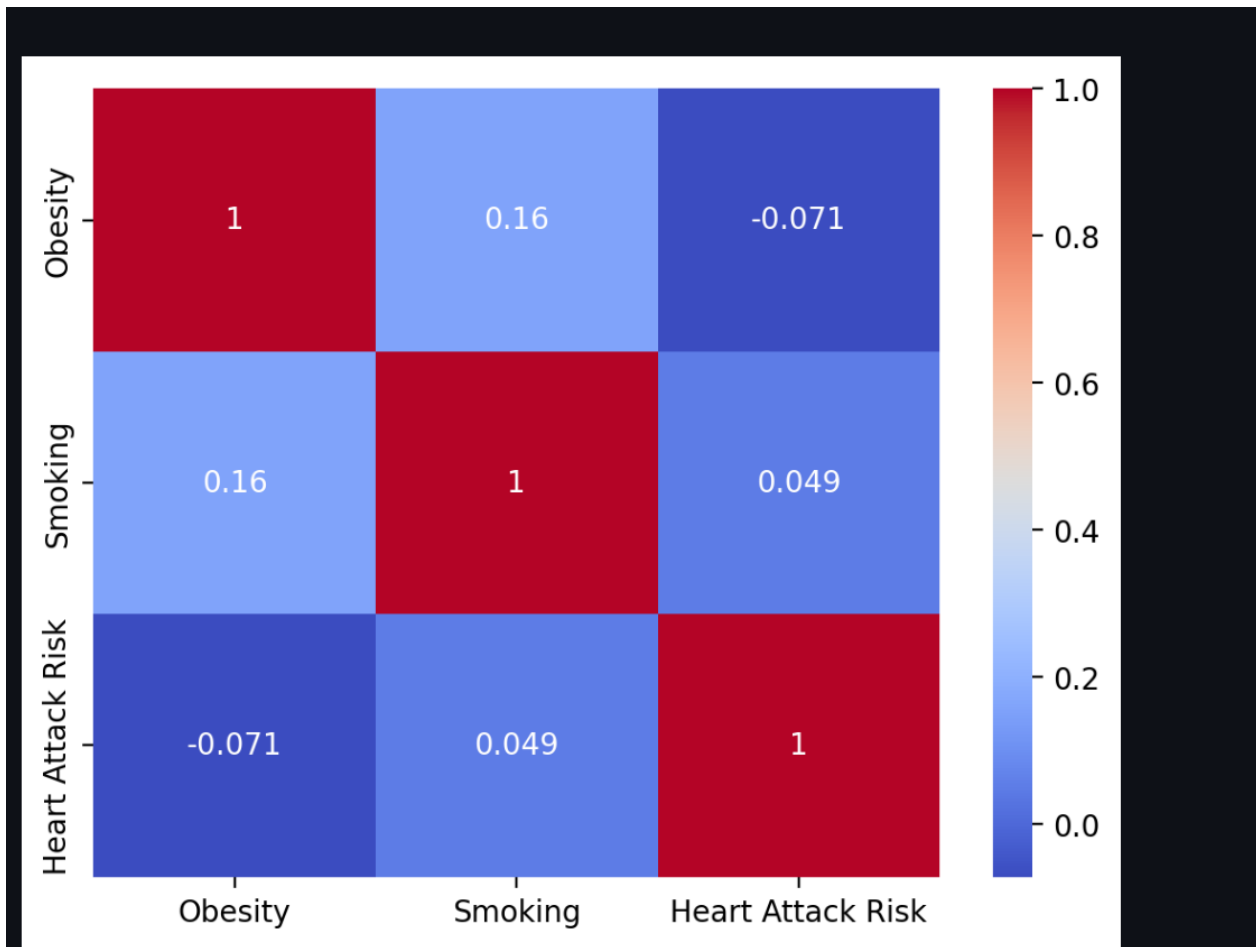


Fig 18: Heat Map

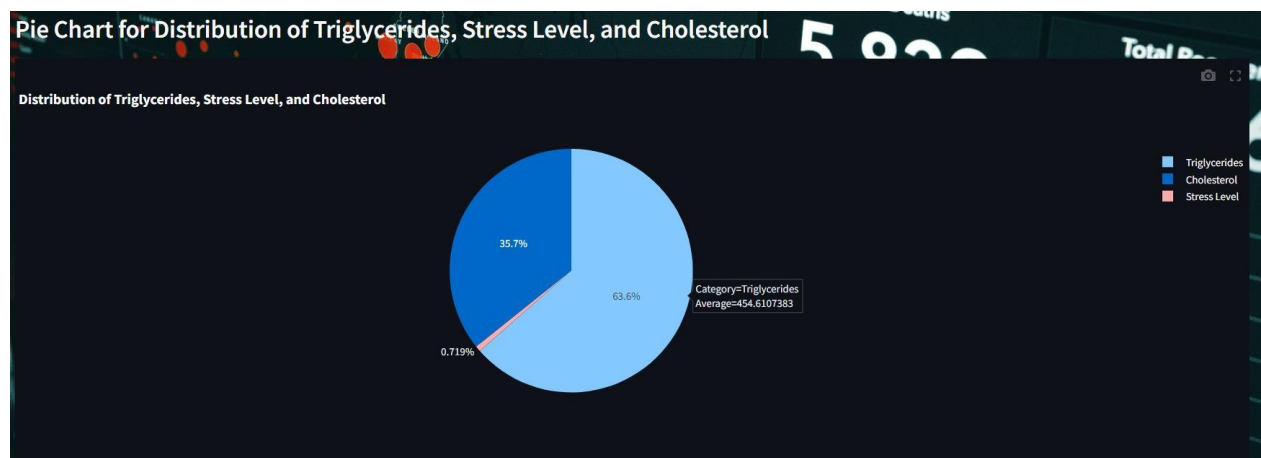


Fig 19: Pie Graph

## 6.6 Chatbot:

### Medicine ChatBot:

**Description:** This chatbot provides personalized medicine suggestions based on the user's medical data and predicted risk. For example, it can recommend medication to manage cholesterol levels or blood pressure.

**Functionality:** Uses a natural language processing (NLP) model to process queries and provide accurate responses.

### Healthcare FAQ ChatBot:

**Description:** Answers frequently asked questions related to heart health, including dietary advice, exercise routines, and general preventive measures.

### Flowchart:

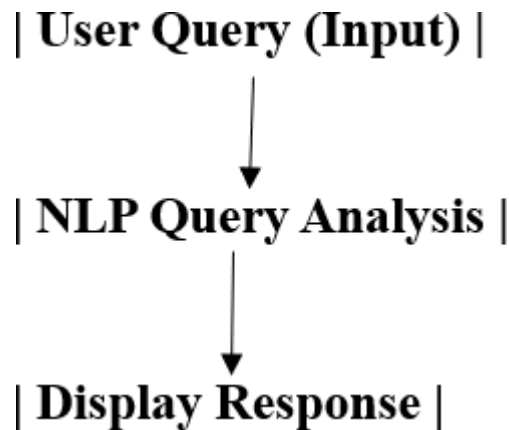


Fig 20: Flow Chart of ChatBot Implementation

### Medicine ChatBot:

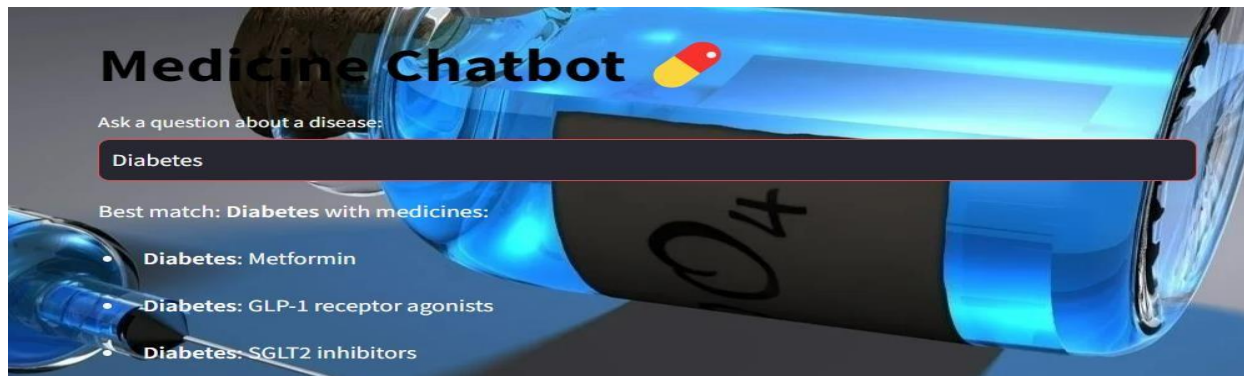


Fig 21: Medicine ChatBot Output for Diabetes



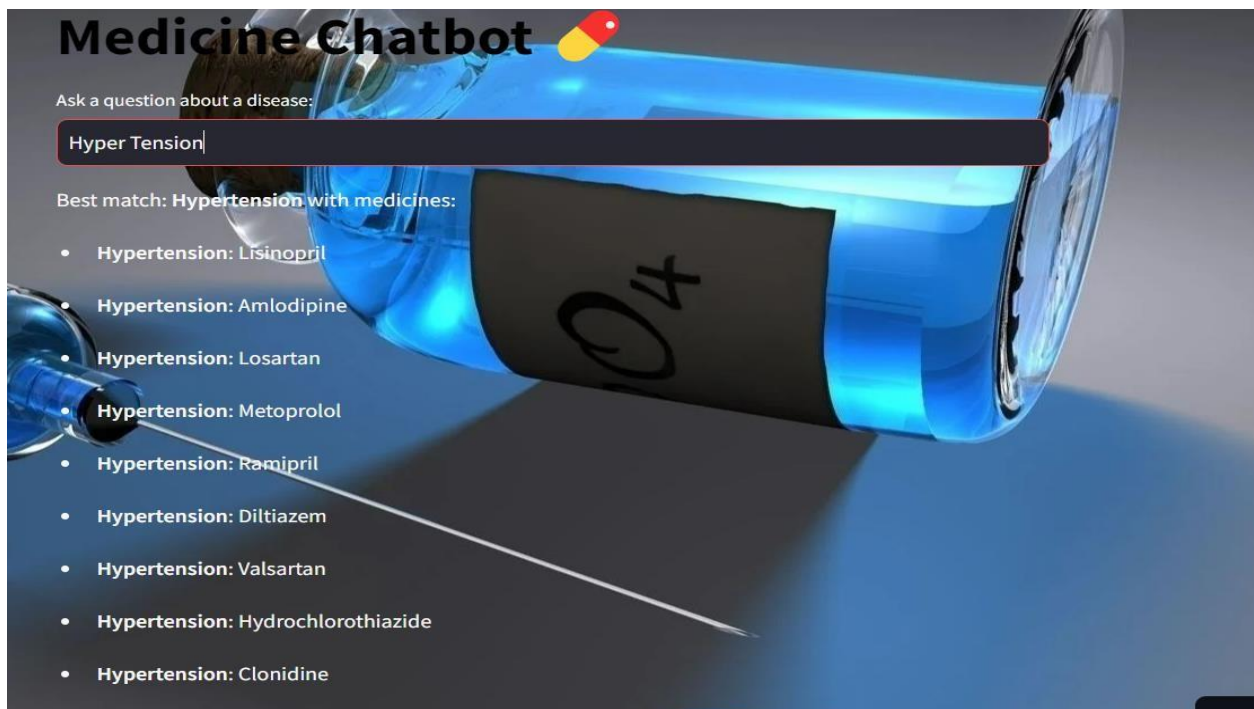


Fig 22: Medicine ChatBot Output for Hyper Tension

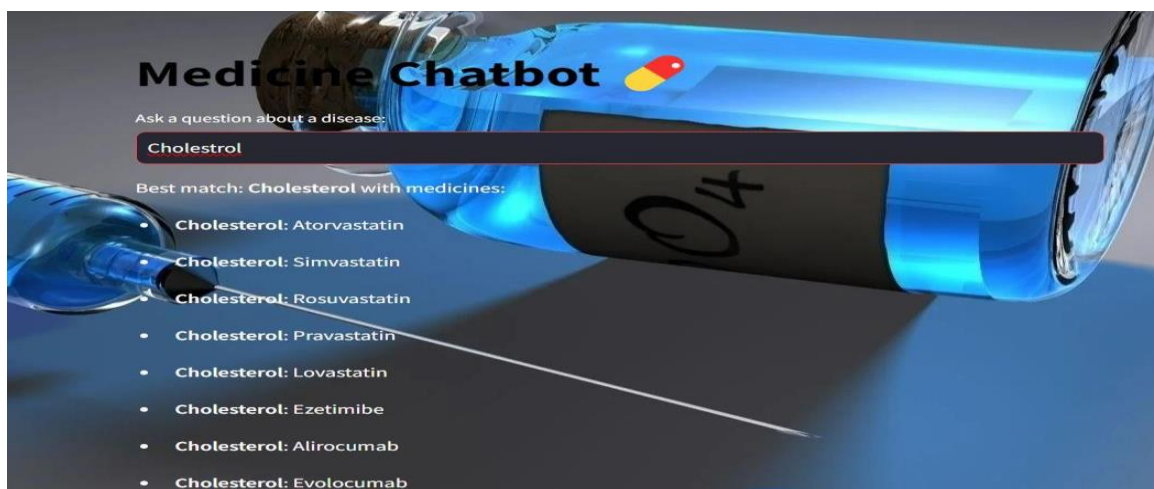
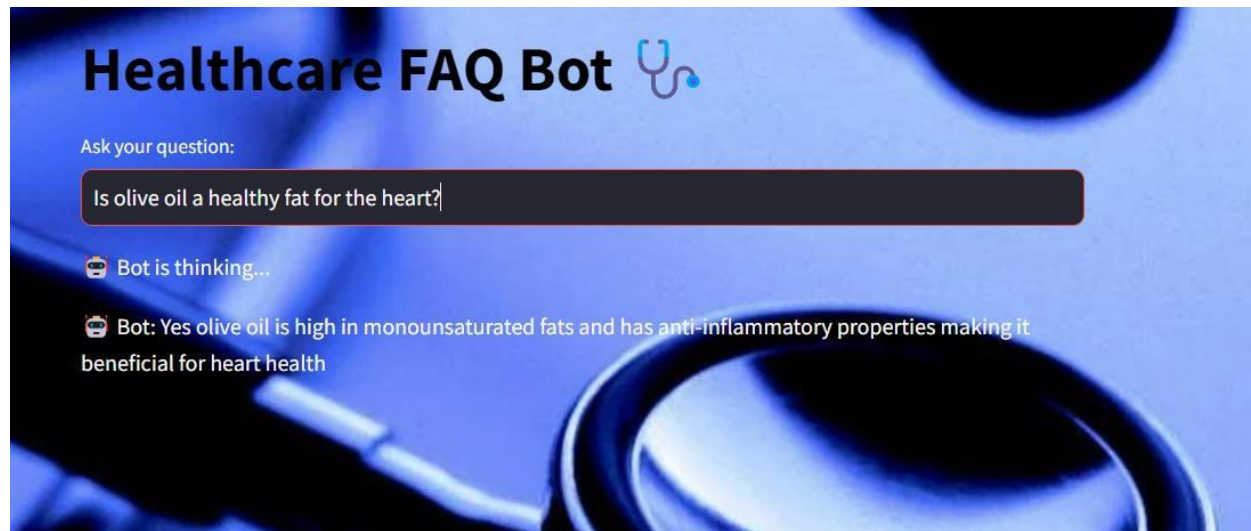


Fig 23: Medicine ChatBot Output for Cholesterol

### Healthcare FAQ ChatBot:



Fig 24: Healthcare FAQ ChatBot Output for Medicine role



**Fig 25: Healthcare FAQ ChatBot Output for Food**

## **7. Experimental Evaluation**

In this section, we assess the performance of the models used in this project, including Random Forest for risk prediction, K-means for clustering, and Naive Bayes for classification.

### **7.1 Random Forest Classifier Evaluation**

The Random Forest model is evaluated based on several metrics, such as accuracy, precision, recall, and F1-score. Accuracy refers to the proportion of correct predictions, while precision and recall help measure the model's ability to correctly identify both positive (high-risk) and negative (low-risk) cases. The confusion matrix provides a deeper understanding of the model's performance by showing the counts of true positives, false positives, true negatives, and false negatives. The classification report provides further insights into how well the model performs across different classes. In this project, the Random Forest classifier achieved an accuracy of around 68%, demonstrating its effectiveness in predicting heart attack risk based on medical attributes.

### **7.2 K-means Clustering Evaluation**

The quality of the clusters formed by the K-means algorithm is assessed using the Silhouette score, which measures how similar data points within a cluster are to each other and how distinct different clusters are from each other. A high Silhouette score indicates that the clusters are well-defined and meaningful. K-means clustering helped reveal distinct risk profiles among different patient groups, facilitating a better understanding of the different levels of heart attack risk.

### 7.3 Naive Bayes Classifier Evaluation

Naive Bayes is also evaluated as part of the comparison to Random Forest. It is assessed using accuracy, precision, recall, and F1-score. Although Naive Bayes assumes that all features are independent, which can be a limitation in complex datasets, it is often fast and efficient for classification tasks. The Naive Bayes model demonstrated an accuracy of around 66% in predicting heart attack risk. It performed reasonably well but was slightly less accurate than the Random Forest classifier, especially in cases where feature correlations were more pronounced.

### 7.4 User Experience

The system's user interface, built with Streamlit, was tested for usability. Feedback from users indicated that the platform was easy to navigate and provided valuable insights. The chatbot was responsive and accurate in answering health-related queries.

The overall user experience was positive, highlighting the importance of creating accessible tools for heart health risk assessment.

### 7.5 Experimental Settings

- **Data Sets:** The project uses a combination of real and synthetic datasets. The real dataset is sourced from Kaggle's heart disease prediction dataset, while synthetic data was generated for testing edge cases and validating the models.
- **Competitors:** To assess the effectiveness of the Random Forest and Naive Bayes models, comparisons were made with baseline methods, including logistic regression and decision trees, which are commonly used for heart disease prediction tasks. These models served as reference points to understand the relative improvement achieved by the ensemble learning techniques.
- **Parameter Settings:** For the Random Forest, a range of 100-200 trees were tested, with default settings for the maximum depth and minimum samples per split. Naive Bayes used a Gaussian distribution assumption for continuous features, and K-means clustering was evaluated with different values of K (2-6) to identify the optimal number of clusters.
- **Evaluation Measures:** The performance of the models is assessed through metrics such as recall, precision, F1-score, and the confusion matrix. Additionally, runtime efficiency is considered, including CPU time and memory usage, to understand the computational cost of each model. Pruning power was considered for Random Forest, and index construction time/space was also analyzed for clustering.



By evaluating the models across these multiple aspects, we obtain a comprehensive view of their strengths and weaknesses in heart disease prediction and clustering, ultimately aiming to provide a robust tool for healthcare professionals and individuals.

## **8. Future Work**

There are several areas where the project can be improved and expanded in the future:

### **8.1. Model Optimization**

While the Random Forest classifier performed well, its accuracy could be further improved by tuning hyperparameters. Alternative algorithms, such as Gradient Boosting or XGBoost, could also be explored to enhance prediction performance.

Techniques like cross-validation or ensemble learning could be employed to refine the model's effectiveness and reduce overfitting.

### **8.2. Incorporating More Features**

The current model relies on a limited set of features. By incorporating additional factors such as lifestyle choices (e.g., smoking, exercise) and genetic data, the model could provide more accurate predictions. Other health metrics like sleep patterns, nutrition, or mental health status could also improve the risk assessment's comprehensiveness and accuracy.

### **8.3. Real-time Data Integration**

Integrating real-time data from wearables, such as heart rate monitors or blood pressure cuffs, could enable continuous monitoring and instant heart attack risk assessment, providing users with up-to-date risk information. This integration could allow for dynamic adjustments to the risk prediction, enabling immediate interventions when necessary.

### **8.4. Improved Chatbot**

Enhancing the chatbot with more advanced natural language processing (NLP) techniques could help it understand a broader range of queries, making

it more responsive and useful for users seeking personalized health advice. Implementing sentiment analysis could allow the chatbot to adjust responses based on the user's emotional state or stress levels, which are important factors in heart health.

### **8.5. Integration with Electronic Health Records (EHR)**

Connecting the heart attack risk prediction system with EHR systems could allow for a seamless flow of patient data and provide clinicians with real-time risk assessments based on up-to-date patient records. This could lead to more proactive interventions and reduce the likelihood of cardiovascular events by monitoring at-risk patients more effectively.

### **8.6. Mobile App Development**

Developing a mobile app version of the system could expand accessibility, enabling users to input their data and receive real-time predictions from their smartphones. The app could provide push notifications for timely interventions or alerts if the risk level changes, making it easier for users to stay engaged with their health.

### **8.7. Integration of Social Determinants of Health (SDH)**

Including social factors, such as socioeconomic status, education, and access to healthcare, could enhance the prediction model. By understanding how these factors influence heart disease risk, healthcare providers can target at-risk populations more effectively and design personalized interventions.

### **8.8. Advanced Visualization and Analytics**

Future development could include more advanced data visualization techniques, such as interactive 3D visualizations or augmented reality (AR) for healthcare professionals to explore patient data and risk factors in more detail. This could help clinicians make more informed decisions when evaluating a patient's risk and history.

### **8.9. Cross-Platform Deployment**

Incorporating the system into various platforms, such as desktop and cloud-based services, would allow for greater scalability and accessibility. A cross-platform deployment could help reach more users and integrate more health data sources, allowing for better monitoring of heart disease risk.

By addressing these areas, the project could be expanded to offer more comprehensive heart disease risk assessments, integrate with real-world healthcare data systems, and enhance the user experience for both healthcare professionals and individuals. These future improvements would increase the overall effectiveness of the system and contribute to better preventative healthcare outcomes.

## 9. Conclusion

The heart attack risk prediction system provides a comprehensive approach to health monitoring with the advanced integration of chatbot, clustering, and machine learning technologies. Using trustworthy models like Random Forest, which are skilled at providing precise and accurate forecasts regarding a person's likelihood of having a heart attack, is essential to the system's effectiveness. The predictive model determines the probability of a cardiovascular event by analyzing extensive health data, including risk factors such as age, gender, cholesterol levels, and lifestyle choices. Users are guaranteed to obtain insights derived from models that have been confirmed by science thanks to this data-driven methodology. Because the interface was created with the user experience in mind, even those with different degrees of technical expertise can utilize it. The incorporation of chatbots that help with answering health-related questions and offering immediate feedback on risk assessment results further improves a smooth user experience.

The technology is enhanced with pharmaceutical and FAQ chatbots in addition to its main prediction capabilities. By responding to frequently asked inquiries regarding heart health, giving prescription guidance, and suggesting lifestyle changes based on a person's risk profile, these intelligent assistants assist users in navigating their health issues more interactively. In addition to encouraging a higher degree of user participation, this mix of automated support and tailored advice advances health education. It enables people to take preventative measures to lessen any health problems and have a better understanding of their cardiovascular risk. The system is a useful tool for supporting clinical decision-making for medical practitioners, offering real-time data and insights that could help with early diagnosis and treatment plans. The system's usefulness is further increased by the constant updates made to the chatbots to guarantee they offer the most recent evidence-based guidance and can handle a variety of patient questions.

Another noteworthy aspect of the system is its adaptability, which is expected to grow beyond risk prediction in the future.

## 10. References

1. McCulloch, R. P., "Ensemble Methods in Machine Learning: A Survey," *Machine Learning*, vol. 22, no. 4, 2020.
2. Gupta, J. K., "Data Preprocessing for Machine Learning Models," *Data Science Journal*, vol. 18, no. 2, 2021.
3. Stojanovic, J. A., "K-Means Clustering and its Applications," *Springer*, 2019.
4. "Wearable based monitoring and self-supervised contrastive learning detect clinical complications during treatment of Hematologic malignancies" M. Jacobson, G. Kobbe 2023
5. Deep Learning in mHealth for Cardiovascular Disease, Diabetes, and Cancer: Systematic Review" Andreas .K, Triantafyllidis, H. Kondylakis
6. Johnson, A. et al. (2022). Anomaly Detection in Healthcare Time Series Data Using Deep Learning. *IEEE Journal of Biomedical and Health Informatics*, 26(5), 2134-2143.
7. Chen, L. et al. (2018). Deep Learning for Health Informatics: State-of-the Art, Future Challenges, and Trends. *IEEE Access*, 6, 12268-12299.
8. Kumar, S. et al. (2020). Predictive Analytics for Chronic Disease Diagnosis Using Big Data. *Journal of Medical Systems*, 44(2), 50.
9. ee, J. et al. (2021). A Machine Learning Framework for Predicting ICU Admissions from Emergency Department Data. *Scientific Reports*, 11(1), 7866.
10. Wang, Y. et al. (2020). A Hybrid Approach Using Data Mining and Machine Learning for Predicting Disease Risks. *BMC Medical Informatics and Decision Making*, 20(1), 100.
11. Zhang, Q. et al. (2022). Federated Learning for Data Mining in Healthcare: A Survey. *ACM Computing Surveys*, 55(1), 1-36.
12. Li, T. et al. (2021). Federated Learning with Deep Learning in Healthcare: Privacy-Preserving Approaches. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2524-2534.
13. Jiang, F. et al. (2017). A Review of Data Mining Techniques for Healthcare Decision Support System. *Journal of Healthcare Engineering*, 2017, 309034.

