TABLE OF CONTENTS

Chapter-1: Introduction

1.1 General Introduction

Cardiovascular diseases (CVDs) are a group of disorders that affect the heart and blood vessels, including conditions such as coronary artery disease, heart failure, and stroke. According to the World Health Organization (WHO), CVDs are among the leading causes of death globally, accounting for millions of deaths each year. Early detection and prediction of CVDs are critical for effective prevention and management, as they allow for timely interventions and lifestyle modifications that can reduce the risk of adverse health outcomes.

In this project, we aim to leverage data mining techniques to develop a predictive model for cardiovascular disease. By analyzing various risk factors, including demographic information, lifestyle habits, and medical indicators, we seek to identify patterns and relationships that can help predict the likelihood of developing CVDs. The ultimate goal is to build a reliable model that can assist healthcare professionals in identifying individuals at high risk of CVDs, thereby enabling targeted interventions and preventive measures.

1.2 Problem Statement

The problem statement revolves around predicting the occurrence of cardiovascular disease based on a comprehensive set of factors. These factors encompass demographic characteristics, such as age and gender, lifestyle behaviors, such as smoking and alcohol consumption, and medical indicators, such as blood pressure and cholesterol levels. By analyzing these diverse attributes collectively, we aim to develop a predictive model capable of identifying individuals who are at increased risk of developing cardiovascular disease.

Early detection of CVDs is essential for effective prevention and management. By predicting the likelihood of CVD occurrence, healthcare providers can implement timely interventions, such as lifestyle modifications, medication management, and regular monitoring, to mitigate the risk of adverse cardiovascular events. Therefore, the primary objective of this project is to develop a robust predictive model that can accurately identify individuals at risk of cardiovascular disease, thus enabling proactive healthcare interventions and improving patient outcomes.

Chapter-2: Dataset Description

2.1 SOURCE OF DATASET

The dataset used in this project is publicly available, obtained from the Cardiovascular Disease dataset on Kaggle.

2.2 Data size, Format and Attributes

The dataset contains 70,000 instances in CSV format. It includes various features such as age, gender, height, weight, blood pressure, cholesterol, glucose levels, smoking habits, alcohol consumption, and physical activity.
Attributes:

Age

Type: Numeric (Objective Feature)

Description: Represents the age of the individual in days.

Height

Type: Numeric (Objective Feature)

Description: Represents the height of the individual in centimeters.

Weight

Type: Numeric (Objective Feature)

Description: Represents the weight of the individual in kilograms.

Gender

Type: Categorical (Objective Feature)

Description: Represents the gender of the individual.

Values: 1: Female 2: Male

Systolic Blood Pressure (ap_hi)

Type: Numeric (Examination Feature)

Description: Represents the systolic blood pressure measured during medical examination.

Diastolic Blood Pressure (ap_lo)

Type: Numeric (Examination Feature)

Description: Represents the diastolic blood pressure measured during medical examination.

Cholesterol

Type: Categorical (Examination Feature)

Description: Represents the cholesterol level measured during medical examination.

Values: 1: Normal 2: Above Normal 3: Well Above Normal

Glucose

Type: Categorical (Examination Feature)

Description: Represents the glucose level measured during medical examination.

Values: 1: Normal 2: Above Normal 3: Well Above Normal

Smoking

Type: Categorical (Subjective Feature)

Description: Represents whether the individual is a smoker or not.

Values: 0: Non-Smoker 1: Smoker

Alcohol Intake

Type: Categorical (Subjective Feature)

Description: Represents whether the individual consumes alcohol or not.

Values: 0: Non-Drinker 1: Drinker

Physical Activity

Type: Categorical (Subjective Feature)

Description: Represents the level of physical activity of the individual.

Values: 0: Inactive 1: Active

Presence or Absence of Cardiovascular Disease (cardio)

Type: Categorical (Target Variable)

Description: Represents the presence or absence of cardiovascular disease.

Values:0: Absence of Cardiovascular Disease1: Presence of Cardiovascular Disease

2.3 Preprocessing

The data you uploaded is a sample of a larger dataset containing health information for 70,000 individuals. It includes features such as age, gender, height, weight, blood pressure, cholesterol, glucose, smoking habits, alcohol consumption, and physical activity level. Here's a breakdown of the data and some observations:

Data Types

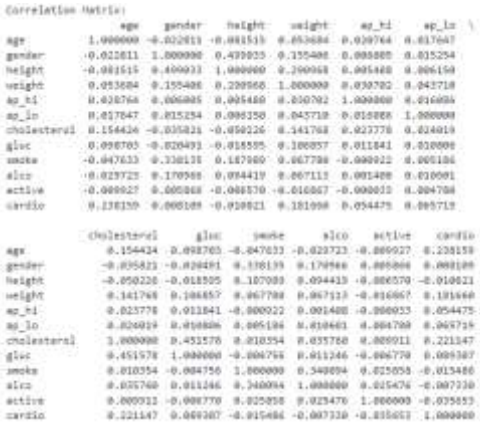
- Numeric: Age (days), height (cm), weight (kg), systolic blood pressure (mmHg), diastolic blood pressure (mmHg).
- Categorical: Gender (coded as 1 for female, 2 for male), cholesterol (coded as 1: normal, 2: above normal, 3: well above normal), glucose (coded as 1: normal, 2: above normal, 3: well above normal), smoking status (coded as 0: non-smoker, 1: smoker), alcohol intake (coded as 0: non-drinker, 1: drinker), physical activity (coded as 0: inactive, 1: active).

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729
std	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179
min	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000
50%	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000
75%	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000
max	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000

- Central Tendency:
  - The average age is approximately 53.5 years old.
  - The average height is 164.4 cm.
  - The average weight is 74.2 kg.
  - The average systolic blood pressure is 128.8 mmHg.
  - The average diastolic blood pressure is 96.6 mmHg.

- Missing Values: The data snippet doesn't reveal any missing values, but it's always a good practice to check for them in the entire dataset.

- Correlations:



Age : Weak positive correlation with cardio (0.238). This suggests that as age increases, the risk of cardiovascular disease also tends to increase. This is likely due to the natural degeneration of the cardiovascular system over time.

Gender: Negligible correlation with cardio (0.008). There is no significant association between gender and cardiovascular disease according to this data.

Height: Very weak negative correlation with cardio (-0.011). This is close to zero and likely not significant. There might be a very slight tendency for shorter people to have a higher risk of cardiovascular disease, but this could be due to chance.

Weight: Weak positive correlation with cardio (0.182). People with higher weight might be at an increased risk of cardiovascular disease. This could be due to factors like increased stress on the heart, inflammation, and metabolic imbalances.

Systolic blood pressure (ap_hi): Weak positive correlation with cardio (0.054). This suggests a possible link between higher systolic blood pressure and increased risk of cardiovascular disease. High blood pressure puts a strain on the heart and blood vessels, leading to damage over time.

Diastolic blood pressure (ap_lo): Weak positive correlation with cardio (0.066). Similar to systolic blood pressure, diastolic blood pressure might also be weakly associated with cardiovascular disease risk.

Cholesterol: Weak positive correlation with cardio (0.221). Higher cholesterol levels might be associated with an increased risk of cardiovascular disease. Cholesterol buildup can lead to narrowing of arteries and restricted blood flow.

Glucose: Very weak positive correlation with cardio (0.089). There might be a possible link between higher glucose levels and increased risk of cardiovascular disease, but the strength of the association is minimal.

Smoking: Very weak negative correlation with cardio (-0.015). This is close to zero and likely not significant. It might suggest a slight protective effect of smoking, but this is counterintuitive and likely due to chance. Smoking is a major risk factor for cardiovascular disease and is unlikely to have a protective effect.

Alcohol intake: Very weak negative correlation with cardio (-0.007). This is close to zero and likely not significant. It might suggest a slight protective effect of alcohol consumption, but this is counterintuitive and likely due to chance. Excessive alcohol consumption is a risk factor for cardiovascular disease.

Physical activity: Very weak negative correlation with cardio (-0.036). This is close to zero and likely not significant. It might suggest a slight protective effect of physical activity, but the strength of the association is minimal. Regular physical activity is well-known to reduce the risk of cardiovascular disease.

Most Relevant Features:

Based on the correlation coefficients, here are the most relevant features for predicting cardiovascular disease in this dataset:

Age: This has a weak positive correlation with cardio, indicating a potential link.

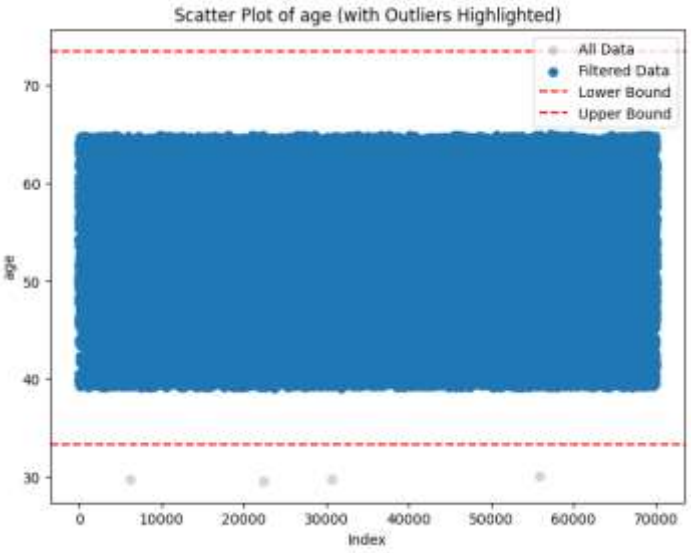
Weight: This shows a weak positive correlation with cardio, suggesting a possible association.

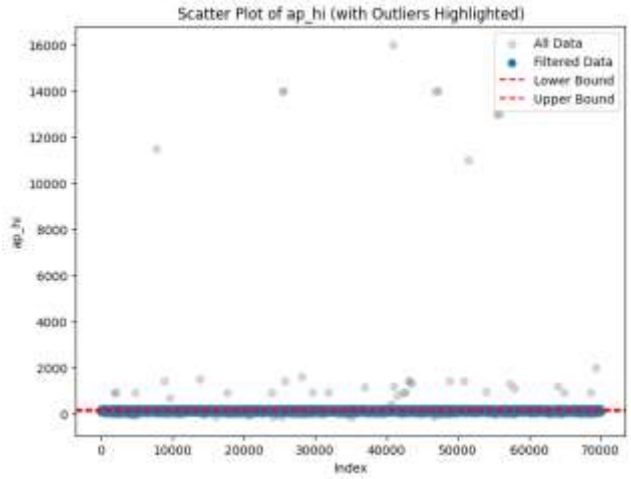
Cholesterol: This has a weak positive correlation with cardio, indicating a potential connection.

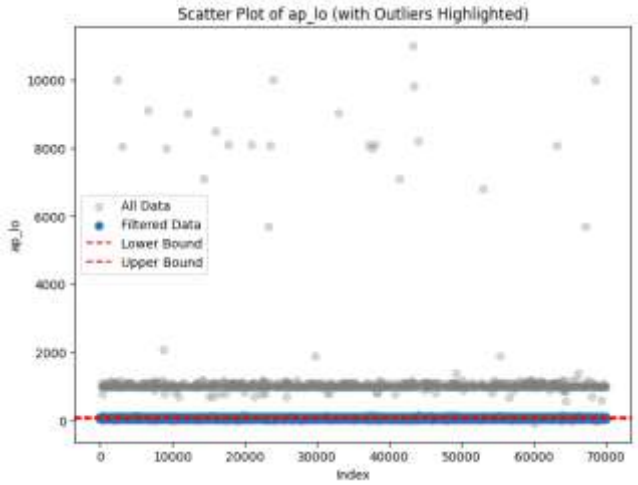
Systolic blood pressure (ap_hi): This shows a weak positive correlation with cardio, suggesting a possible association.

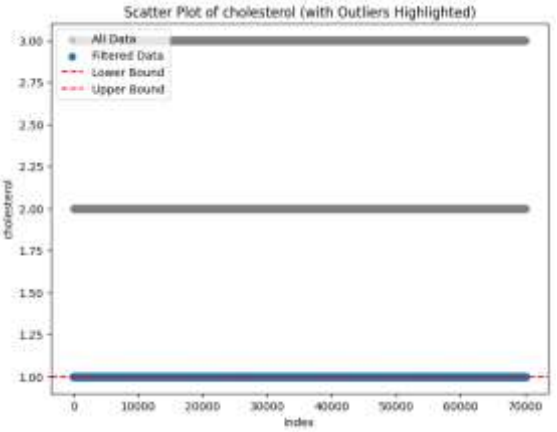
Diastolic blood pressure (ap_lo): This has a weak positive correlation with cardio, indicating a possible link.

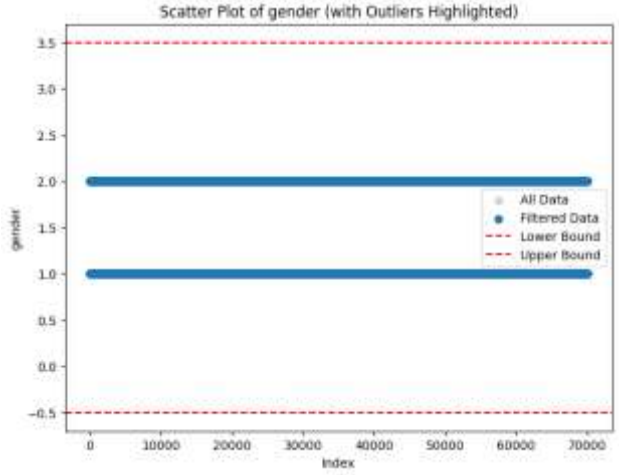
- Outliers

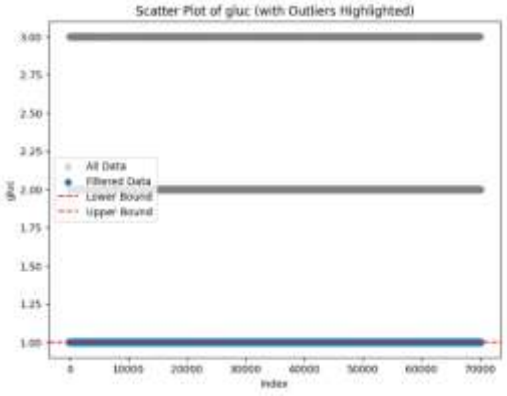


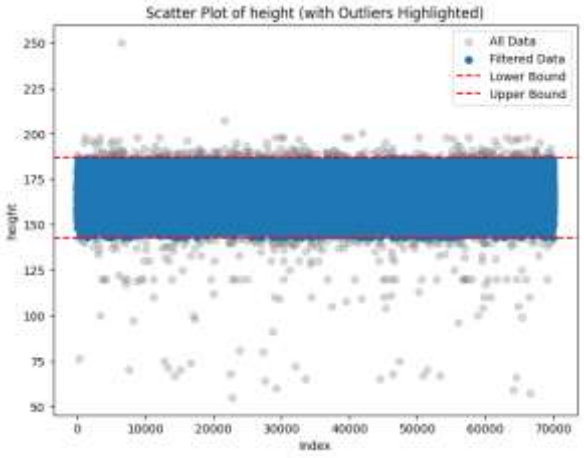


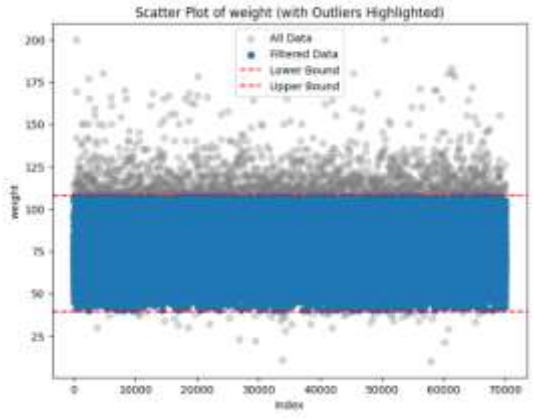

Scatter Plot of ap_lo (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
ap_lo
Index


Scatter Plot of cholesterol (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
cholesterol
Index


Scatter Plot of gender (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
gender
Index


Scatter Plot of gluc (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
gluc
Index


Scatter Plot of height (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
height
Index

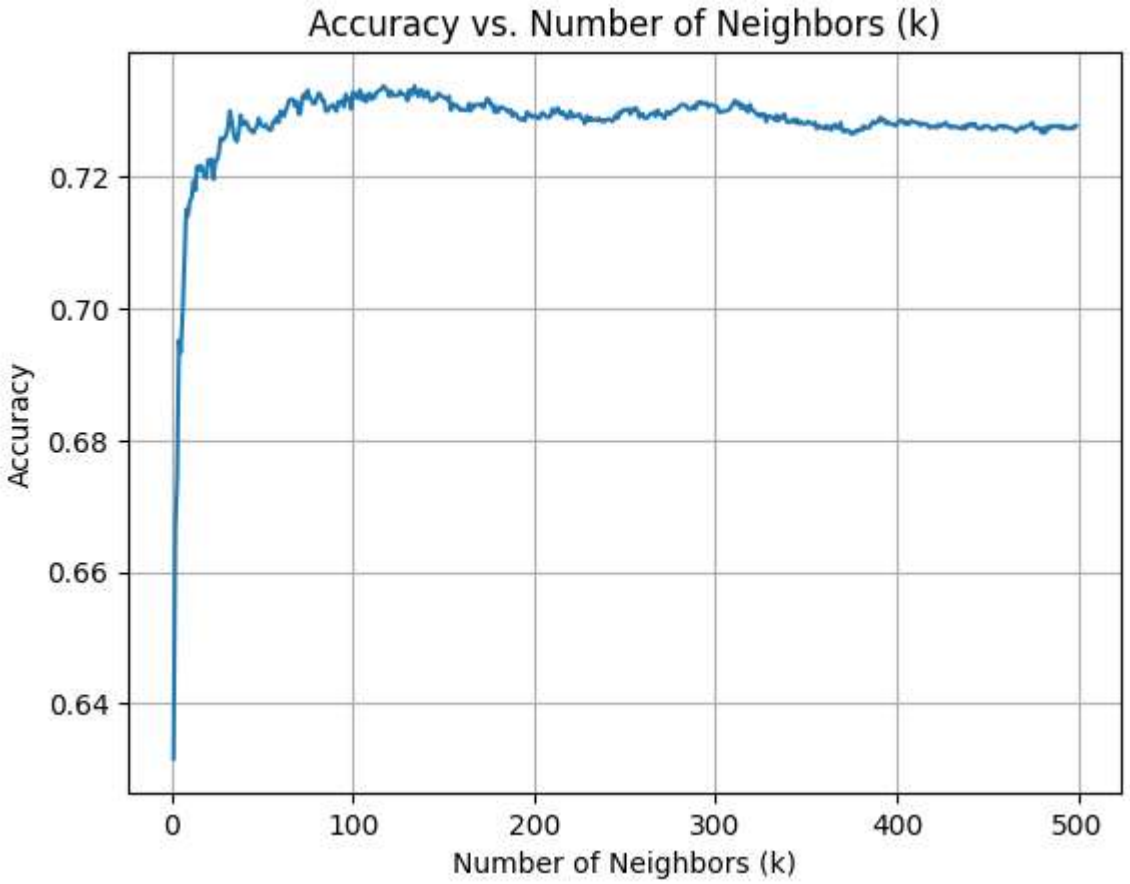

Scatter Plot of weight (with Outliers Highlighted)
All Data
Filtered Data
Lower Bound
Upper Bound
weight
Index

Chapter-3: Model selection , training and evaluation

3.1 K-Nearest Neighbors (KNN) Classifier

The K-Nearest Neighbors (KNN) algorithm is a simple and intuitive method used for classification. It classifies a data point by comparing it with its k nearest neighbors in the feature space. The class label of the majority of the nearest neighbors is assigned to the data point being classified. KNN is based on the assumption that similar data points tend to belong to the same class.

Optimal value of K

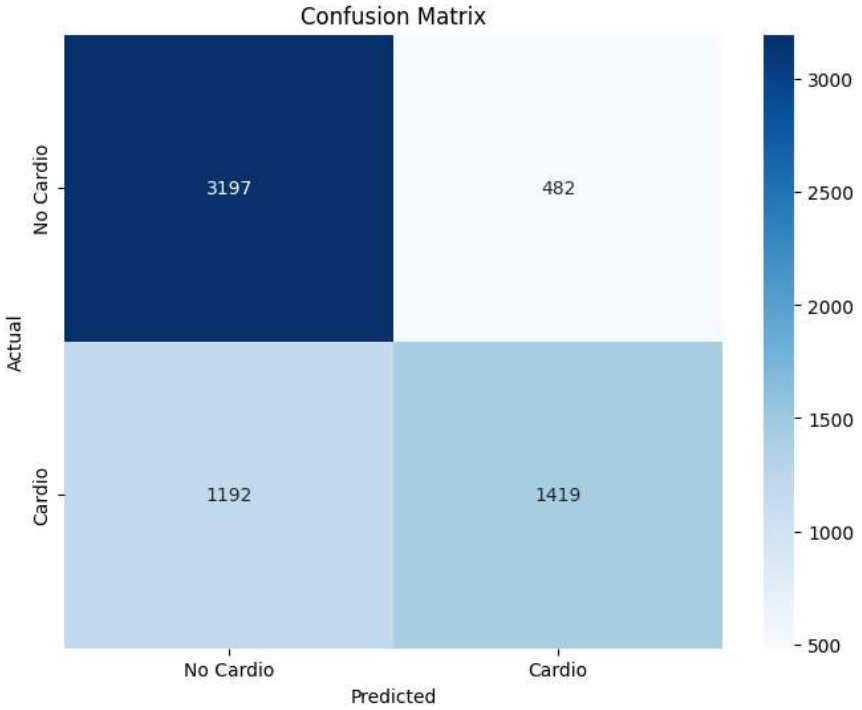


K = 117

Results

```
Confusion Matrix:
[[3197  482]
 [1192 1419]]

Accuracy: 0.7338632750397456
```



Accuracy = (True Positives + True Negatives) / Total Number of Individuals

= 0.73386

Precision = True Positives / (True Positives + False Positives)

= 0.74644

Recall = True Positives / (True Positives + False Negatives)

` **= 0.54346**

This model seems to have a decent accuracy, but the recall is lower than the precision. This suggests that the model might be good at avoiding false positives (predicting someone has cardiovascular disease when they don't) but might miss a significant number of true cases (individuals with the disease who are classified as healthy).

Here are some potential reasons for the lower recall:

> **Class Imbalance:** If the dataset has a higher proportion of healthy individuals compared to those with cardiovascular disease, the model might be biased towards predicting the majority class (healthy) and miss a higher number of disease cases.

Data Quality: Issues with data quality, like missing values or outliers, could affect the model's ability to learn the true relationships between features and the target variable.

Model Complexity: An overly complex model might struggle to generalize well to unseen data, leading to missed cases.

Chapter 4 : Web application

4.1 Technologies Used

Front-end :

HTML,CSSJS,EJS

Back-end :

Node.js, express.js

NPM packages used:

nodemon: Auto-restart server during development.
connect-flash: Store and display flash messages.
connect-mongo: Store session data in MongoDB.
dotenv: Load environment variables from .env file.
ejs: Templating engine for dynamic HTML.
ejs-mate: Layout engine for EJS templates.
express: Web application framework for Node.js.
express-session: Manage user sessions in Express.
fs: Built-in module for file system operations.
method-override: Simulate HTTP methods like PUT and DELETE.
mongodb: Official MongoDB driver for Node.js.
mongoose: ODM library for MongoDB and Node.js.
passport: Authentication middleware for Node.js.
passport-local: Passport strategy for local authentication.
passport-local-mongoose: Mongoose plugin for local authentication.

Deployment:

Docker, aws(ec2)

4.2 Website Functionality

Prediction:

Description: Users are provided with a form where they can input various cardiovascular health parameters such as age, gender, blood pressure, cholesterol levels, and other relevant factors. Upon submission, the backend processes this data using the K-Nearest Neighbors (KNN) algorithm to predict the potential risk of cardiovascular diseases for the user.

Implementation:

User Input Form: The frontend presents a user-friendly form using HTML and EJS templates, allowing users to input their health data.

Data Processing: Upon form submission, the backend controller receives the user input, validates it, and passes it to the KNN algorithm for prediction.

Prediction Outcome: The prediction outcome is then returned to the user interface, where it is displayed along with an explanation of the risk level.

Registration and Login:

Description: Users can create an account by registering with their email address and password. Once registered, they can log in securely to access personalized features and manage their health data.

Implementation:

Registration Form: The frontend provides a registration form for users to create an account. It validates user inputs such as email format and password strength.

Authentication Middleware: Passport.js middleware handles user registration, login, and session management securely.

Password Encryption: User passwords are encrypted using hashing algorithms before storing them in the database to enhance security.

4.3 Project structures

Frontend:

HTML, CSS, JavaScript (with EJS templating): These technologies are used for creating the user interface and adding interactivity to the website.

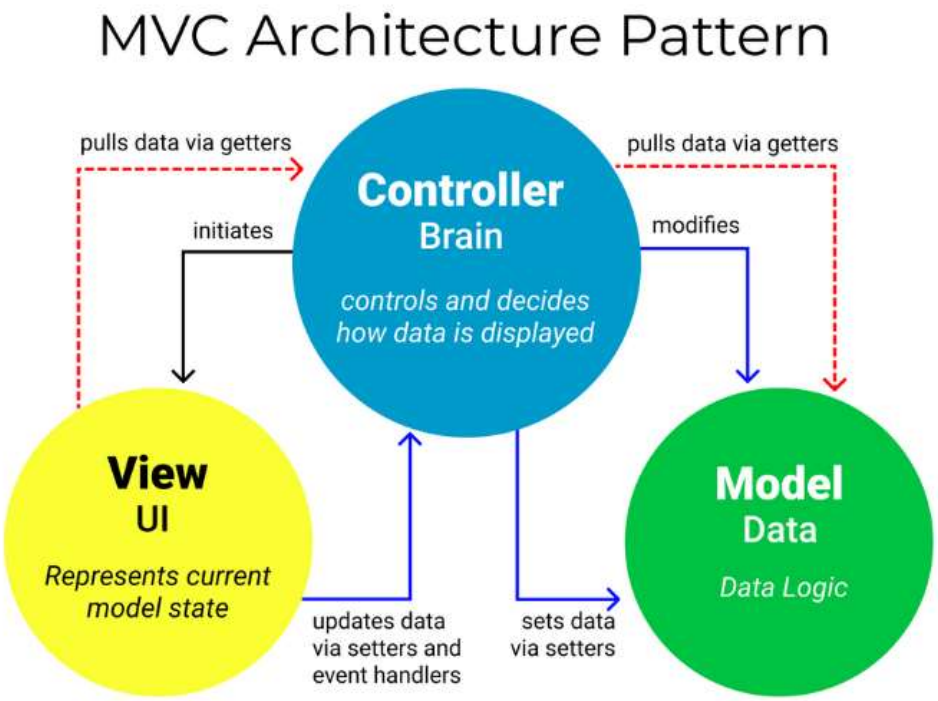
EJS (Embedded JavaScript): EJS templating engine is used for generating dynamic HTML content based on data from the backend.

Backend:

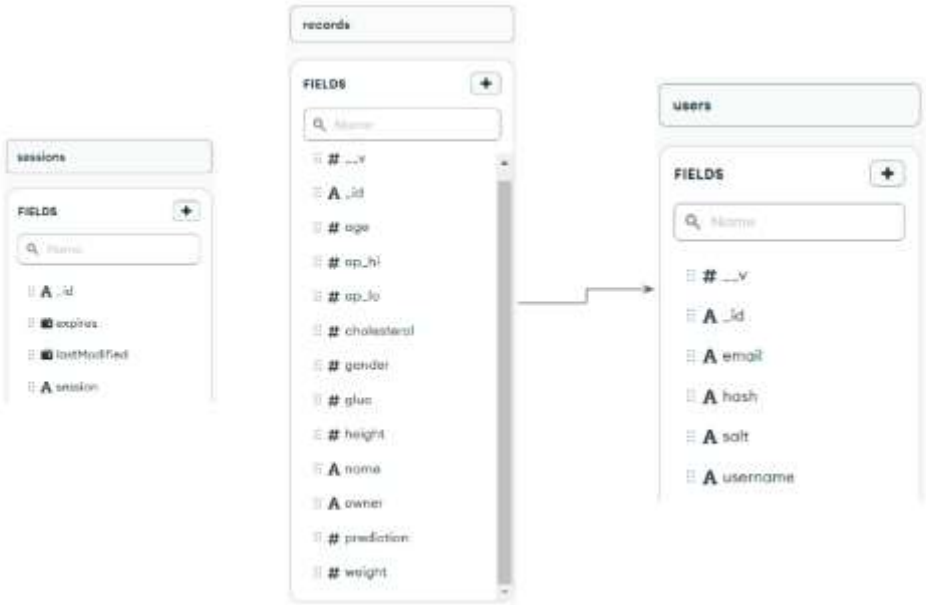
Node.js: Node.js is the server-side JavaScript runtime used to build the backend of the application.

Express.js: Express.js is a web application framework for Node.js, facilitating the creation of APIs and handling HTTP requests.

MVC Model: The application follows the Model-View-Controller architectural pattern for organizing code into separate components responsible for data, presentation, and business logic.



Models (Mongoose): Define Mongoose schemas to structure and validate data stored in MongoDB. Utilize schema types for defining field types, required fields, default values, and validation rules.



Views (EJS): Configure Express to render EJS templates using the render() method, passing data from controllers to templates for dynamic content generation.

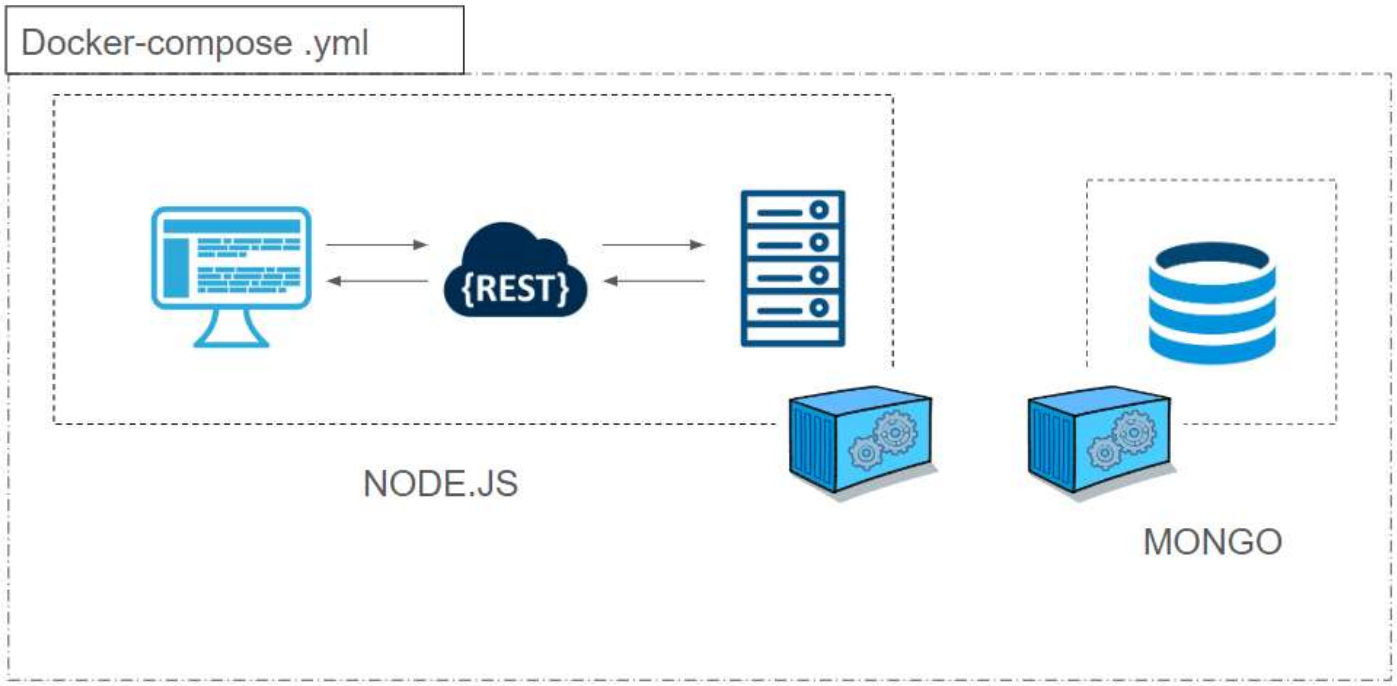
Controllers: Write controller functions to handle route logic, interact with models for data retrieval and manipulation, and send appropriate responses to clients. Ensure separation of concerns by keeping business logic separate from route definitions.

Authentication:

Passport.js: Passport.js is used for user authentication and session management, providing a flexible authentication middleware for Node.js applications.

Deployment:

Docker containers on AWS EC2: Docker is used for containerizing the application and its dependencies, allowing for consistent deployment across different environments. AWS EC2 (Elastic Compute Cloud) is utilized for hosting the Docker containers, providing scalable and reliable compute capacity in the cloud.



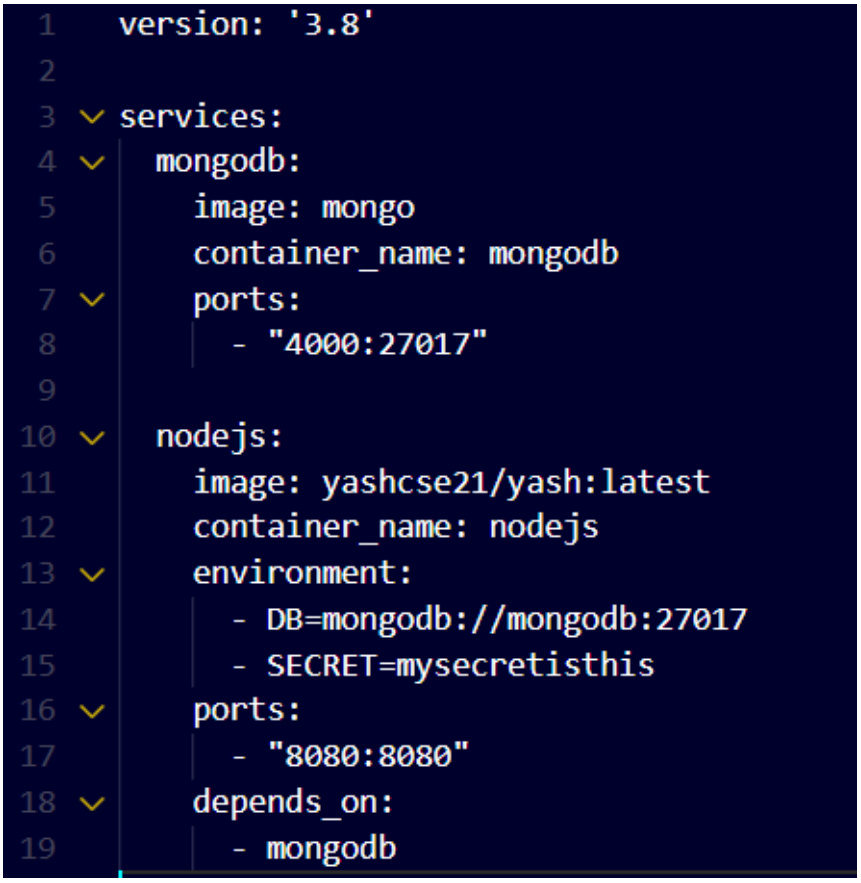
HOW TO BUILD :

Requirement:

Docker-engine
Docker-desktop(optional)
Docker-compose

Build command :

mkdir app
cd app
touch docker-compose.yml
nano docker-compose.yml



```yaml
version: '3.8'

services:
  mongodb:
    image: mongo
    container_name: mongodb
    ports:
      - "4000:27017"

  nodejs:
    image: yashcse21/yash:latest
    container_name: nodejs
    environment:
      - DB=mongodb://mongodb:27017
      - SECRET=mysecretisthis
    ports:
      - "8080:8080"
    depends_on:
      - mongodb
```

docker-compose up

AWS (ec2)

LAUNCH A CONTAINER (UBUNTU AS THE OS IMAGE)
CONNECT THE EC2 INSTANCE WITH CLI

sudo apt-get update
sudo apt-get install docker.io -y
sudo systemctl start docker
[REPEAT THE ABOVE MENTION BUILD COMMANDS]

CHANGE NETWORK SETTING (INBOUND RULES)
ADD NEW RULE (ALL TRAFFIC)

Chapter 5 : Conclusion and Future Work

5.1 Conclusion

In this project, we have addressed the critical issue of predicting cardiovascular diseases (CVDs) using a comprehensive dataset and advanced data mining techniques. By leveraging the K-Nearest Neighbours (KNN) algorithm and a rich set of health-related features, we have developed a predictive model capable of identifying individuals at risk of developing CVDs. The integration of this model into a web application provides users with a convenient platform for assessing their cardiovascular health and receiving personalized predictions.

The utilization of various technologies, including Node.js, Express.js, MongoDB, and Docker, has facilitated the development and deployment of the web application. The modular architecture, following the MVC model, ensures scalability, maintainability, and separation of concerns, enhancing the overall robustness of the system. Additionally, the incorporation of user authentication and session management using Passport.js ensures data security and privacy, enabling users to securely access and manage their health information.

5.2 Future Work

Despite the successful development of the predictive model and web application, there are several avenues for future improvement and expansion:
Enhanced Prediction Models: Continuously refine and optimize the predictive model by incorporating additional features, exploring alternative algorithms, and fine-tuning hyperparameters to improve accuracy and reliability.
User Engagement Features: Introduce interactive features such as personalized health recommendations, progress tracking, and community forums to enhance user engagement and support proactive health management.
Integration with Wearable Devices: Integrate with wearable devices and health tracking platforms to collect real-time health data, allowing for more accurate and timely predictions and facilitating seamless user experience.

REFERENCES

- https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data
- https://scikit-learn.org/0.21/documentation.html