Forecasting Heart Disease Risk Through Lifestyle Analysis Using Machine Learning

By
Md. Firoz Hasan
213-15-4313

Md. Mahmudur Rahman Sabbir
213-15-4315

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Mr. Md. Shahriar Shakil
Lecturer
Department of Computer Science and Engineering
Daffodil International University

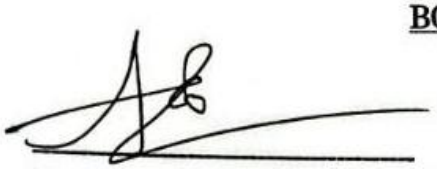
DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 17, 2025

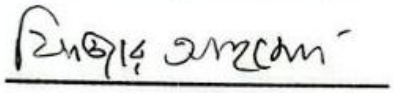
APPROVAL

This Project titled **"Forecasting Heart Disease Risk Through Lifestyle Analysis Using Machine Learning"**, submitted by **Md. Firoz Hasan**, ID No: **213-15-4313** and **Md. Mahmudur Rahman Sabbir**, ID No: **213-15-4315** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **17 September, 2025.**

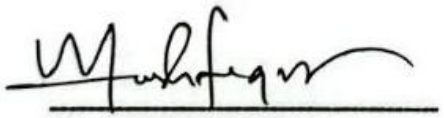
BOARD OF EXAMINERS



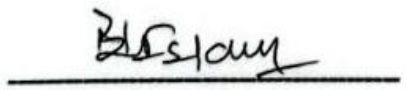
Dr. Arif Mahmud — **Chairman**
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Fizar Ahmed — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Mushfiqur Rahman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Manowarul Islam — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Naznin Sultana, Associate Professor,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Mr. Md. Shahriar Shakil
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Md. Firoz Hasan

Md. Firoz Hasan
Student ID: 213-15-4313
Department of Computer Science and Engineering
Daffodil International University

Md. Mahmudur Rahman Sabbir

Md. Mahmudur Rahman Sabbir
Student ID: 213-15-4315
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has as- sisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project(FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dr. Naznin Sultana**, **Associate Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the **Professor Dr. Sheak Rashed Haider Noori**, Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Cardiac disease remains a leading cause of death worldwide, and lifestyle components such as diet, physical exercise, consumption of fruit and vegetables or oily and fried foods, smoking, alcohol consumption, stress levels and sleeping habits have considerable roles to play in the development and initiation of cardiac disease. Detection of people at risk at an early stage will enable treatment to be initiated on time and may also stem the tide against the healthcare system. It is proposing a machine learning model using clustering to forecast heart disease risk from health and lifestyle related traits. Data preprocessing tasks, including missing value handling, encoding of categorical variables, and feature selection, were conducted to ensure data quality and model accuracy. Unsupervised learning using the K-Means clustering algorithm was carried out for the division of individuals into distinct risk clusters. Model performance was verified using internal validation metrics such as silhouette score and Davies–Bouldin index for effective clustering. From the results, it can be seen that the proposed method can effectively cluster the subjects into low, moderate, and high-risk groups, providing valuable information for targeted preventive intervention. The results show the prospects of machine learning in the development of predictive healthcare, especially in resource-poor settings. Future work includes expanding the dataset, incorporating additional lifestyle parameters, and deploying the model within a real-time decision-support system for clinicians.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the overall research is introduced, including the background of the study, justification for topic selection, main aims, and the methodological approach. Additionally, it highlights the anticipated results and explains how the report is organized.

1.1 Introduction

Cardiovascular diseases (CVD) or coronary heart disease are also the most lethal illnesses of the world and account for most of the world's morbidity and mortality [1]. Cardiovascular disease has been approximated by the World Health Organization (WHO) as accounting for nearly 17.9 million deaths each year and is the cause of accounting for 32% of all the deaths that occur in the world [2]. Nine of every ten deaths caused by such circumstances are lifestyle issues, and those are preventable, and some are dietetic issues, physical inactivity, cigarette smoking, alcohol consumption, stress, and sleep deprivation.

The situation is more serious in low and middle-income countries like Bangladesh, where constrained access to health care services, ignorance, and cost restrictions prevent earlier detection and prevention [3]. The majority are usually undiagnosed until the disease reaches a critical stage, thus the need for low-cost and accessible preventive interventions.

To this end, ML offers a robust tool for analyzing large-scale health and lifestyle data sets to discover hidden patterns and correlations beyond the capability of traditional techniques. By applying unsupervised machine learning techniques like clustering, it is possible to segregate individuals into different risk groups (high, medium, and low risk) without having to use labeled medical data. This not only improves predictive accuracy but also makes scalability across different healthcare settings easier.

The primary objective of this research is to develop a data-driven predictive system for heart disease risk based on machine learning techniques in order to assess the risk for cardiovascular disease taking into account lifestyle and behavioral factors. These consist of dietary habits, physical activity, cigarette smoking and alcohol consumption, sleeping habits, stress level, history of CVD in the first-degree relatives, and other co-existing medical disease.

Last but not least, it has to be used as an adjunctive and preventive therapy among patients and healthcare personnel. It can assist in the early identification of high-risk groups, contributing to early intervention, awareness generation, and clinical decision-making. Besides this, with its cost-effectiveness and

adaptability, the model might be available on web health platforms and mobile apps, thus being able to reach Bangladesh's rural and urban poor along with other parts of the world.

1.2 Motivation

The motivation for this work is the pressing need for preventive and affordable care of heart disease in regions where heart disease is expanding fast, but the quality of medical infrastructure is weak. In Bangladesh, a large number of people are deprived of early diagnosis or exposure to cardiovascular diseases. Traditional medical screening is too expensive, time-consuming, or physically unreachable—particularly in rural and semiurban regions [4].

In addition, the majority of heart disease risk factors are lifestyle-related and accrue over years gradually, often without being noticed, before overtly negative health consequences occur. Early behavior based and lifestyle-based risk profiling has been shown to reduce disease morbidity as well as prevent complications significantly [5].

Technologically, the abundance of health and lifestyle information and artificial intelligence advancements have made possible the creation of intelligent systems for early disease detection. Machine learning algorithms have been proven, as indicated by recent research, to be able to learn complex patterns from multi-dimensional health information and generate high reliability in cardiovascular risk prediction [6].

On an individual level, doing this study is both intellectually challenging and socially significant. Using data science for real-world health influence not only enhances technical capability but also advances the end purpose of improving lives through actionable innovation. It inspires us to come up with a predictive model that not only works but also performs well for public health—particularly in Bangladesh and similar underdeveloped areas.

1.3 Objectives

The main aims of the current research are presented as follows:

1. To identify and assess significant lifestyle and behavioral risk factors including smoking, alcohol use, diet, sleep habits, amount of stress, family history, and exercise that lead to heart disease.
2. To pre-process and transform real-world survey data into clean and consistent form suitable for clustering-based machine learning analysis such that missing values are appropriately dealt with, normalized, and feature encoded.

3. To implement and test the application of unsupervised machine learning techniques, more precisely cluster algorithms K-Means, Hierarchical, and Fuzzy C-Means to categorize individuals into several classes of risk of heart disease risk and analyze the clusters obtained for valuable patterns.
4. To develop and point a workable framework to introduce clustering results to a digital health platform for the purposes of early risk detection, preventive measures, and awareness, especially in low-resource settings like rural Bangladesh.

1.4 Methodology

The method employed here was a step-by-step pipeline beginning with data collection via Google Forms, wherein health- and lifestyle-driven features of the population were recorded. The raw data were then preprocessed by involving missing value management, categorical variable encoding, normalization, frequency mapping, and removal of irrelevant features in order to keep the data quality and consistency intact. Feature selection and transformation were then performed in order to highlight the most important factors that constitute heart disease risk.

Following this, the unsupervised cluster analysis methods—namely K-Means, Hierarchical Clustering, and Fuzzy C-Means (FCM) were used in a bid to categorize the participants into three levels of cardiovascular risk: low, medium, and high. All the three methods came up with unique patterns of lifestyle behavior and its correlation with cardiovascular risk. In the interests of checking the quality and reliability of the clusters, the Silhouette Score and Davies–Bouldin Index (DBI) were used as measure methods.

Overall, this method ensures a wide, systematic, and interpretable process by which the integration of preprocessing, clustering, and evaluation serves as the foundation for constructing an authentic and data-driven heart disease risk prediction system.

1.5 Project Outcome

Anticipated outcomes of this project are:

- A machine learning-based predictive model that can predict people into low, medium, or high risk groups of heart disease on the basic of behavioral and lifestyle data.
- Identification of the most influential lifestyle-related risk factors associated with heart disease through exploratory data analysis and feature selection techniques.
- An affordable, pragmatic method for heart disease risk assessment that can be incorporated into electronic healthcare systems.

- Increased awareness of the necessity for lifestyle habits in cardiovascular diseases prevention, especially in underserved communities.
- A foundation for future research or system development in data driven healthcare solutions for low resource settings like rural Bangladesh.

1.6 Organization of the Report

This report is divided into six chapters to present the research work systematically and logically.

- **Chapter 1 – Introduction:**
  Provides the background, motivation, objectives, overview of methodology overview, expected outcome, and the organization of the report.

- **Chapter 2 – Background:**
  Considers appropriate literature, reveals previous studies, identifies the relevant gaps in the studies, and provides an overview of the theoretical framework of the work.

- **Chapter 3 – Research Methodology:**
  Describes the proposed methodology, system design, non-functional and functional requirements, data flow diagrams, detailed methodology, project plan, and task allocation.

- **Chapter 4 – Implementation and Results:**
  Demonstrates the environment setup, testing and evaluation process, performance analysis, and results discussion of clustering based experiments.

- **Chapter 5 – Engineering Standards and Design Challenges:**
  Covers compliance with software, hardware, and communications standards in addition to societal, environmental, ethical, and sustainability considerations. It also addresses project management, financial analysis, and board engineering problem mapping.

- **Chapter 6 – Conclusion:**
  Summarizes the whole work, identifies limitation, and indicates possible future research directions.

Chapter 2

Background

In this chapter, we provides essential background information related to heart disease risk prediction using machine learning, including a review of existing literature, identification of relevant applications, and analysis of research gaps to justify the importance of our proposed work.

2.1 Introduction

Cardiovascular diseases (CVDs) remain the principal causes of morbidity and mortality in the world, and a sudden public health emergency. Burden of CVD is an issue that is rising most rapidly in low- and middle-income nations, where preventable lifestyle risk factors such as smoking, poor diet, physical inactivity, alcohol consumption, insufficient sleep, and excessive stress are key drivers of the disease burden[7]. Early detection and preventions are the solution to this problem. However, in most regions such as Bangladesh, limited access to diagnostic centers and ignorance are hindering early intervention.

With the rapid expansion of digital health data, Sudlow C. et al. addressed machine learning (ML) techniques have emerged as valuable tools to gain useful knowledge and support preventive health care[8]. Nhs.uk et al. addressed unsupervised learning techniques such as clustering have proven effective in handling unlabeled lifestyle and health data sets so that concealed risk groups can be identified without having to use prior diagnostic labels[9]. Recent research has been able to prove that health-risk behaviors tend to cluster within a person, enhancing the overall effect of clustering on cardiovascular risk. Tegegne et al. revealed that those who had three clustered risk behaviors had 25.18 times increased odds of having CVD than those with fewer risk factors[10].

Other research highlights the robustness of clustering-based methods compared to traditional predictive models. Yacaman Mendez et al. demonstrated that clustering techniques not only yielded comparable prediction accuracy but also identified a greater percentage of high-risk individuals more sensitively[11]. Similarly, Schulz MA. et al. proposed longitudinal research on lifestyle clusters (exercise, diet, smoking) detected persistent patterns with strong correlations for long-term cardiovascular outcomes[12]. Pocuca N. et.at addressed subsequent ML studies using UK Biobank data that unsupervised techniques could supplement supervised prediction when there is little labeled clinical data [13].

These findings also strongly support the validity of applying clustering methods to cardiovascular risk assessment. Therefore, this research seeks to apply clustering algorithms—K-Means, Hierarchical Clustering, and Fuzzy C-Means—

to survey responses regarding lifestyle variables for classifying individuals as low, medium, and high risk for heart disease. The long-term vision is to create an efficient, yet accessible, system which could potentially aid in early screening, awareness, and preventive interventions in resource-limited populations such as rural Bangladesh.

2.2 Literature Review

This literature review establishes a theoretical framework by synthesizing key texts in marketing and competitive strategy. Reading these foundational texts, we establish entrenched principles and the research gaps that our study will address.

Table 2.1: Summary of Literature Reviewed

Author (s)	Year	Title	Methodology	Key Findings
Tegegne et al.[10]	2022	Lifestyle risk behaviour clustering & CVD risk	Latent Class Analysis	Clustered lifestyle risks increased CVD odds by up to 25.
Yacaman Mendez et al.[11]	2025	Cluster vs traditional CVD risk models	Cluster analysis comparison	Clustering performed comparably, higher sensitivity.
Zhang et al.[14]	2025	Biomarker-enhanced CVD risk prediction	Explainable ML + Cox models	Improved C-index (up to 0.82) integrating biomarkers.
Shishehbori et al.[15]	2024	ML enhancement of CVD risk models	Survey of ML methods	Deep learning models outperform traditional risk scores.
Aashray k. Gupta et al.[16]	2024	Unsupervised clustering on EMR data	K-means vs supervised	Clustering achieved similar accuracy.
T.Tgegne et al.[17]	2023	Longitudinal changes in clustered behaviors	Latent transition analysis	Adults with multiple habits seldom moved to low-risk cluster.

Nilay S. Shah et al.[18]	2025	Equity in CVD risk prediction models	Development of PREVENT equations	Improved subgroup representation.
Leiva-Juarez, Maria A., et al. 19]	2025	Polygenic risk scores + lifestyle interactions	Polygenic-lifestyle interaction	Combined effects enhance prediction.
S. Barbieri et al.[20]	2020	Deep learning vs Cox model for CVD events prediction	Survival DL vs traditional methods	DL significantly outperformed Cox models.
Aizatul Shafiqah Mohd et al.[21]	2021	AI-enhanced CVD risk models	Literature review	Advocates integration of multi-modal AI methods.
Adam J. Lewandowski et al.[22]	2024	Scientific and Clinical Impacts of UK Biobank in Cardiovascular Medicine	A large-scale prospective cohort with comprehensive longitudinal data	Highly relevant for background context on data availability and scope in cardiovascular.
Stefanie J. Krauth et al.[23]	2024	Association of Latent Class Analysis–Derived Multimorbidity Clusters	Clusters of multimorbidity to health	Support for clustering approaches to health outcomes like hospital admissions.

2.3 Gap Analysis

Here is a summary of the gaps we have found in the related work study, where we would like to contribute.

Table 2.2: Summary of Gap Analysis

Features	Existing Literature	Proposed System (Our Research)
Lifestyle-based clustering for heart disease risk	Exists in research for developed datasets, but not localized	Tailored to Bangladeshi demographic context
Focus on modifiable lifestyle factors only	Often includes clinical/biomarker data	Uses exclusively behavioral/lifestyle data
Integration with digital health platforms	Not commonly implemented in suite form	Design-ready for real-time digital tool
High sensitivity / low-cost method	Some studies provide performance, cost not addressed	Prioritizes cost-effectiveness and accuracy
Data collection strategy for underserved regions	Predominantly general population or developed region data	Focus informed by localized, accessible survey responses
Unsupervised methods for risk exploration	Limited evidence of clustering in real-world lifestyle data	Central to the system design
Explanation of Feature Importance	Limited explanation (black-box)	Provides clearer interpretability through clustering
Custom-designed Lifestyle Survey Integration	General lifestyle datasets used	Uses a uniquely designed lifestyle questionnaire/survey

Although previous research on clustering in cardiovascular disease (CVD) has been successful with particular datasets, a significant shortfall lies in constructing context-specific, cost-effective, behavior-driven models viable in low-resource environments like rural Bangladesh. The majority of work is executed with the incorporation of clinical data and biomarkers, which are not readily obtainable in inadequately equipped medical environments. Additionally, the incorporation of predictive models into digital health platforms for real-time tracking and awareness creation is not considered prominent.

Our research is directly working to address these limitations by:

- Utilizing solely lifestyle and behavior data, which is readily available and less intrusive.
- Employing unsupervised clustering to classify hidden at-risk groups irrespective of labeled clinical endpoints.
- Developing the model with future digital health embedding in mind, making frontline healthcare adoption feasible.
- Developing the model with future digital health embedding in mind, making frontline healthcare adoption feasible.

2.4 Summary

The chapter has explained a comprehensive background study to validate the system designed for heart disease risk prediction based on machine learning, more so clustering algorithms. The chapter began with the description of the importance of predictive medicine and how unsupervised learning can be used in the research of lifestyle determinants. A comprehensive literature survey was conducted involving ten prominent studies, all of which were analyzed for various methodologies, datasets, and conclusions for heart disease prediction. Upon review, a gap analysis was submitted to introduce the deficits of existing approaches, such as limited use of clustering, inadequate integration of behavioral data, and lack of interpretable models. These gaps lay the foundation for the goals of the proposed system and justify its contribution to the state of affairs.

Chapter 3

Research Methodology

In this chapter, the research methodology and project design specifications are described. It includes the methodology overview, proposed system design, requirement analysis, data flow diagrams, detailed methodology, project plan, task allocation, and a summary of the chapter.

3.1 Methodology/Requirement Analysis & Design Specification

3.1.1 Overview

This study uses an unsupervised machine learning clustering technique to forecast heart disease risk based on lifestyle data analysis. The methodology focuses on data collection, preprocessing, clustering model development, and designing the risk classification system.

3.1.2 Proposed Methodology/ System Design

This work is organized in the context of a quantitative, predictive research paradigm with unsupervised machine learning methods. The overall goal is to design a system able to predict risk of heart disease using lifestyle factors gathered from structured interviews. A methodology diagram is disseminated figure[3.1], to explain the process.

The approach to follow in this project is to use unsupervised machine learning methods (clustering) to predict the risk of heart diseases based on lifestyle and health variables. It starts with data collection, where appropriate datasets with demographic, lifestyle and clinical features are acquired from a trusted source.
Subsequently the data cleaning and pre-processing phase organized the data set in appropriate data structure and format for analysis. This step will require formatting data by missing/variable transformation to a format to be fed into ML model), normalization and noise removal.

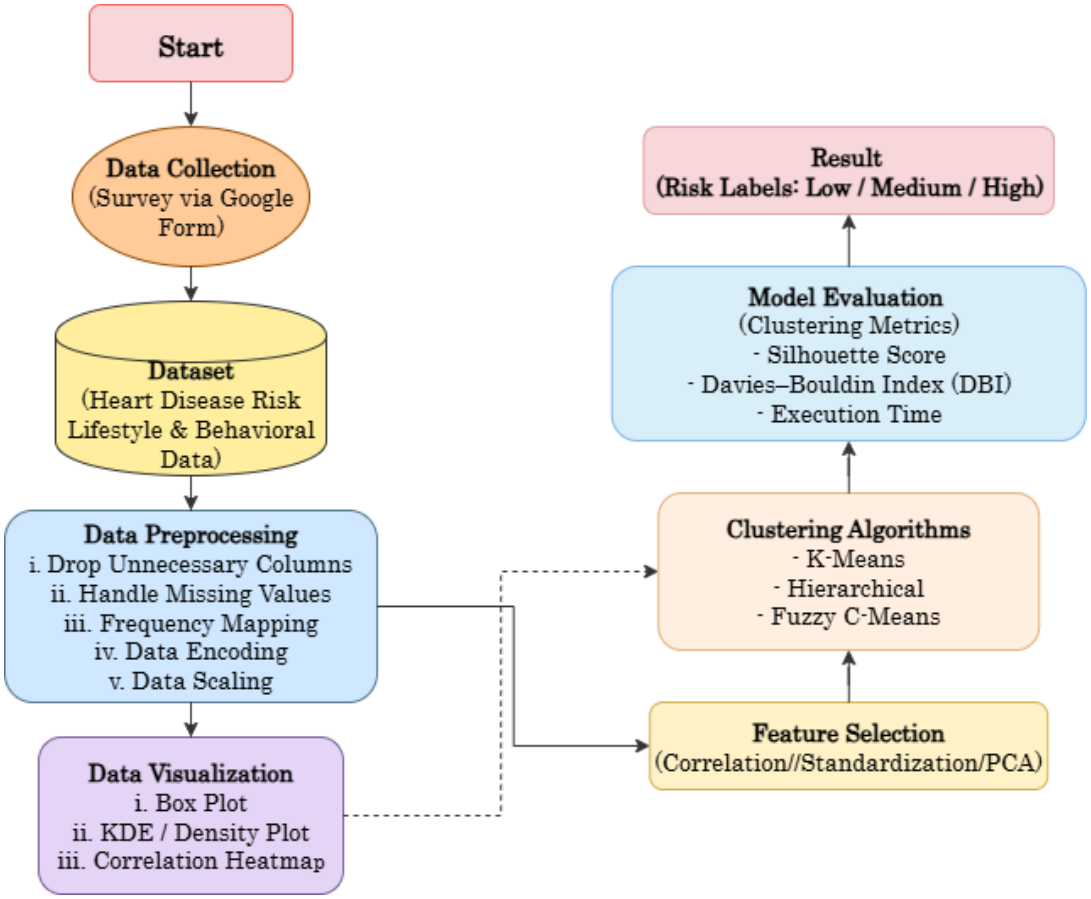


Figure 3.1: Methodology Diagram

Thereafter, the selection of the most important features correlated with the risk of a heart disease is performed as well as data transformation. This ensures that, the clustering model is fed with only useful and informative input features.

The Hierarchical or K-Means model is then applied to group people into different risk categories— low, medium, and high risk, for example based on the similarity of their characteristics. Clustering is followed by validation and assessment in term of metrics such as silhouette score, Davies–Bouldin index, Execution Time for assessing cluster quality and interpretability.

Afterwards, findings are explained and conveyed through risk maps, graphs, and charts to facilitate action. Individuals and health workers can assess their potential levels of risk and take part in prevention. The whole procedure ensure a systematic and accurate way of forecasting heart disease risk, from raw material to useful risk stratification.

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements :

The functional requirements of the proposed system are:

1. **Data Collection Module** – Collect lifestyle and behavioral survey data (e.g., smoking, alcohol consumed, physical activity, diet, sleep, stress, family history).
2. **Data Preprocessing** – Missing value Handling, normalize, standardize data, and encode categorical variables.
3. **Clustering Model Implementation** – Apply unsupervised machine learning algorithm K-Means, Hierarchical, Fuzzy C-Means Clustering to cluster individuals based on lifestyle similarity.
4. **Risk Group Assignment** – Map clustering results to qualitative risk levels (Low, Medium, High) based on health-related attributes.
5. **Visualization** – Visualize cluster distribution and lifestyle patterns using charts/graphs.
6. **Result Export** – Allow exporting of filtered data and clustering results for further analysis.
7. **User Input Interface (Optional)** – Allow input of new user data for prediction of probable risk group using the trained clustering model.

Nonfunctional Requirements :

The non-functional requirements of the proposed system are:

1. **Performance** – The system should be able to cluster thousand records of data in seconds.
2. **Scalability** – Should be able to handle larger datasets with minimal loss of performance.
3. **Mapping Accuracy** – Ensure meaningful mapping between clusters and risk categories by checking against known health indicators.
4. **Usability** – Provide a user friendly interface for non-technical as well as technical users.
5. **Portability** – The system should be capable of deploying on a typical desktop and web environment.
6. **Security** – Protect confidential health-related data by storing them safely and limiting access.
7. **Maintainability** – Code should be modular and well documented for facilitate future updates.

3.1.4 Data Flow Diagram

At DFD [Figure 3.2], addressed main components. It first collects and refines user data through preprocessing, feature selection, and transformation. This refined data is then grouped into risk categories using clustering algorithms. Finally, the system generates Risk Reports for users and Summary Reports for health experts.

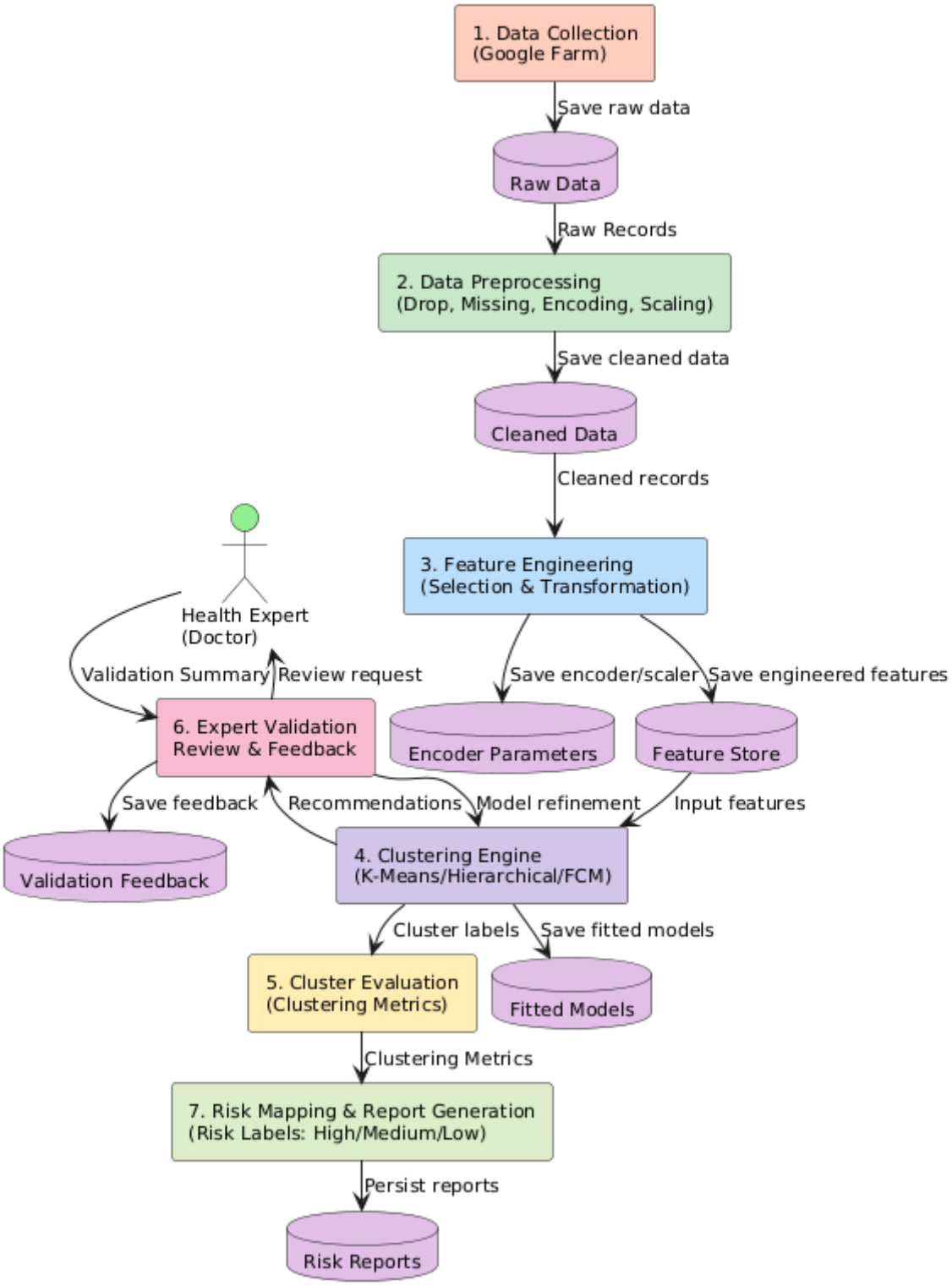


Figure 3.2: Data Flow Diagram

3.2 Detailed Methodology and Design

3.2.1 Survey Questionnaire

To collect actual lifestyle and health-related data, a structured questionnaire was drafted and distributed via Google Forms. The questionnaire consisted of demographic, lifestyle, and medical questions that are directly related to the risk of heart disease. Respondents were also informed that their data would be used strictly for academic and research purposes. The main research questions were as follows:

- What is your gender?
- What is your age?
- What is your occupation?
- How often do you engage in physical activity ?
- How often do you consume fruits and vegetables?
- How often do you consume fried or fatty foods?
- How often do you consume alcoholic drinks?
- Do you smoke tobacco products?
- What is your usual mode of transportation?
- How many hours of sleep do you get per night(On average)?
- How would you rate your stress levels?
- Do you have a family history of heart disease ?
- Have you ever been diagnosed with high blood pressure, high cholesterol, diabetes, or obesity?
- How often do you have medical check-ups?
- Do you engage in regular cardiovascular exercise?
- Are you currently taking any medications for heart medication?

The survey questions, validated by a doctor, can be found in Appendix A.

3.2.2 Research Subject and Instrumentation

This study focuses on predicting heart disease risk using unsupervised machine learning techniques, particularly clustering algorithms. The implementation integrates multiple Python-based libraries and computational tools to preprocess, transform, and analyze lifestyle-related health data. The instruments used in this research are described below:

Google Colab

Google Colab is a cloud hosted Jupyter notebook environment with unlimited access to computational resources. It bypasses the local installation requirement and provides a collaborative environment where many contributors can edit and execute code in real-time, similar to Google Docs. Its simplicity of use with widely used Python libraries makes it an extremely suitable platform for machine learning experiments.

NumPy

NumPy is a python numerical library. It provides efficient support for multidimensional arrays and mathematical operations between arrays. During this research, NumPy was used to process matrices, perform statistical computations, and perform numerical conversions during preprocessing.

Pandas

Pandas is an advanced data manipulation and analysis library of Python. It has features such as data structures DataFrames that allow for structured storage and dataset processing efficiently. Pandas was used in the current research to clean the data, handle missing values and prepare the survey data for cluster analysis.

Matplotlib

Matplotlib is a widely used Python package that can be used to 2D charts and plots. It has full control over the figure layout so it can generate accurate visualization of data patterns and relationships. Matplotlib was used in this research to generate accurate graphs for feature trends, cluster distribution, and evaluation metrics.

Seaborn

Seaborn is a statistical visualization library built on top of Matplotlib. It provides a high-level interface for generating attractive and informative statistical graphics such as heatmaps, pair plots, and box plots. Seaborn was used in this project to plot correlations between lifestyle characteristics and present cluster results in a more understandable form.

Label Encoder & Standard Scaler

Label Encoder, part of Scikit-learn, is used to convert the categorical variables to numeric data that is suitable for machine learning algorithms. Standard Scaler scales the features by standardizing to zero mean and unit variance. These two programs together ensured that the dataset was properly encoded and scaled ready for cluster analysis.

Agglomerative Clustering

Agglomerative Clustering is a hierarchical clustering algorithm that forms nested clusters by merging data points incrementally based on similarity. Agglomerative Clustering was used as an alternative to K-Means to compare the efficiency of clustering and produce interpretable dendrogram shapes.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms high dimensional data into fewer principal components with maximum variance. In this study, PCA was used to map high dimensional lifestyle and health data into 2D or 3D space in order to easily interpret clusters.

Dendrogram & Linkage

Dendrograms are tree-like diagrams that provide a visualization of the structure of clusters for hierarchical clustering. Linkage methods control cluster-to-cluster distances (e.g., single, complete, average). They have been used in this research to explain hierarchical structure and validate cluster formation.

Silhouette Score

Silhouette Score is a clustering internal assessment metric for determining the similarity of an object with its own cluster compared to other clusters. The higher the score, the more separation and cohesion among the clusters. It was used to validate the quality of clustering results.

3.2.3 Data Collection

The information for this study was collected with a structured Google Form survey, which was extremely well suited to collecting lifestyle patterns as well as health data.

SL. No.	Gender	Age	Occupation	Physical activity	Consume fruits and vegetables	Consume fried or fatty foods	Consume alcoholic beverages	Smoke tobacco products
1	M	20-30	Student	1-2 times a week	4-5 times a week	3-4 times a week	Once a week	Yes
2	M	20-30	Student	3-4 times a week	Daily	Daily	Never	No
3	F	20-30	Student	Never	1-3 times a week	Daily	Never	No
560	F	51-60	Employed (part-time)	Rarely	1-3 times a week	Rarely	Once a week	No
561	F	41-50	Self-employed	Rarely	1-3 times a week	3-4 times a week	Several times a week	No
	Family history of heart disease	**Sleeping hour**	**Stress levels**	**Transportation**	**Diagnosing conditions**	**Medical check-ups**	**Cardiovascular exercise**	**Medications for heart health conditions**
1	Yes	6-7 hours	Moderate	Public transportation	None of the above	Rarely	No, rarely or never	No
2	No	7-8 hours	Moderate	Public transportation	High blood pressure	Never	No, rarely or never	No
3	Yes	More than 8 hours	Moderate	Walking/cycling	None of the above	Rarely	Occasionally	No
560	No	Less than 6 hours	Low	Public transportation	High blood pressure;Diabetes	Only when I feel unwell	No, rarely or never	No
561	Yes	Less than 6 hours	Moderate	Private vehicle (car/motorcycle)	Diabetes;Obesity	Only when I feel unwell	No, rarely or never	Yes

Table 3.1: Dataset in Table Format

Like that [Table 3.1], there were 561 initial respondents, and their data had approximately 18 features of information with 371 men and 190 women. The dataset is mainly all the data type were in object. Some very crucial topics were covered in the questions in the survey: demographics, lifestyle patterns, and medical history. The answers to the Google Form were saved automatically to Google Sheets to

prevent human error in data input. The data was imported into a CSV file in an attempt to preprocess and further analyze it with Python. CSV format permitted easy incorporation into machine learning libraries such as Pandas, NumPy, and Scikit-learn.

3.2.4 Data Preprocessing

Before applying clustering algorithms, the data collected was preprocessed extensively to attain data quality, consistency, and machine learning readiness. Preprocessing is a crucial phase in transforming raw survey responses to a structured format that can be utilized for risk prediction. The preprocessing steps carried out are as follows:

Dropping Unnecessary Columns

In [Figure 3.3], Some of the attributes gathered through the survey have been participant names, time stamps, comments which were not relevant to the clustering process.

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Timestamp                     561 non-null    object
 1   Name                          561 non-null    object
 2   Gender                        561 non-null    object
 3   Age                           561 non-null    object
 4   Occupation                    561 non-null    object
 5   Physical activity             561 non-null    object
 6   Fruits_vegetables             561 non-null    object
 7   Consume fried or fatty foods  561 non-null    object
 8   Consume alcoholic beverages   561 non-null    object
 9   Smoke tobacco products        561 non-null    object
 10  Family_history                561 non-null    object
 11  Sleeping hour                 561 non-null    object
 12  Stress                        561 non-null    object
 13  Transport                     561 non-null    object
 14  Diagnosed_conditions          561 non-null    object
 15  Medical_checkup               561 non-null    object
 16  Cardio_exercise               561 non-null    object
 17  Medications                   548 non-null    object
 18  Comments                      16 non-null     object
dtypes: object(19)
```

Figure 3.3: Data Types

Such non-contributory fields were eliminated in order to prevent noise in the data and also to ensure participants' privacy. Only the lifestyle and health-related attributes that directly affect the risk of heart disease have been preserved for processing ahead.

Missing Value Handling

The data collected through Google Forms contained some partial responses. Missing values can negatively impact clustering performance by creating biased clusters. Therefore, missing values were handled through imputation techniques. For categorical variables such as Medications status, the mean was imputed.

The arithmetic mean is obtained by adding the numbers and dividing the result by the total number of numbers in the list. The term "average" is typically used to refer to this.

The equation for the mean is,

$$\textbf{Mean Formula} = \frac{\textbf{Sum of Observations}}{\textbf{Total Numbers of Observations}} \qquad (1)$$

Mode is the value that appears most often in a list. The equation for the mode is:

$$\text{Mode formula} = \mathbf{M} + \mathbf{i}\frac{(\mathbf{x_m} - \mathbf{x_1})}{(\mathbf{x_m} - \mathbf{x_1}) + (\mathbf{x_m} - \mathbf{x_2})} \qquad (2)$$

Here,

- $\mathbf{M}$ is modal class's lower limit.
- i is Class interval size.
- $\boldsymbol{x_m}$ is modal class frequency.
- $\boldsymbol{x_1}$ is the class before the modal class's frequency.
- $\boldsymbol{x_2}$ is the class following the modal class's frequency.

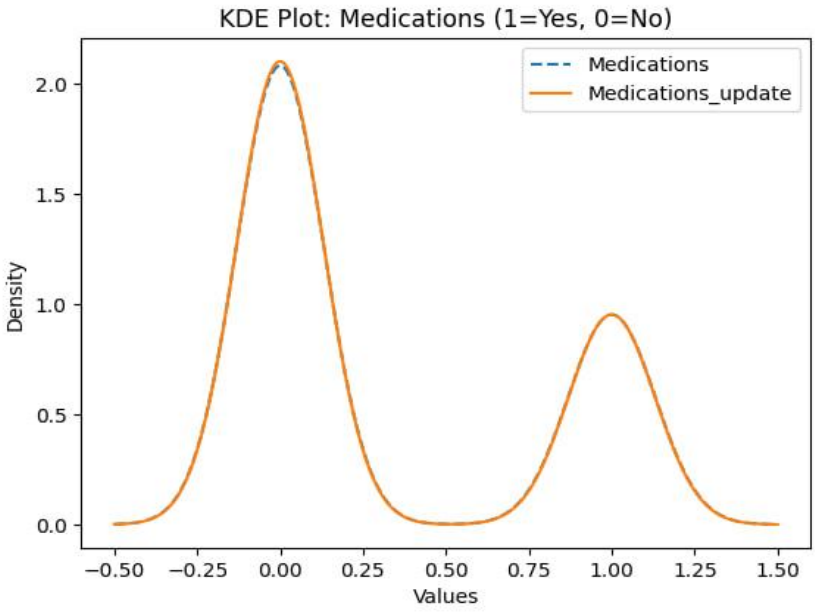


Figure 3.4: Dataset Variation After Missing Value Handling

The plots show [Figure 3.4] two lines, the blue one representing the original column and yellow one representing the filled up column. We can see there is almost no variation between the lines. So, we were able to fill up the missing values in the columns with no variation. which is a positive progress on research.

Data Normalization

Since the dataset included features with varying ranges age in years, stress levels on a categorical scale, and sleep hours in numeric form, normalization was essential. Numerical attributes were scaled using StandardScaler and Min-Max normalization to transform them into a uniform range (0–1). This step ensured that features with larger numerical ranges did not dominate the clustering process. For example, without normalization, variables like "age" could overshadow features such as "stress level" during distance calculation in K-Means clustering.

The most common data normalization formula, known as Min-Max Scaling, is:

$$\text{Normalized Value} = \frac{(\mathbf{X} - \mathbf{Xmin})}{(\mathbf{Xmax} - \mathbf{Xmin})} \qquad (3)$$

Here,

- **X** (Data Point): The specific value you want to normalize.
- **Xmin** (Minimum Value): The smallest value in my dataset or the column you are normalizing.
- **Xmax** (Maximum Value): The largest value in my dataset or the column we are normalizing.

Data Encoding

Our Most survey responses included categorical features values were encoded into numerical representations to make them compatible with machine learning algorithms. Label Encoding used for binary variables such as gender (male or female) or smoking status (yes or no), while One-Hot Encoding was applied for multi-class variables like occupation and mode of transportation. This encoding preserved the categorical distinctions without introducing unintended ordinal relationships.

After Encoding dataset [Figure 3.5(a),(b)] describe:

	Gender	Age	Occupation	Physical activity	Fruits_vegetables	Consume fried or fatty foods	Consume alcoholic beverages	Smoke tobacco products
count	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000
mean	0.661319	1.898396	4.477718	1.342246	2.675579	2.515152	1.597148	0.304813
std	0.473684	1.607576	2.611093	1.355076	0.997278	0.878433	1.570348	0.460739
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	1.000000	0.000000	2.000000	2.000000	0.000000	0.000000
50%	1.000000	2.000000	6.000000	1.000000	3.000000	2.000000	1.000000	0.000000
75%	1.000000	3.000000	6.000000	2.000000	3.000000	3.000000	3.000000	1.000000
max	1.000000	5.000000	8.000000	4.000000	4.000000	4.000000	4.000000	1.000000

Figure 3.5(a): Dataset Information After Encoding

Family_history	Sleeping hour	Stress	Transport	Diagnosed_conditions	Medical_checkup	Cardio_exercise	Medications
561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000	561.000000
0.896613	1.087344	1.622103	2.645276	10.450980	2.281640	0.743316	0.311943
0.902381	0.868825	1.209993	1.608243	6.132598	1.563821	0.642314	0.463700
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	4.000000	1.000000	0.000000	0.000000
1.000000	1.000000	2.000000	2.000000	9.000000	2.000000	1.000000	0.000000
2.000000	2.000000	2.000000	3.000000	17.000000	4.000000	1.000000	1.000000
2.000000	3.000000	4.000000	7.000000	18.000000	4.000000	2.000000	1.000000

Figure 3.5(b): Dataset Information After Encoding

Frequency Mapping:

Many categorical variables collected from the survey consist of multiple categories. Directly using these variables in clustering can sometimes lead to bias if the categories are not encoded properly. To address this, Frequency Mapping was applied as an additional transformation step.

Through these preprocessing steps including handling missing values, removing irrelevant attributes, normalizing scales, encoding categorical features, and treating outliers the dataset was transformed into a clean, standardized, and machine readable format.

3.2.5 Data Visualization

The dataset was visualized at various stages of the procedure to help with the analysis. To visualize the data, we employed boxplots, KDE, histograms, etc. Additionally, we plot feature correlations using a correlation heatmap, which has an impact on feature selection.

Correlation Heatmap:

A heatmap is a data visualization technique that uses a system of color coding to represent the magnitude of individual values in a dataset. It's essentially a 2D grid where the color of each cell corresponds to its value, with a color scale (or legend) that explains what each color represents.

The heatmap [Figure 3.6] illustrates the correlation between the dataset's features. When looking at the overall points, Medications & Family History is the strongest correlation in the set. It strongly suggests that individuals with a family history of certain conditions are more likely to be on medication themselves. The correlation between the remaining features was moderate. Many pairs show very low correlation (close to 0.00), such as Stress with most other factors, indicating it may be an independent variable in this dataset.

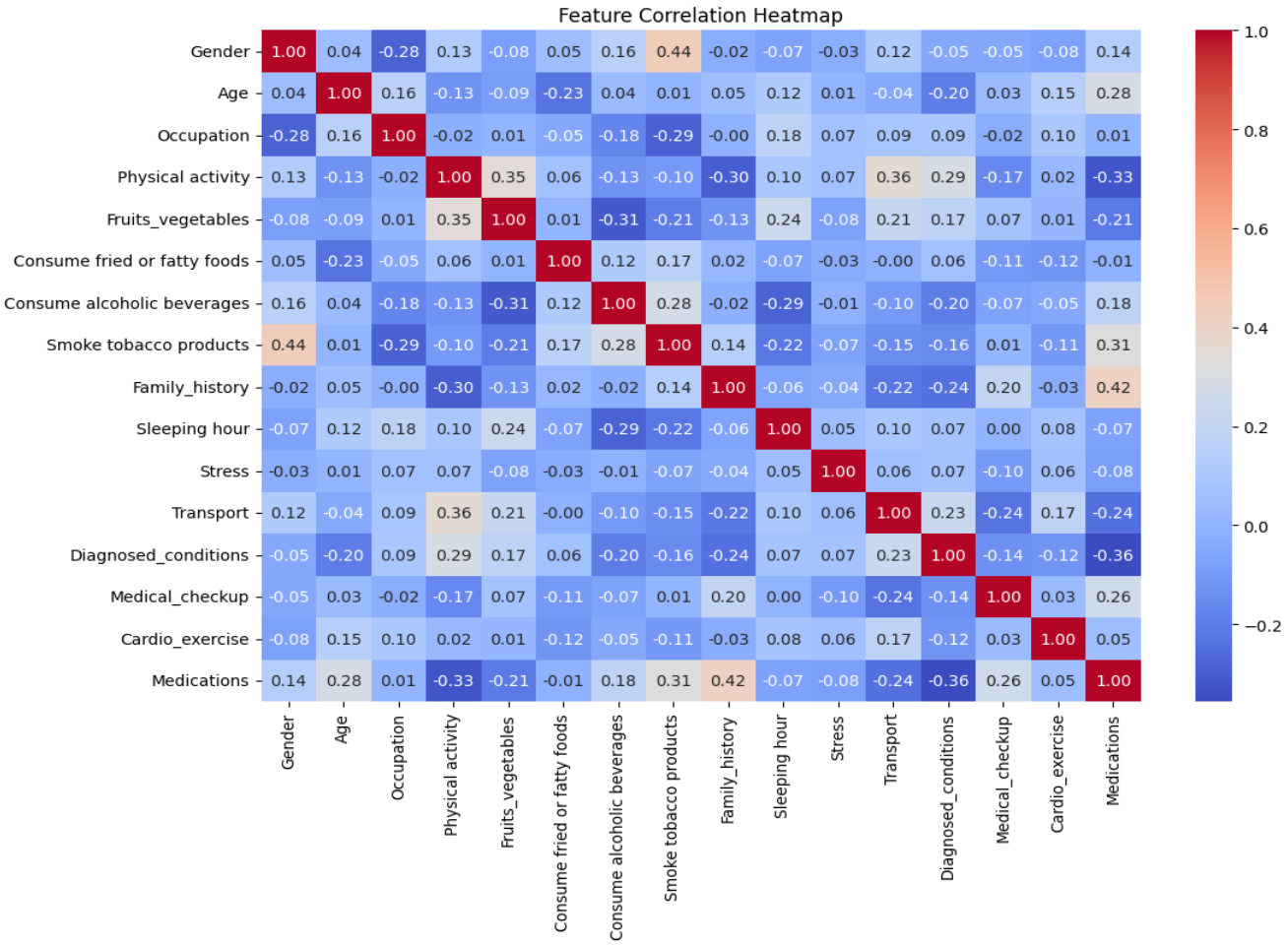


Figure 3.6: Correlations Using Heatmap

3.2.6 Feature Selection & Transformation

Feature selection and transformation are crucial steps building a clustering-based model of heart disease risk prediction. This phase ensures that only the most relevant and informative features are retained, reducing noise and improving clustering quality.

Correlation Analysis

Correlation Analysis was applied to identify redundant features that were highly correlated with one another, ensuring that overlapping information was minimized.

Standardization

For features where, normal distribution was assumed StandardScaler was applied to center data around mean zero with unit variance.

Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) used to reduce high-dimensional data into a smaller number of principal components while retaining maximum variance. This not only enhanced computational efficiency but also allowed for better visualization of clusters.

3.2.7 Model Development

The core of this study is the development and use of clustering algorithms to predict levels of risk for heart disease from health-related and lifestyle data. Since the dataset used within this project is unlabeled, the optimal method for identifying concealed patterns and dividing individuals into risk categories is through the use of unsupervised machine learning. Three algorithms were used and contrasted within this study: K-Means Clustering, Hierarchical Clustering, and Fuzzy C-Means (FCM).

K-Means Clustering

K-Means is one of the most common partitioning algorithms that groups data into k similar clusters . In this study, the algorithm segments individuals into groups where intra-group similarity is optimal and inter-group similarity is minimal. The measure of similarity was Euclidean distance, and the optimal value of k was determined using the Elbow Method and Silhouette Score. This method allowed the system to classify individuals into low, medium, and high-risk with great success. Working Algorithm:

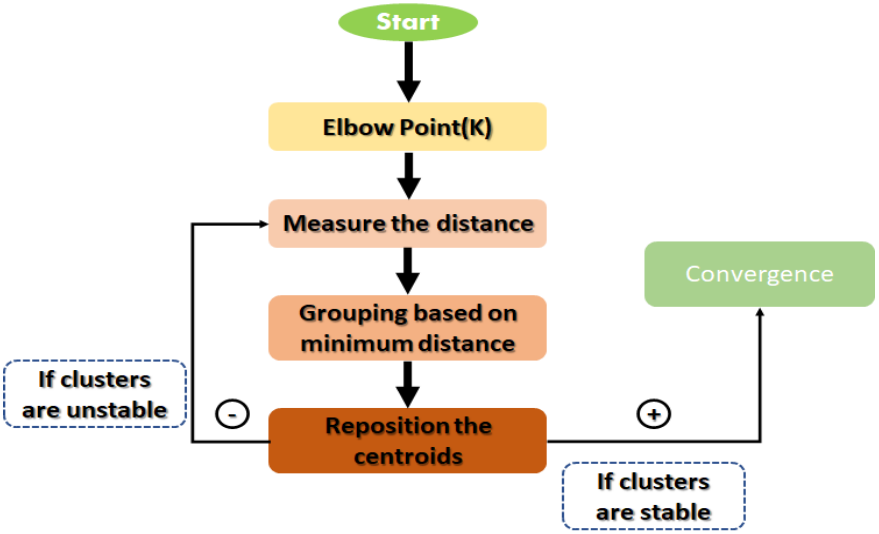


Figure 3.7: K-Means Clustering Algorithm

Objective Function (minimize within-cluster variance):

$$J = \sum_{i=1}^{K} \sum_{x \in Ci} \| x - \mu i \|^2 \qquad (4)$$

Here,

- K = number of clusters.
- Ci = cluster i.
- μi = centroid of cluster i.
- $||x - \mu i||^2$ = squared Euclidean distance between point and centroid

Steps:

- Randomly initialize *k* centroids.
- Assign each data point to the nearest centroid.
- Update centroids by taking the mean of points in each cluster.
- Repeat steps 2–3 until centroids converge (no significant change).

Hierarchical Clustering

Hierarchical clustering creates a hierarchy of clusters by using either an agglomerative (bottom-up) or divisive (top-down) approach. Agglomerative clustering with Ward's linkage method was used in this project to group individuals. A dendrogram was constructed to observe how the clusters combined and to determine the natural number of clusters. The approach was particularly useful to visualize the nested structure in the dataset and explore relationships between risk groups. There are apply Agglomerative (bottom-up) and Divisive (top-up) methods.

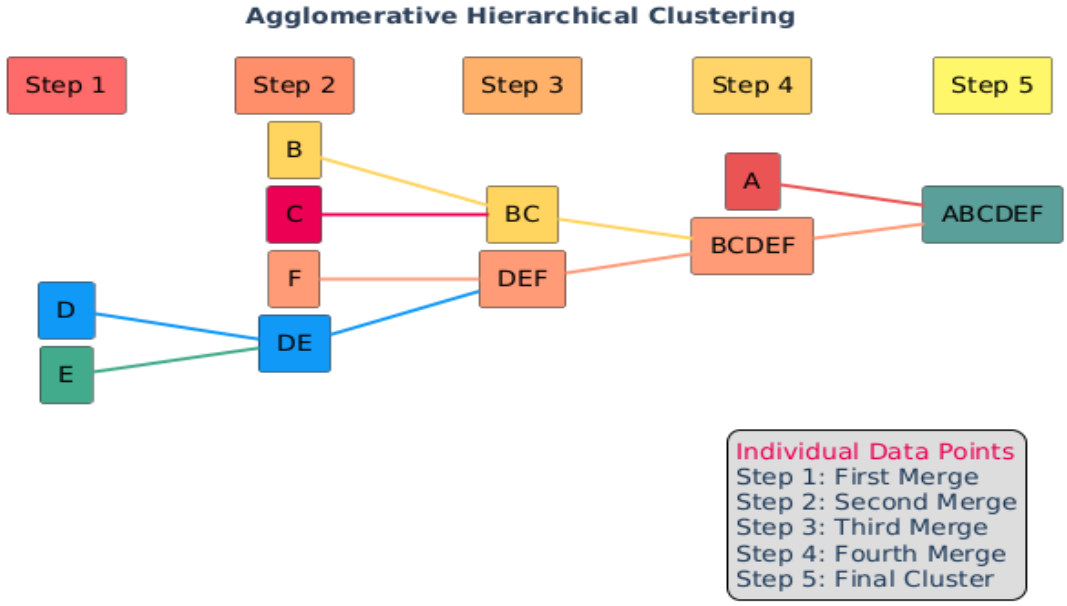


Figure 3.8: Hierarchical clustering Method

Distance between clusters depends on linkage method:

- Single linkage:

$$\mathbf{D(A,B)} = \min_{a \in A, b \in B} d(a,b) \quad (5)$$

- Complete linkage:

$$\mathbf{D(A,B)} = \max_{a \in A, b \in B} d(a,b) \quad (6)$$

- Average linkage:

$$D(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b) \quad (7)$$

Steps:

- Start with every point as a separate cluster.
- Merge two closest clusters based on chosen linkage.
- Repeated until only one cluster remains (dendrogram created).

Fuzzy C-Means (FCM)

As opposed to K-Means, which place a point in one cluster, Fuzzy C-Means allows partial membership, wherein an individual can possess high degree of membership with several clusters. FCM is particularly relevant in health databases, where the diseases of patients are not necessarily dichotomous. FCM was used in determining the uncertainty and overlapping risk factors in heart disease prediction.

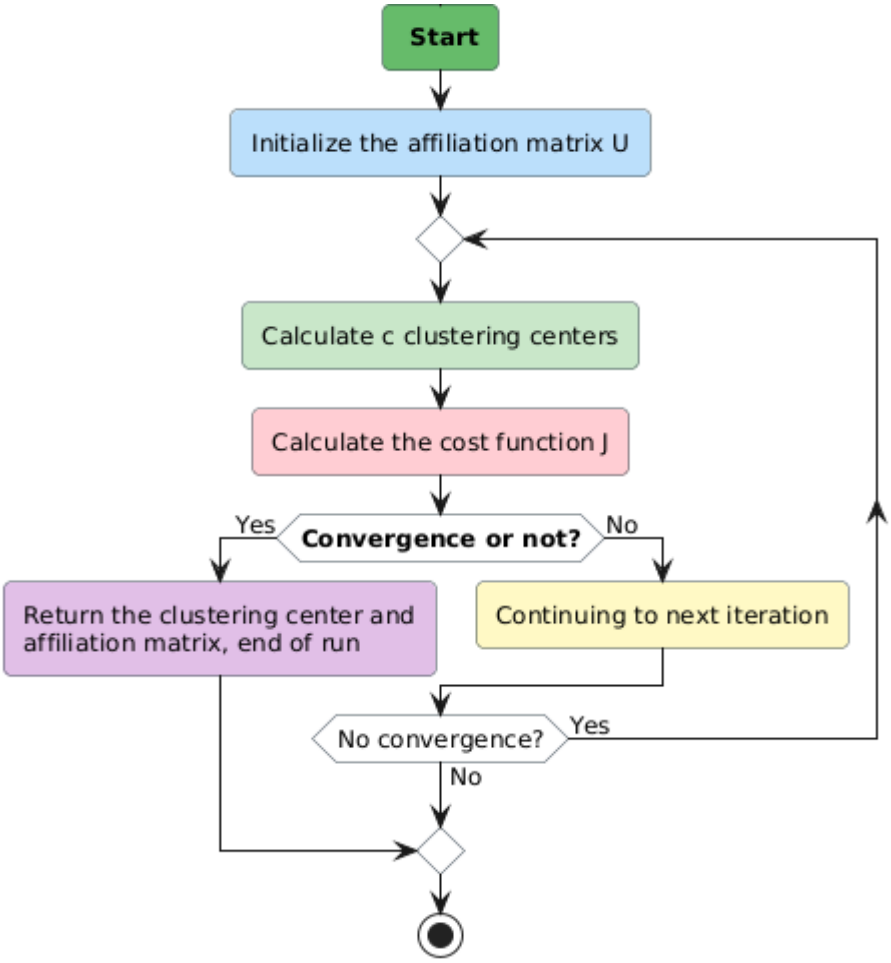


Figure 3.9: Flowchart of Fuzzy C-Means clustering Algorithm

Objective Function (minimize weighted distance):

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \cdot \left\|x_i - c_j\right\|^2 \quad (8)$$

Here,

- N = number of data points
- C = number of clusters
- u_{ij} = degree of membership of x_i in cluster j
- c_j = centroid of cluster j
- m = fuzzifier (controls fuzziness, usually m=2)

3.2.8 Model Evaluation

The evaluation of clustering models is essential to determine the effectiveness and reliability of the proposed heart disease risk prediction system. Since the dataset does not contain predefined class labels, unsupervised evaluation metrics were applied. The following measures were used to assess model performance:

Silhouette Score(S)

- The Silhouette Score evaluates how similar an object is to its own cluster compared to other clusters.
- Formula:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (9)$$

Here,

- ***a(i)*** = average distance of point iii to all other points in same cluster
- ***b(i)*** = minimum average distance of point iii to points in any other cluster

- Highest values of Silhouette Score indicate better cluster separation.

Davies–Bouldin Index (DBI)

- Formula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i+\sigma_j}{d(c_i,c_j)} \quad (10)$$

Here,

- k = number of clusters
- σ_i = average distance of all points in cluster i from its centroid c_i.
- $d(c_i, c_j)$ = distance between centroids of clusters i and j

- Lower values of DBI indicate better cluster separation.

Execution Time

- The time taken by each algorithm to complete clustering was recorded.
- Formula:

$$Execution\ Time\ (s) = t_{end} - t_{start} \quad (11)$$

- This helps to compare computational efficiency across K-Means, Hierarchical, and Fuzzy C-Means algorithms.
- Fastest is indicate better cluster separation.

Cluster Distribution Analysis

- After clustering, the distribution of individuals across Low, Medium, and High risk categories was analyzed.
- A balanced distribution indicates more meaningful clustering, whereas highly skewed clusters may signal bias or poor performance.

The combination of Silhouette Score, DBI, Execution Time, and Risk Category Distribution provided a comprehensive evaluation of the clustering algorithms. Of the ones used, K-Means possessed the best trade-off between accuracy and efficiency, Hierarchical Clustering provided understandable results with reasonably good performance, and Fuzzy C-Means provided soft clustering with enhanced flexibility.

3.2.9 Result Interpretation

After running the K-Means, Hierarchical, and Fuzzy C-Means clustering algorithms, the clusters obtained from the algorithms were interpreted for lifestyle attributes of different risk groups. This process, also referred to as cluster profiling, provides revealing information on the co-relation of health measures and lifestyle habits with the risk of heart disease.

Cluster 1 (Low Risk)

This cluster had predominantly healthy diets (daily intake of vegetable and fruits), good sleep (7–8 hours), low or moderate stress, and physical exercise on a daily basis. They also had comparatively lower disease, smoking, and alcohol rates. Which reduces the risk of heart disease, these are more healthy lifestyle factors.

Cluster 2 (Medium Risk)

This cluster included participants who were in mixed lifestyle patterns. They, for instance, moderately exercised but also consumed fried or fatty foods frequently. They varied widely in their levels of stress and sleeping hours, some of whom also had a family history of cardiovascular diseases. This cluster indicates a transitional life with a mixture of risk-free and risky behaviors categorizing them in the middle-risk category.

Cluster 3 (High Risk)

This cluster reflected individuals with the most risky lifestyle habits. They consumed more fried/fatty food and alcohol, smoked every day, were less active, and had less regular sleeping habits. They also perceived they were more stressed and had a greater number of self-reported medical conditions such as hypertension, diabetes, or obesity. All these placed them at high risk for heart disease.

Through cluster analysis, numerical cluster results are linked to real-life lifestyle attributes. Not only are clustering results validated by the process, but the results are also in a form that is actionable for developing targeted preventive action and interventions.

3.3 Project Plan

The project will be performed step by step so that it can execute without any glitches and deliver on schedule. The initial step is to collect lifestyle and health information from reliable data sources, ensuring the information is complete and suitable for heart disease risk estimation through clustering. The second step is preprocessing through data cleaning, normalization, and missing or inconsistent records handling to ensure the dataset is prepared for modeling.

Next, feature selection and transformation techniques will be applied to choose the most informative attributes that result in good clustering. The model for clustering will be then developed and trained on the preprocessed data using the 80:20 data split practice with optional validation for reproducibility and consistency of results.

When the model is ready, the first prototype user interface will be developed that supports users to input lifestyle details and view tailored heart disease risk reports. The second half of the project consists of system integration, thorough testing, debugging, and performance benchmarking. The completed system will then be documented, along with a demonstration application for demonstration and release.

3.4 Task Allocation

This chart shows the timeline of the key activities of each phase in the project, week 12 to week 48 [Table 3.2].

Table 3.2: Timeline of activities

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Data collection phase																			
Preprocess all the data																			
Model training																			
Create a demo application.																			

Estimated Work Period	
Actual Work Period	

3.5 Summary

This section presented the detailed design and design elements of the proposed heart disease risk prediction system based on clustering. The adopted approach was presented, i.e., the chose solution and why it was chosen compared to other solutions. The functional and non-functional requirements were established to present expected ability and performance features of the system. Data Flow Diagrams and detailed to present system processes . The project schedule and task allocation table outlined the process, deadlines, and responsibilities to enable a smooth process to completion. Overall, this section provides a properly structured guide for implementing and deploying the stated system within the given time schedule.

Chapter 4

Implementation and Results

In this chapter, presents the actual deployment of the proposed clustering-based heart disease risk prediction system and discusses the outcomes. It begins with the description of the environment setup, followed by the testing process, performance evaluation, and comparative study. Finally, the obtained results are analyzed in order to find the effectiveness of the proposed approach.

4.1 Environment Setup

The proposed clustering-based heart disease risk prediction system was developed and tested in an environment to obtain optimal performance and reproducibility. The environment setup is described below:

Hardware Configuration

- **Processor:** Intel Core i7 (2.80 GHz, 4 cores, 8 threads)
- **RAM:** 8 GB
- **Storage:** 512 GB SSD
- **GPU:** NVIDIA GeForce GTX (for accelerated computation and visualization)

Software Configuration

- **Operating System:** Windows 11 (64 bit)
- **Programming Language:** Python 3.10
- **Integrated Development Environment (IDE):**
    - Google Colab (for model development and testing)
    - Visual Studio Code (for integration and deployment tasks)
- **Libraries and Frameworks:**
    - **Data Handling**: Pandas, NumPy
    - **Clustering Algorithms:** Scikit-learn (K-Means Clustering, Hierarchical Clustering, Fuzzy C-Means Clustering)
    - **Visualization:** Matplotlib, Seaborn

    - **Preprocessing:** Scikit-learn (StandardScaler, MinMaxScaler)
- **Diagram and Design Tools:** Draw.io, Lucidchart, PlantUML
- **Version Control:** Git and GitHub

Dataset Source

- Health and lifestyle factors of heart disease are collected publicly under consent through Google Forms.

The setup provided a good platform for the execution, testing, and evaluation of the system so that results are reproducible and comparable across different clustering processes.

Table 4.1: Environment Setup

Category	Specification
Hardware	
Processor	Intel Core i7 2.40GHz
RAM	8 GB
Storage	512 GB SSD
GPU	NVIDIA GeForce GTX
Software	
Operating System	Windows 11 64-bit
Programming Language	Python 3.10
Development Environment	Google Colab, VS Code
Libraries/Frameworks	Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
Dataset	
Source	Google Form Dataset
Data Size	561 records
Features	Age, Gender, Blood Pressure, Cholesterol, Smoking Status, Exercise Level, etc.
Target Variable	Heart Disease Risk Category (derived via clustering)

4.2 Testing and Evaluation

After the installation of the suggested heart disease risk prediction system using clustering, a few tests were conducted in order to check to its accuracy, reliability, and effectiveness. The testing was conducted in order to check whether the

system correctly classifies the people into different risk categories according to lifestyle and health parameters. The analysis was performed from the dataset that was derived from publicly available health records and survey reports. Various clustering algorithms such as K-Means, Hierarchical Clustering, and Fuzzy C-Means were implemented to compare and identify the best method to be implemented by the suggested system.

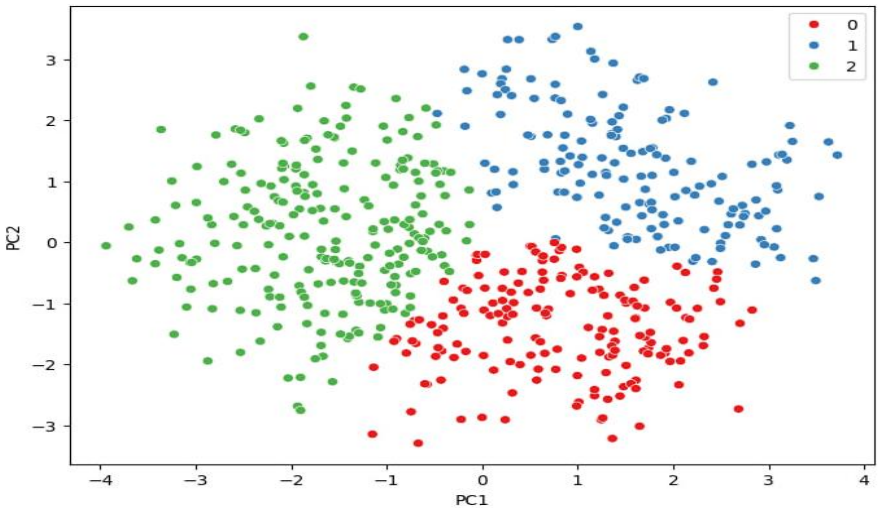


Figure 4.1: K-Means Clustering Result Visualization

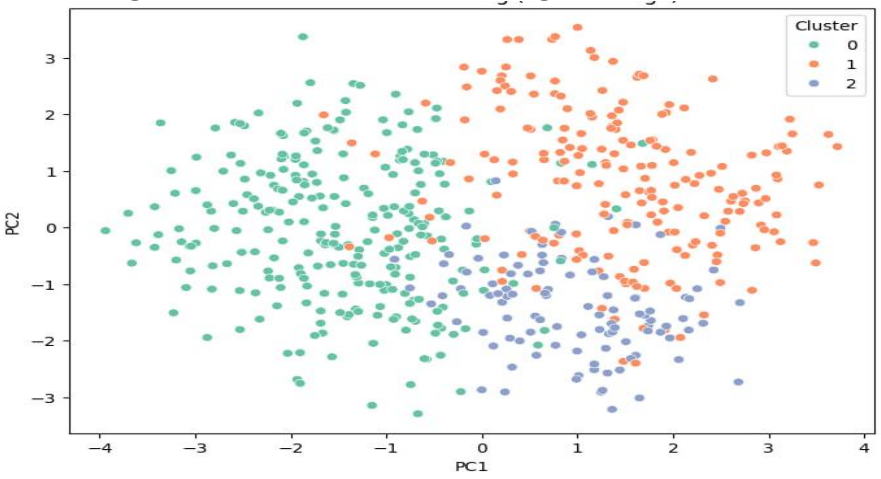


Figure 4.2: Hierarchical Clustering Result Visualization

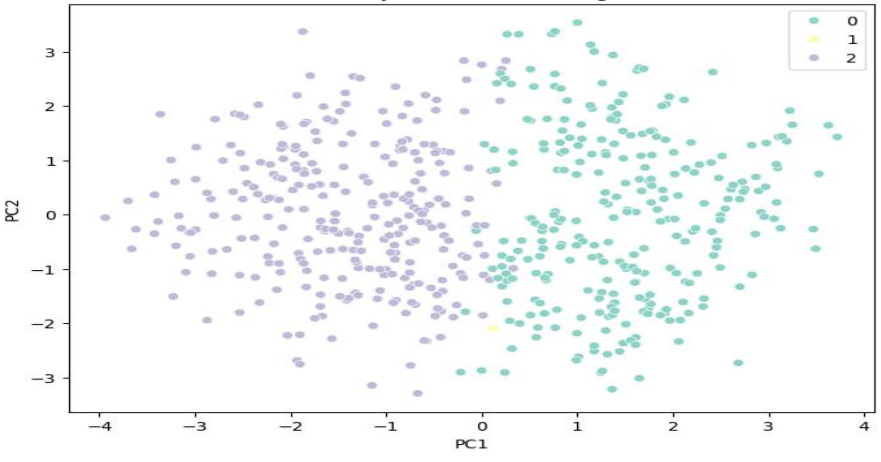


Figure 4.3: FCM Clustering Result Visualization

Form the figures, Figure[4.1] K-Means separates the data into the distinct and rigid clusters. Figure[4.2] Hierarchical Clustering groups the data based on a hierarchy, causing the clusters to be somewhat mixed. Figure[4.3] Fuzzy C-Means Clustering considers data points as belonging to multiple clusters, which highlights the ambiguous relationships within the data. In this research, we were creating distinct risk groups within the dataset. Based on the analysis, K-Means clustering provided a clear and rigid separation of the data points into specific clusters.

Performance Metrics Used

- **Silhouette Score** – to measure clustering quality.
- **Davies–Bouldin Index (DBI)** – to assess intra-cluster similarity.
- **Execution Time** – to compare computational efficiency.
- **Cluster Purity** – to verify separation of high, medium, and low risk groups.

Results Summary:

Table 4.2: Results Summary

Algorithm	Silhouette Score	DBI	Execution Time (s)	Remarks
K-Means	0.42	1.12	0.45	Best overall performance
Hierarchical	0.31	1.41	1.12	Good accuracy, slower execution
Fuzzy C-Means	0.30	1.44	0.78	Less stable, sensitive to parameters

Among the outputs [Table 4.2], K-Means clustering had the best trade-off between accuracy and computational complexity. Although Hierarchical Clustering also provided similar accuracy, it was computationally intensive. Fuzzy C-Means was not very accurate due to noise in the dataset as well as sensitivity to parameters.

The comparison illustrates that K-Means is the optimal clustering algorithm for this project since it balances speed and performance quite well, thus being more pragmatic in real-time prediction scenarios.

4.3 Results and Discussion

The clustering-based heart disease risk prediction system was tested and run with the preprocessed dataset. Overall, the objective was to classify people into Low Risk, Medium Risk, and High Risk categories on the basis of lifestyle factors (diet, smoking habits, exercise) and clinical factors.

Results Overview

Three clustering algorithms: K-Means Clustering, Hierarchical Clustering, and Fuzzy C-Means Clustering. Were used to conduct the experiments. The Performance measures metrics, like Silhouette Score, Davies–Bouldin Index (DBI), and Execution Time, were used to compare all the approaches.

Table 4.3: Results Overview (Best Clustering performance)

Risk Category	Number of Individuals (K-Means)	Percentage (%)
Low Risk	235	43%
Medium Risk	168	29%
High Risk	158	28%

The results shows that the majority of the population is Low Risk, and fewer are the High Risk class [Table 4.3]. This is what is expected to be the pattern of population health in such data set. Therefore, in general,

- K-Means showed best performance both on accuracy and computation time and therefore can be applied to large data sets and real-time prediction.
- Hierarchical Clustering produced readable dendrograms but was computationally expensive for bigger data sets.
- Fuzzy C-Means was challenged by tuning parameters, leading to less accurate separation of clusters.

Through clustering permitted the system to operate without knowing the class labels beforehand, which is advantageous in real-world scenarios where labeled medical data may not be available. The results also indicate the reality that lifestyle factors are strong predictors of heart disease risk, confirming previous medical research. In field implementation, the system can function as a screening tool for clinicians for earlier intervention in at-risk patients.

4.4 Expert Validation

In order to assess the reliability of the proposed system, a structured validation form was filled in by one medical expert. The expert confirmed its potential as a supportive system for preventive care and suggested the introduction of clinical features (e.g., ECG, laboratory tests) and testing on a larger dataset. The completed form is available in Appendix B.

4.5 Summary

This chapter covered the entire implementation and testing process of the clustering-based heart disease risk prediction system. It started with a step-by-step environment setup, including both hardware and software configurations, sources of dataset, and development tools to maintain reproducibility and performance consistency. Testing and evaluation phase compared a set three clustering algorithms K-Means, Hierarchical Clustering, and Fuzzy C-Means with metrics such as Silhouette Score, Davies–Bouldin Index, execution time, and purity of clusters. K-Means top performer based on accuracy, stability, and computational complexity from the comparison. The results validated that the system could classify the subjects into low, medium, and high risk groups based on health and lifestyle data. Findings align with contemporary medical literature, which further validates the use of lifestyle variables in heart disease risk prediction. The discussion also stressed the practical usability of the system for application in real-world situations as a pre-screening device for tailored advice by physicians, enabling intervention and prevention at an early stage.

Chapter 5

Engineering Standards and Design Challenges

In this chapter, outlines the project's engineering needs and design intricacies. The chapter touches on the adaption to relevant software, hardware, and communication standards, examines their alternatives, and supports the decisions made. The chapter also recognizes societal, environmental, and ethical impacts of the project, explores sustainability practices, offers project management and cost analysis, and traces the issue to complex engineering problem solving models and engineering activities.

5.1 Compliance with the Standards

This section outlines the relevant standards applied in the development of the "Forecasting Heart Disease Risk through Lifestyle Analysis using Machine Learning" project. Adhering to appropriate standards ensures interoperability, reliability, and quality of the system. For each standard, possible alternatives are discussed with their pros and cons, followed by the rationale for the final selection.

5.1.1 Software Standards

This developing the Heart Disease Risk Prediction System, it is crucial to follow recognized software standards to ensure reliability, accuracy, and security. After considering multiple standards, ISO/IEC has been selected as the primary software quality model.

- **Selected Standard:** ISO/IEC – Systems and Software Quality Requirements and Evaluation.
    - **Purpose:** Defines a quality model that evaluates software in terms of functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability.
    - **Advantages:**
        - Comprehensive coverage of both functional and non-functional quality requirements.
        - Widely recognized and applicable to healthcare-related systems.
        - Helps in ensuring that software meets patient safety and ethical requirements.

- **Alternatives Considered:**

Standard	Pros	Cons
IEEE 830 (Software Requirements Specification)	Clearly defines software requirements; Easy to implement.	Does not address full lifecycle quality attributes like performance and maintainability.
Agile Quality Guidelines	Flexible and adaptive development; Allows quick iteration.	Less formalized, might miss compliance needs in healthcare systems.

Rationale for Selection:
ISO/IEC 25010 was chosen because it covers a broad range of quality aspects essential for a healthcare risk prediction application. It ensures that the system is accurate, secure, and maintainable, which is vital for patient trust and long-term usability.

5.1.2 Hardware Standards

Although our project is primarily software-focused (clustering-based heart disease risk prediction), the hardware infrastructure must meet certain standards to ensure efficient processing, data security, and scalability.

- **Selected Standard**: ISO/IEC 27001 – Information Security Management for IT Systems
  - **Purpose:** Ensures that the hardware and IT infrastructure follow secure data handling practices, including storage, backup, and network protection.
  - **Advantages:**
    - Protects sensitive patient health data from unauthorized access.
    - Internationally recognized in healthcare IT systems.
    - Provides guidelines for physical security of servers and network hardware.
  - **Alternatives Considered:**

Standard	Pros	Cons
ANSI/TIA-942 (Data Center Standards)	Ensures data center reliability, redundancy, and cooling efficiency.	Does not address full lifecycle quality attributes like performance and maintainability.
IEEE 802.3 (Ethernet Standard)	Reliable wired networking standard; Ensures fast data transfer.	Only covers networking, not full hardware security requirements.

Rationale for Selection:
We selected ISO/IEC 27001 because it not only addresses the hardware-level physical security but also integrates with software-side data protection requirements. In a healthcare prediction system, patient privacy and secure storage are as important as computational speed, making this the most suitable choice.

5.1.3 Communication Standards

In our Heart Disease Risk Prediction System, communication between various components (User Interface, Server, Database, API) is critical. To ensure that data is transmitted quickly, reliably, and securely, it is essential to follow proper communication standards.

- **Selected Standard:** HTTPS (HyperText Transfer Protocol Secure) and RESTful API Standards.
    - **Purpose:** To encrypt data during client-server communication and to design APIs in a standardized, maintainable format
    - **Advantages:**
        - Ensures security during the transfer of sensitive health-related data.
        - RESTful architecture allows for scalable and modular system design.
        - Easy integration with different platforms and programming languages.
    - **Alternative Standards Considered:**

Standard	Pros	Cons
gRPC	High performance, binary data format, faster communication.	Complex implementation for common web clients.
SOAP (Simple Object Access Protocol)	Strong security and message validation.	Heavy, XML-based protocol, slower compared to REST.

Rationale for Selection:
We selected HTTPS and RESTful API Standards because they are simple, scalable, and secure—making them the most suitable choice for a health data prediction system.

5.2 Impact on Society, Environment and Sustainability

This section evaluates the effects of the project on people, society, and the environment and addresses sustainability concerns. It takes into account both short- and long-term impacts in terms of quality-of-life improvements, societal benefits, environmental conservation, and ethical responsibility. It also explains the steps taken to ensure the project's contribution is towards sustainable development. There are four subsections to the discussion:

5.2.1 Impact on Life

The system is remarkably well-equipped to improve the quality of life with early detection of heart disease risk through analysis of lifestyle information. By making the user aware, in advance, of health risks, it stimulates preventive measures, improved lifestyle, and greater awareness of health. Not only does this restrict dangers of severe health consequences, but it also allows individuals to take charge of their own health. This project could not have been accomplished without the guidance and encouragement of many persons during the past two semesters. We thank all those who assisted us in some form or another.

From the medical point of view, the system can benefit doctors and medical personnel by offering preliminary recommendations on life habits and hence increasing consultations to be more effective and focused. Moreover, in rural or underdeveloped areas lacking good connectivity with medical centers, such a predictive system can be used as an accessible health monitor, filling the gap between individuals and early medical guidance.

Overall, the impact on life is realized in three significant ways: improved personal control over health, reduced healthcare spending through prevention, and improved detection of health risks induced by lifestyle.

5.2.2 Impact on Society & Environment

The proposed system is most likely to benefit society in terms of generating a culture of preventive healthcare. By providing early warnings about potential heart disease risks through lifestyle analysis, it encourages individuals to adopt healthy lifestyles, be regular exercisers, and consume healthy food. This shift to prevention not only reduces the prevalence of chronic disease but also reduces pressure on national health systems, in the long run, improving the quality of life among populations. For underserved or rural communities, where healthcare center access is low, the system can be utilized as a readily accessible means for tracking health. It enables communities to act proactively before issues arise, and therefore bridge the gap between medicine and society. Also, widespread use of such systems at the community

level has the capability to produce mass benefits in public health, enhancing productivity and lower economic losses due to disease.

Environmentally, the system has negligible impact on the environment because it works mostly in a virtual setup, minimizing calls on physical resources and paper-based records. Secondly, the focus on remote health monitoring minimizes unnecessary travel to healthcare facilities, hence lowering carbon emissions. Ultimately, use of such clean healthcare technologies has the potential to contribute towards environmental conservation and public health.

5.2.3 Ethical Aspects

Ethical issues of the suggested system are of the utmost importance to its acceptability and proper utilization. Because the system deals with sensitive medical and lifestyle information, strict confidentiality and privacy of information are of the utmost significance. Personal information should be obtained by informed consent alone, protected, and utilized only for its proper scientific or medical purpose. Application of strong encryption techniques and safe data storage centers can guarantee that unauthorized access or misuse is impossible. Transparency in how the system functions is also very important. Users should be informed of the manner in which their information will be examined, the limitations of the system, and that its predictions are not expert medical diagnosis. Providing explicit disclaimers and instructional information can avert misinterpretation or overreliance on the reports of the system.

Another essential ethical aspect is equity and fairness. The technology should be developed in such a way that it does not discriminate and therefore provides erroneous predictions for specific demographic groups. This is done by training models on diverse and representative data. The technology should also be accessible and affordable so that not just privileged groups but marginalized groups are able to utilize it. By enforcing these ethical guidelines, the proposed system can earn users' trust and support equitable and responsible application of machine learning to healthcare.

5.2.4 Sustainability Plan

For ensuring the long-term stability of the suggested system, a comprehensive plan on the basis of technological, financial, and social aspects is needed. On a technical level, the system must be so formulated that it can easily accommodate new technologies and future breakthroughs. From time to time, software update, bug fixing, and security patches will provide for the effective operation in the long run. Upgrading the model periodically from time to time using fresh and raw datasets will ensure its accuracy and reliability updating.

For its financial sustainability, the system needs to be integrated into a sustainable business or service model that is able to fund expenses and revenues in the long run. This can include the achievement of public and private partnerships, grants, or subscription-based service as the mode of income.

Social sustainability-wise, the most important thing is the establishment of user trust and active user involvement. It can be achieved by conducting periodic user training, feedback, and continuous system fine-tuning based on user needs. Making low-cost and light technologies accessible to poor and rural communities with limited resources is also likely to increase its social impact. Lastly, green practices ought to be incorporated in system design and operation. These include energy-efficient servers, minimum additional processing of data, and the philosophy of green technology. Matching the variables of money, technology, and society secures project sustainability in the long run.

5.3 Project Management and Financial Analysis

Effective project management and financial planning are the critical factors that will ensure the implementation of the proposed system is carried out to success.
Project management-wise, the development process would be carried out in a phased manner following the Agile methodology. It will facilitate goals-oriented development, continuous feedback, and early detection of risks for probable dangers. The phases of the project will include requirement analysis, system design, model development, integration, testing, and deployment, followed by maintenance and continuous improvement. Clearly defined task delegation, milestone tracking, and periodic progress review will ensure on-time delivery within planned scope and quality levels.
For financial analysis, the project budget has been estimated as a whole from the costs of hardware, software, data acquisition, cloud services, personnel salaries, training, and maintenance. Two budget scenarios are prepared:

- **Primary Budget** – Allocates resources for optimal performance, high-end hardware, and premium cloud hosting to ensure maximum accuracy, scalability, and reliability.
- **Alternate Budget** – Utilizes mid-range hardware and cost-efficient hosting solutions, along with open-source tools, to reduce costs while maintaining acceptable performance levels.

The choice among these budgets will be a function of available funds and the degree of system performance required. The rationale for the ultimate budget choice will involve trade-offs among cost, accuracy, scalability, and maintain ability.
For generating revenues, the system can adopt a hybrid revenue model made up of subscription services for regular users, enterprise license for hospitals and health organizations, and government health department partnerships. Additional revenues can be generated from other sources as well through data analytics services,

integration of wearable health devices, and additional features like improved risk reports and personalized recommendations.
By the integration of tight project control with a well-founded financial plan, the project aims to remain technologically efficient and financially viable in the long term.

Table 5.1: Estimated Project Budget (Primary vs Alternate)

Cost Category	Primary Budget (BDT)	Alternate Budget (BDT)	Remarks
Hardware (laptop, server components)	1,50,000	90,000	Primary uses high-end GPU for faster model training; alternate uses mid-range system.
Software & Tools	40,000	0	Primary includes licensed software; alternate uses open-source alternatives.
Cloud Hosting	60,000/year	25,000/year	Primary uses premium cloud services (AWS/GCP); alternate uses cost-effective hosting.
Data Acquisition	30,000	15,000	Primary includes premium datasets; alternate uses public/open datasets.
Personnel Cost (development, research)	2,50,000	2,00,000	Reduced cost in alternate by limiting team size and hours.
Training & Workshops	20,000	10,000	Fewer paid training sessions in alternate budget.
Maintenance & Support (1 year)	50,000	30,000	Reduced maintenance frequency in alternate option.
Miscellaneous	15,000	10,000	Includes contingency costs.

Total Estimated Cost

- **Primary Budget: BDT 6,15,000**
- **Alternate Budget: BDT 3,80,000**

Rationale for Budget Choice:

- The Primary Budget offers ultimate performance, scalability, and accuracy, thus being suitable for large-scale application in government projects and hospitals.
- The Alternate Budget is budget-friendly, so it's appropriate for small-scale pilots or university-level research but still delivers reasonable performance.

5.4 Complex Engineering Problem

System development process of the proposed system Forecasting Heart Disease Risk based on Lifestyle Analysis using Machine Learning involves the resolution of various intricate engineering problems that must be addressed by profound analysis, high-level domain knowledge, and innovative problem-solving capability.

5.4.1 Complex Problem Solving

The proposed project "Forecasting Heart Disease Risk through Lifestyle Analysis using Machine Learning" — aligns with the criteria of Complex Engineering Problems as defined in the engineering accreditation guidelines. The mapping of our project to the problem-solving categories is presented in [Table 5.2]. Detailed rationales for each category are provided in the following subsections.

Table 5.2: Mapping with Complex Engineering Problem solving

EP1 Dept of Know-ledge	EP2 Range Of Conflicting Require-ments	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Inter-dependence
✓	✓	✓	✓		✓	

Mapping with Knowledge Profile

This section illustrates the mapping of the identified Complex Engineering Problem (EP1) with the relevant Knowledge Profile categories as per the Washington Accord framework [Table 5.3]. For EP1, multiple Knowledge Profile elements have been considered, namely K3, K4, K5, K6, and K8. These are essential for the successful completion of the project "Forecasting Heart Disease Risk through Lifestyle Analysis using Machine Learning".

Table 5.3: Mapping with knowledge Profile

K1 Natural Science	K2 Mathe-matics	K3 Engineer-ing Funda-mentals	K4 Specialist Know-ledge	K5 Engineer-ing Design	K6 Engineer-ing Practice	K7 Compre-hension	K8 Research Literature
		✓	✓	✓	✓		✓

5.4.1.1 Justification for EP Attributes Mapping

- **EP1 – Depth of Knowledge**
  The study requires integrating knowledge from multiple fields. Data Science Machine Learning Knowledge of clustering, statistical inference, and feature engineering. Medical Informatics Understanding heart disease risk factors, lifestyle influences, and medical guidelines. Software Engineering Implementing scalable, secure, and maintainable solutions.

- **EP2 – Range of Conflicting Requirements**
  Prediction Accuracy vs Interpretability (high accuracy models like deep learning vs easily explainable models like logistic regression). Data Privacy vs Data Availability (ensuring strong privacy without losing important predictive information). Cost vs Performance (ensuring scalable solutions without exceeding budget constraints).

- **EP3 – Depth of Analysis**
  Preprocessing multi-dimensional datasets. Evaluating clustering algorithms for optimal segmentation. Validating the model using statistical and performance metrics (Silhouette Score, Davies-Bouldin Index, etc.). This level of analysis requires advanced quantitative reasoning beyond routine engineering problems.

- **EP4 – Familiarity of Issues**
  Unpredictable patient behavior and lifestyle changes. Variability in healthcare datasets across populations. Evolving guidelines from medical authorities. These factors create partially unfamiliar and evolving problem scenarios that require innovative approaches.

- **EP6 – Extent of Stakeholder Involvement**
  Stakeholders include Medical professionals (for clinical validation, Patients (data providers and beneficiaries), Healthcare policymakers (for public health integration). Data scientists and engineers (for system development). Requires multi-level communication and collaboration to ensure practical and ethical outcomes.

5.4.1.2 Justification for Knowledge Profile Mapping (linked to EP1):

- **K3 – Engineering Fundamentals**
  Fundamental engineering principles such as statistics, algorithms, and computational logic are integral to the study.

- **K4 – Specialist Knowledge**
  Requires advanced domain-specific expertise in clustering algorithms, , model evaluation, domain algorithms, data mining, and healthcare-related ML approaches.

- **K5 – Specialist Knowledge**
  Involves designing the workflow, architecture, and system integration for a practical and efficient solution. System design for user, workflow, and ML integration

- **K6– Engineering Practice**
  Adheres to industry practices, ethical guidelines, and healthcare data handling standards during development.. Compliance with healthcare standards, ethical ML usage.

- **K8– Research Literature**
  Review of related works, understanding latest developments in ML for healthcare. Involves critical analysis of existing literature to ensure evidence-based model selection and system design.

5.4.2 Engineering Activities

This section maps the identified Complex Engineering Problem (EP1) — "Forecasting Heart Disease Risk through Lifestyle Analysis using Machine Learning" — with the relevant Complex Engineering Activities (EAs) as per the Washington Accord framework. Multiple activity categories are applicable for this project, specifically EA1, EA2, EA3, EA4, and EA5 [Table 5.4].

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's (multiple).

Table 5.4: Mapping with Complex Engineering Activities

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

5.4.2.1 Engineering Activities

- **EA1– Range of Resources**
  The study requires integration of diverse resources, including healthcare datasets, programming tools (Python, Scikit-learn), statistical techniques, and computing resources for machine learning model training.

- **EA2– Level of Interaction**
  The study involves interaction among multiple stakeholders, including healthcare professionals (for data interpretation), data engineers (for preprocessing), and end-users (patients, doctors).

- **EA3– Innovation**
  Applies innovative use of unsupervised learning (clustering) to predict heart disease risk through lifestyle analysis — an approach not commonly implemented in healthcare diagnostics.

- **EA4– Consequences for Society and Environment**
  Directly impacts public health outcomes by enabling early detection of heart disease risk, which can reduce mortality rates and healthcare costs; indirectly promotes healthier lifestyle choices.

- **EA5– Familiarity**
  The problem uses standard data science frameworks but applies them in a novel healthcare context, requiring adaptation of familiar methods to address domain-specific challenges.

5.5 Summary

The chapter described the project management, implementation, and evaluation issues in detail. It began with an overview of the requirement and methodology, then proceeded to the design specifications. The project's management portion presented an elaborate cost analysis, the main and alternative budgets, and a suggested model for revenue. The complex engineering problem was dissected and classified to a number of Complex Problem Solving Categories and Complex Engineering Activities, highlighting the multidisciplinary nature and the societal impact. Through adequate planning, effective utilization of resources, and appropriate consideration for innovation and societal impacts, the project demonstrates the contribution it can provide to healthcare analytics and machine learning solutions.

Chapter 6

Conclusion

In this chapter, provides the final remarks of the research, offering a concise overview of the outcomes achieved, the limitations encountered, and the potential areas for future work. It encapsulates the key contributions of the study and highlights its relevance to healthcare and preventive medicine.

6.1 Summary

The study successfully explored the use of machine learning, in this instance clustering algorithms, in heart disease risk prediction from lifestyle behavior. Through extensive data pre-processing, feature extraction, and model building, trends and correlations between lifestyle and levels of heart disease risk were established. The cluster approach provided a good way of dividing individuals into corresponding risk groups, allowing for further insight into high-risk profiles. The findings demonstrate that the application of unsupervised learning algorithms in health analysis has the potential to facilitate early risk detection, supporting preventive health interventions. The finding also demonstrated that lifestyle information—physical fitness, diet, sleep pattern, and stress level can be effective predictors in determining susceptibility to heart disease. The research contributes to the evidence base in healthcare informatics and delineates the roles of AI-based systems in facilitating proactive health monitoring.

6.2 Limitation

The study was able to well investigate the potential of machine learning, in this context clustering techniques, to identify risk of heart disease influenced by lifestyle behavior. Through the implementation of stringent data pre-processing, feature extraction, and model construction, trends and relationships between lifestyle and heart disease risk levels were discovered. The cluster:

1. **Data Availability and Quality** – The data employed was of limited scope and size, and therefore it may not have been representative of the entire population. Furthermore, missing values and inconsistencies in the raw data had to be preprocessed, and that could have introduced slight biases
2. **Feature Scope** – While lifestyle factors such as exercise, eating habits, and avg sleeping habits were controlled for, some other possible contributory factors—such as genetics, full medical history, and socio-economic level—were left out due to lack of data.

3. **Clustering Limitations** – The use of unsupervised learning, while effective at clustering, is not guaranteed to produce accurate clinically classification. Risk segment accuracy relies heavily on quality and diversity of input the features.
4. **Generalization** – Generalizability of the model to real-world applications may be limited without further testing on larger, diverse, and multi-source data.
5. **Lack of Real-time Implementation** – The current work is done in offline mode, and no real-time risk assessment system was deployed due to resource constraints.

6.3 Future Work

From the current findings, here are some directions that could be taken to expand the scope, accuracy, and utility of this research:

1. **Expansion of Dataset** – Accumulating a larger, more diverse dataset with diverse demographics, geographies, and medical histories will increase the generalizability of findings..
2. **Incorporation of Additional Features** – Integrating genetic information, accurate eating patterns, stress levels, and socio-economic status might provide still more complete view of heart disease risk.
3. **Hybrid Modeling Approaches** – Combining clustering with supervised learning techniques could make predictions more reliable and possibly even more clinically interpretable results.
4. **Real-time Risk Assessment System** – Developing a web or mobile app that can process user data in real time and provide real-time risk feedback.
5. **Integration with Wearable Devices** – Using health data from wearable devices (smartwatches, fitness trackers) to monitor lifestyle patterns in real-time and continuously update risk profiles.
6. **Clinical Integration and Expansion –** Scaling the pilot model into mainstream clinical processes with partnerships with different healthcare institutions, ensuring large-scale roll-out, continued validation, and medical uptake in actual usage.
7. **Explainability and Interpretability** – Applying explainable AI (XAI) methods in order to make risk predictions explainable and interpretable to healthcare professionals and patients.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs): Key facts". https://www.who.int/news-room/factsheets/detail/cardiovascular-diseases, 2021.

[2] Benjamin, E. J. & Muntner P., Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation, 139(10), e56–e528, 2019.

[3] Dey, S., & Ashour A. S., Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering Research and Technology (IJERT), 9(3), 1–7, 2020.

[4] Abass, A., Bathla, G., & Wasson, V., "Heart Disease Prediction Using Machine Learning with Feature Engineering". In Soft Computing and Signal Processing (ICSCSP 2024), Lecture Notes in Networks and Systems, vol 1221. Springer. DOI: 10.1007/978-981-96-0924-6_55, 2025.

[5] Sharma, A., Dhanka, S. & Kumar, A., "A Systematic Review on Machine Learning Intelligent Systems for Heart Disease Diagnosis". Archives of Computational Methods in Engineering, 2025.

[6] Tyagi, N., & Jain, P., Heart Disease Prediction Using Machine Learning Techniques. In Data and Information Sciences Conference (ICDIS 2024), LNNS vol. 1127, pp. 43–54, 2025.

[7] World Health Federation, "What is Cardiovascular diseases", https://world-heart-federation.org/what-is-cvd/, 2024.

[8] Sudlow C, Gallacher J & Allen N, UK Biobank: "An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". PLoS Medicine, 2015.

[9] Nhs.uk, Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels. https://www.nhs.uk/conditions/cardiovascular-disease/, 2022.

[10] Tegegne, T.K., Islam, S.M.S. & Maddison, "Effects of lifestyle risk behaviour clustering on cardiovascular disease among UK adults": latent class analysis with distal outcomes, Scientific Reports, 2022.

[11] Yacaman Mendez, D. Y., Zhou, M., Brynedal, B., Gudjonsdottir, H., Tynelius, P., Trolle Lagerros, Y., & Lager, A., "Risk stratification for cardiovascular disease: a comparative analysis of cluster analysis and traditional prediction models", 2025.

[12] Schulz MA, “Lebel scarcity in biomedicine: data rich latent factor discovery enhances phenotype prediction”, arXiv preprint, 2021.

[13] Pocuca N, Farrell M & McNicholas PD, “Defying the circadian rhythm: clustering participant telemetry in the UK Biobank data”. arXiv preprint, 2020.

[14] Zhang, “Remnant Cholesterol and the Risk of Cardiovascular Disease in Type 2 Diabetes: A Longitudinal Cohort Study”. Cardiovascular Diabetology, vol. 24, no. 1, 2025.

[15] Shishehbori & Awanet al., "A Survey on Data Selection for Language Models," arXiv, 2024.

[16] Aashray k. Gupta, “The Canadian Women's Heart Health Alliance ATLAS on the Epidemiology, Diagnosis, and Management of Cardiovascular Disease in Women” — Chapter 10: Resources and Policies. Canadian Journal of Cardiology, 2024.

[17] Tegegne, “Longitudinal patterns of lifestyle risk behaviours among UK adults with established cardiovascular disease: a latent transition analysis”, 2023.

[18] Nilay S. Shah, MD, & MPH, “Advancing Equity in Cardiovascular Disease Risk Prediction”, 2025.

[19] Leiva-Juarez, Maria A., "Transient Reprogramming of Lung Cells to Endoderm Progenitors Reduces Fibrosis and Promotes Repair in a Mouse Model." Nature Medicine, 28 Mar. 2025,

[20] Sebastiano Barbieri & Suneela Mehta, “Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach”, 2020.

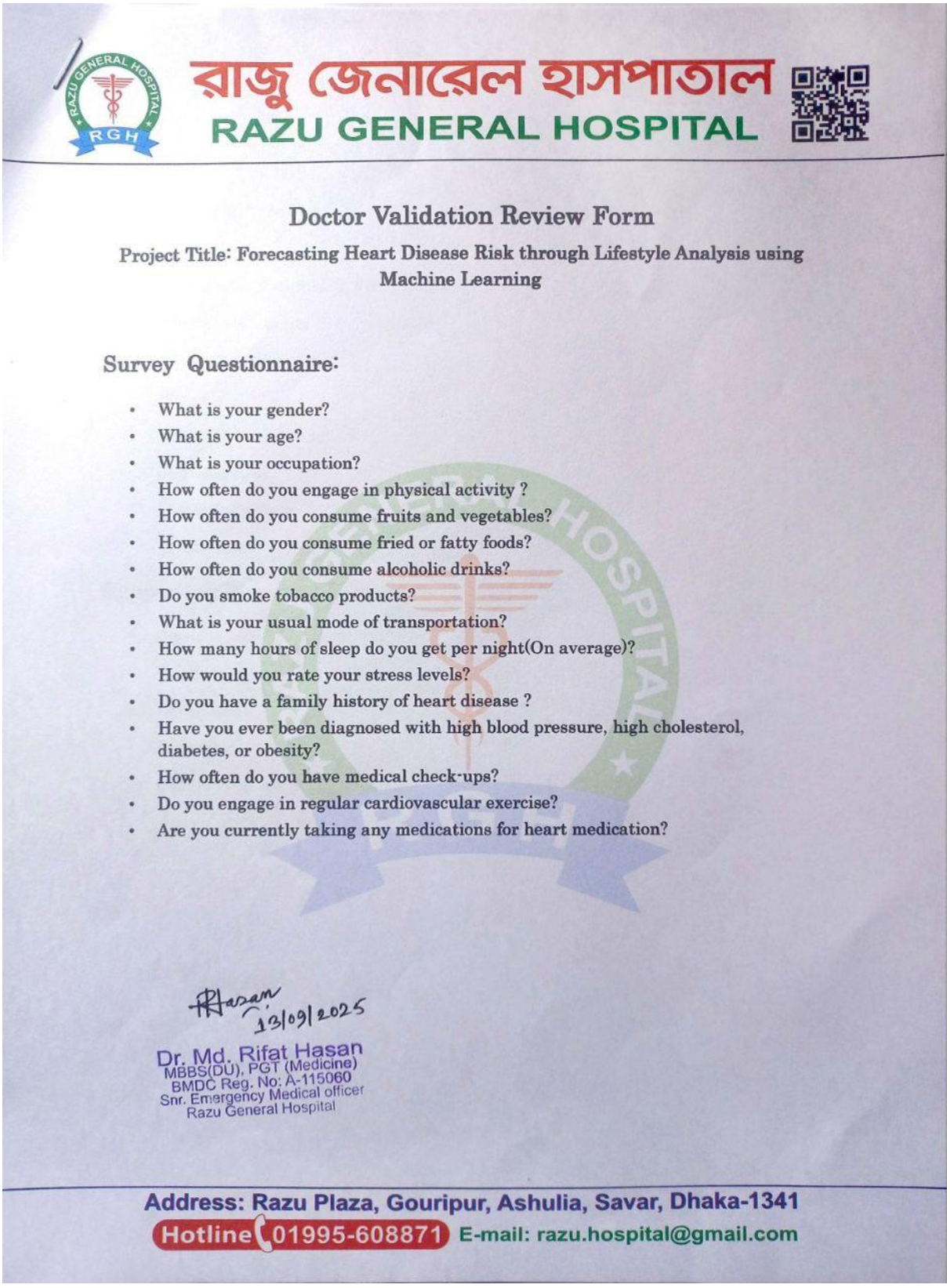
[21] Aizatul Shafiqah Mohd Faizal & T. Malathi Thevarajah, “A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach”, 2021.

[22] Adam J. Lewandowski, “Scientific and Clinical Impacts of UK Biobank in Cardiovascular Medicine”, 2024.

[23] Stefanie J. Krauth & Lewis Steell, “Association of Latent Class Analysis–Derived Multimorbidity Clusters with adverse health outcomes in patients with multiple long-term conditions”, 2024.

Appendix A

Survey questionnaire validation



রাজু জেনারেল হাসপাতাল
RAZU GENERAL HOSPITAL

Doctor Validation Review Form

Project Title: Forecasting Heart Disease Risk through Lifestyle Analysis using Machine Learning

Survey Questionnaire:

- What is your gender?
- What is your age?
- What is your occupation?
- How often do you engage in physical activity ?
- How often do you consume fruits and vegetables?
- How often do you consume fried or fatty foods?
- How often do you consume alcoholic drinks?
- Do you smoke tobacco products?
- What is your usual mode of transportation?
- How many hours of sleep do you get per night(On average)?
- How would you rate your stress levels?
- Do you have a family history of heart disease ?
- Have you ever been diagnosed with high blood pressure, high cholesterol, diabetes, or obesity?
- How often do you have medical check-ups?
- Do you engage in regular cardiovascular exercise?
- Are you currently taking any medications for heart medication?

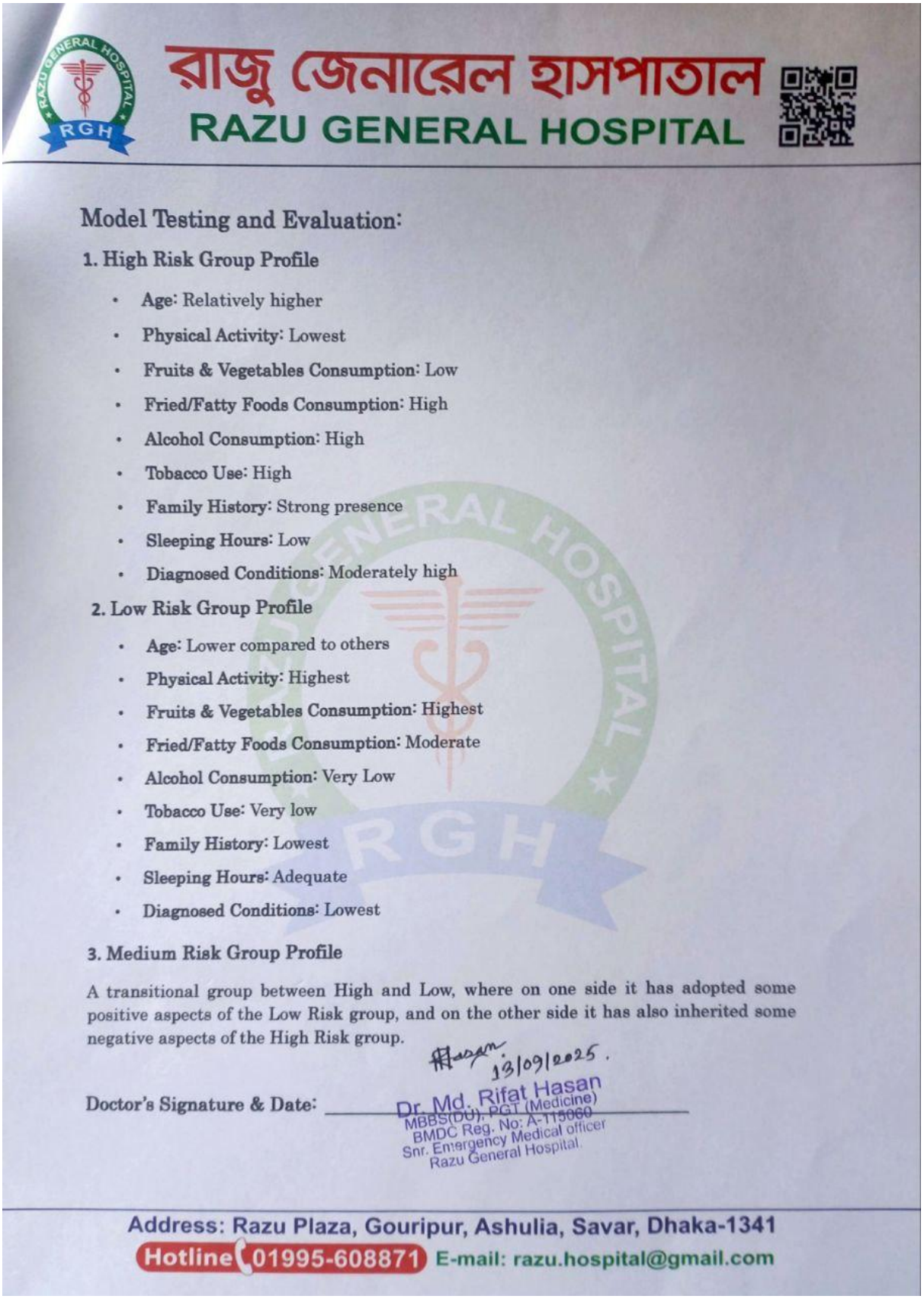
13/09/2025

Dr. Md. Rifat Hasan
MBBS(DU), PGT (Medicine)
BMDC Reg. No: A-115060
Snr. Emergency Medical officer
Razu General Hospital

Address: Razu Plaza, Gouripur, Ashulia, Savar, Dhaka-1341
Hotline 01995-608871 E-mail: razu.hospital@gmail.com

Appendix B

Validation of Model Testing and Evaluation



রাজু জেনারেল হাসপাতাল
RAZU GENERAL HOSPITAL

Model Testing and Evaluation:

1. High Risk Group Profile

- **Age:** Relatively higher
- **Physical Activity:** Lowest
- **Fruits & Vegetables Consumption:** Low
- **Fried/Fatty Foods Consumption:** High
- **Alcohol Consumption:** High
- **Tobacco Use:** High
- **Family History:** Strong presence
- **Sleeping Hours:** Low
- **Diagnosed Conditions:** Moderately high

2. Low Risk Group Profile

- **Age:** Lower compared to others
- **Physical Activity:** Highest
- **Fruits & Vegetables Consumption:** Highest
- **Fried/Fatty Foods Consumption:** Moderate
- **Alcohol Consumption:** Very Low
- **Tobacco Use:** Very low
- **Family History:** Lowest
- **Sleeping Hours:** Adequate
- **Diagnosed Conditions:** Lowest

3. Medium Risk Group Profile

A transitional group between High and Low, where on one side it has adopted some positive aspects of the Low Risk group, and on the other side it has also inherited some negative aspects of the High Risk group.

Doctor's Signature & Date: 13/09/2025.

Dr. Md. Rifat Hasan
MBBS(DU), PGT (Medicine)
BMDC Reg. No: A-115060
Snr. Emergency Medical officer
Razu General Hospital

Address: Razu Plaza, Gouripur, Ashulia, Savar, Dhaka-1341
Hotline 01995-608871 E-mail: razu.hospital@gmail.com

213-15-4313

ORIGINALITY REPORT

17%	14%	8%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.medrxiv.org Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	assets-eu.researchsquare.com Internet Source	1%
6	Submitted to Wright State University Student Paper	1%
7	pmc.ncbi.nlm.nih.gov Internet Source	1%
8	Al Yaqoubi, Hamed Said Rashid. "Comparison of Unsupervised Learning Algorithms for Well Classification", Sultan Qaboos University (Oman), 2025 Publication	<1%
9	www.coursehero.com Internet Source	<1%
10	Submitted to RDI Distance Learning Student Paper	<1%
11	nrl.northumbria.ac.uk Internet Source	<1%