Cardiovascular Diseases Risk In People with Mental Illnesses

A Thesis Submitted for the Degree of Master of Computer Science

P. G. N. S Indika

University of Colombo School of Computing

2024

Declaration

Name of the student: Pahala Gamage Nidu Sihani Indika	
Registration number: 2020mcs034	
Name of the Degree Programme: Master of Computer Science	
Project/Thesis title: Cardiovascular Diseases Risk In People with Mental Illnesses	

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	**Date (DD/MM/YYYY)**
	26/10/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	**Supervisor 2**	**Supervisor 3**
Name	Prof. G.K.A.Dias		
Signature			
Date	26/10/2024		

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Professor G.K.A. Dias, my supervisor, whose guidance, expertise, and unwavering support have been invaluable throughout the course of this study. His insightful feedback and encouragement have played an important role in shaping the direction and execution of this research.

I am also deeply grateful to Dr. Kasun Karunanayake, our project coordinator, for his assistance and valuable insights throughout the duration of this project.

Furthermore, I extend my heartfelt thanks to the university staff and the Postgraduate Division for their support and assistance throughout this study.

I would also like to acknowledge the support and understanding of my family and friends, whose unwavering encouragement and patience have been a constant source of strength throughout this endeavor.

Lastly, I extend my appreciation to all the authors, researchers, and institutions whose work has paved the way for advancements in the fields of mental health and cardiovascular diseases, upon which this study builds.

This research would not have been possible without the collective efforts and support of all those mentioned above, and for that, I am sincerely thankful.

ABSTRACT

Cardiovascular diseases (CVD) pose a significant global health burden, with various types such as coronary heart disease, stroke, peripheral arterial disease, and aortic disease contributing to mortality and morbidity worldwide. Evidence suggested a strong association between mental health disorders and CVD, wherein conditions like mood disorders, anxiety, PTSD, and chronic stress contributed to increased risk and poorer outcomes. Conversely, CVD events could also precipitate mental health disorders, creating a complex interplay between the two domains. This study aimed to explore this relationship, identify common risk factors, and develop a predictive system for assessing CVD risk among individuals with mental illnesses. Through a review of relevant literature, the study examined prevalence rates, shared risk factors, and the impact of mental health disorders on CVD management and prognosis. Utilizing machine learning techniques, a decision support web-based system was constructed to predict CVD risk factors based on patients' mental health histories. While the scope included data collection and algorithm development, the system did not offer medical consultancy services. By illuminating the nexus between mental health and CVD, this research sought to enhance risk assessment and inform preventive interventions for vulnerable populations.

This study aimed to explore this relationship, identify common risk factors, and develop a predictive system for assessing CVD risk among individuals with mental illnesses. Through a review of relevant literature, the study examined prevalence rates, shared risk factors, and the impact of mental health disorders on CVD management and prognosis.

Utilizing machine learning techniques, a decision support web-based system was constructed to predict CVD risk factors based on patients' mental health histories. Machine learning algorithms analyzed comprehensive datasets to identify patterns and correlations between mental health conditions and CVD risk factors. These models were trained to recognize how specific mental health disorders and their severities influence the likelihood of developing CVD.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.......... ii
ABSTRACT.......... iii
TABLE OF CONTENTS.......... iv
LIST OF FIGURES.......... vi
LIST OF TABLES.......... vii
LIST OF ABBREVIATIONS.......... ix
CHAPTER 1 INTRODUCTION.......... 1
1.1 Statement of the problem.......... 1
1.2 Project Aim and Objectives.......... 2
1.3 Scope.......... 3
CHAPTER 2 LITERATURE REVIEW.......... 4
2.1 Related Work/ Background Study.......... 4
2.2 Novelty of the Research and Expected Research Contribution to the Field.......... 7
CHAPTER 3 METHODOLOGY.......... 8
3.1 Research Philosophy.......... 8
3.2 Research Design.......... 8
3.3 Dataset.......... 8
3.4 Machine Learning Model.......... 12
3.5 Web Based System.......... 32
3.6 Ethical Considerations.......... 33
3.7 Limitations.......... 33
CHAPTER 4 EVALUATION AND RESULTS.......... 34
4.1 Evaluation Approach of Machine Learning Model.......... 34
4.2 Web Based System Validation and Testing.......... 46
CHAPTER 5 CONCLUSION AND FUTURE WORK.......... 47
5.1 Conclusion.......... 47
5.2 Future Work.......... 47
5.3 Deliverables.......... 47
REFERENCES.......... 48
APPENDICES.......... 50
APPENDIX A: Data obtained from the Personal Key Indicators of HeartDiseases dataset... 50
APPENDIX B: Data obtained from the Questionnaire.......... 52
APPENDIX C: UI of the Web Based System.......... 59

LIST OF FIGURES

Figure 3.4.1 - High Level System Architecture........13
Figure 3.4.2 - Code used to import and visualize the dataset........13
Figure 3.4.3 - Section of the imported dataset........14
Figure 3.4.4 - Summary of the dataset........14
Figure 3.4.5 - Code to drop the columns from the dataset........15
Figure 3.4.6 - Modified dataset........15
Figure 3.4.7 - Converting data to factors........15
Figure 3.4.8 - Converted data........16
Figure 3.4.9 - Code used to plot the bar graphs........16
Figure 3.4.10 - Plot for Difficulty Walking........17
Figure 3.4.11 - Plot for Physical Health affects the Mental Health........17
Figure 3.4.12 - Plot for Physical Activity affects the Mental Health........18
Figure 3.4.13 - Plot for General Health affects Mental Health........18
Figure 3.4.14 - Plot for Sleep Time affects Mental Health........19
Figure 3.4.15 - Relationship between the Mental Health Risk vs Heart Disease Risk........20
Figure 3.4.16 - Relationship between the Mental Health Risk vs Stroke Risk........21
Figure 3.4.17 - Code used to generate the relationship between the Mental Health Risk vs Heart Disease Risk and Mental Health Risk vs Stroke Risk........21
Figure 3.4.18 - Code for creating new columns for Heart Diseases Risk, Stroke Risk and Mental Health Risk........22
Figure 3.4.19 - Chi-square Test mental diseases risk vs heart disease risk........23
Figure 3.4.20 - Chi-square Test mental diseases risk vs stroke risk........23
Figure 3.4.21 - Code to generate the heatmaps for mental diseases risk vs heart disease risk and stroke risk........23
Figure 3.4.22 - Heatmap of mental diseases risk vs heart disease risk........24
Figure 3.4.23 - Heatmap of mental diseases risk vs stroke........24
Figure 3.4.24 - Code for training mental diseases risk vs heart disease risk data........25
Figure 3.4.25 - Code for training mental diseases risk vs stroke risk data........25
Figure 3.4.26 - Output of the training data of heart disease risk........26
Figure 3.4.27 - Output of the confusion matrix of heart disease risk........26
Figure 3.4.28 - Output of the training data of stroke risk........27
Figure 3.4.29 - Output of the confusion matrix of stroke risk........27
Figure 3.4.30 - Code used to get the estimations of the predictor variables of the heart diseases risk........28
Figure 3.4.31 - Summary of the heart diseases risk model........28
Figure 3.4.32 - Code to calculate the probability of heart diseases risk........29
Figure 3.4.33 - Code used to get the estimations of the predictor variables of the stroke risk.29
Figure 3.4.34 - Summary of the stroke risk model........30
Figure 3.4.35 - Code to calculate the probability of stroke risk........30
Figure 3.4.36 - Code used to get the estimations of the predictor variables of the mental health

risk....31
Figure 3.4.37 - Summary of the mental health risk model.... 31
Figure 3.4.38 - Code to calculate the probability of mental health risk....32
Figure 3.4.39 - Code to calculate the heart diseases and stroke risk based on the mental health risk probability.... 32
Figure 4.1.1 - Output of the training data of heart disease risk using Random Forest Trees....37
Figure 4.1.2 - Output of the confusion matrix of heart disease risk using Random Forest Trees 37
Figure 4.1.3 -Confusion matrix of heart disease risk using Random Forest Trees....37
Figure 4.1.4 - Output of the training data of stroke risk using Random Forest Trees.... 38
Figure 4.1.5 - Output of the confusion matrix of stroke risk using Random Forest Trees.... 39
Figure 4.1.6 - Confusion matrix of stroke risk using Random Forest Trees....39
Figure 4.1.7 - Output of the confusion matrix of heart diseases risk using Decision Trees.... 40
Figure 4.1.8 - Confusion matrix of heart diseases risk using Decision Trees....41
Figure 4.1.9 - Output of the confusion matrix of stroke risk using Decision Trees....42
Figure 4.1.10 - Confusion matrix of stroke risk using Decision Trees....42
Figure 4.1.11 - Output of the confusion matrix of heart diseases risk using KNN....44
Figure 4.1.12 - Confusion matrix of heart diseases risk using KNN....44
Figure 4.1.13 - Output of the confusion matrix of stroke risk using KNN....45
Figure 4.1.14 - Confusion matrix of stroke risk using KNN.... 45
Figure A - 1 - Personal Key Indicators of HeartDiseases dataset.... A - 2
Figure B - 1 - Questionnaire.... B - 1
Figure B - 2 - Questionnaire.... B - 2
Figure B - 3 - Questionnaire.... B - 3
Figure B - 4 - Questionnaire.... B - 4
Figure B - 5 - Responses....B - 4
Figure B - 6 - Responses....B - 5
Figure B - 7 - Responses....B - 5
Figure B - 8 - Responses....B - 6
Figure B - 9 - Responses....B - 7
Figure C - 1 - UI of the Home Page....C - 1
Figure C - 2 - UI of the Factors Page....C - 1
Figure C - 3 - Login Page.... C - 2
Figure C - 4 - Registration Page.... C - 2
Figure C - 5 - Data entry screen to enter the user details.... C - 3
Figure C - 6 - Data entry screen to fill the questionnaire.... C - 3
Figure C - 7 - Risk Analysis screen (Heart Diseases and Stroke Risk Percentages)....C - 3
Figure C - 8 - Risk Analysis screen (Mental Diseases Risk and CVD Risk Percentages)....C - 4
Figure C - 9 - Risk Analysis screen (Association of Each Risk Factor in Heart Diseases and Stroke)....C - 4
Figure C - 10 - Risk Analysis History.... C - 4

LIST OF TABLES

Table 2.1.1 - Comparison between the similar work related to the cardiovascular diseases prediction and proposed project.......... 7

Table 3.3.1 - Fields and details of the BRFSS dataset.......... 10

Table 3.3.2 - Fields and details of the The Personal Key Indicators of Heart Disease dataset. 12

Table 4.1.1 - Confusion matrix for cardiovascular disease risk among people with mental illnesses.......... 31

LIST OF ABBREVIATIONS

BMI - Body Mass Index

CVD - Cardiovascular Diseases

ML - Machine Learning

UI- User Interfaces

CHAPTER 1
INTRODUCTION

1.1 Statement of the problem

Cardiovascular Diseases (CVD) are a group of diseases and disorders in heart and blood vessels and one of the leading causes of deaths worldwide.

There are four main types of CVD:

- Coronary heart disease - occurs when the blood supply of heart muscles is blocked or interrupted by a build-up of fatty substances (atheroma) in the coronary arteries
- Stroke - A stroke is a serious medical condition that occurs when the blood supply to the brain is disturbed
- Peripheral arterial disease - Peripheral arterial disease, happens when there is a blockage in the arteries to your limbs (usually your legs)
- Aortic disease - The aorta is the largest blood vessel in the body. It carries blood from the heart to the body. The most common type of aortic disease is aortic aneurysm, which is where the wall of the aorta becomes weakened and bulges outwards (18 November 2022, Cardiovascular disease)

People who are suffering from mood disorders, depression, anxiety, PTSD (Post-traumatic stress disorder), chronic stress for a long period of time may experience disorders in the heart due to the increase in blood pressure, increased heart rate, and reduced blood flow to the heart. And people with severe mental illness (SMI), including schizophrenia and bipolar disorder are more prone to CVD (Nielsen, R.E., Banner, J. & Jensen, S.E), (Osborn DPJ, Hardoon S, Omar RZ, et a, 2014l)

Mood Disorders: People living with mood disorders, such as major depression or bipolar disorder, find that their mood affects both psychological and mental well-being nearly every day for most of the day

Anxiety Disorders: People respond to certain objects or situations with fear, dread, or terror. Anxiety disorders include generalized anxiety, social anxiety, panic disorders, and phobias

Post-Traumatic Stress Disorder (PTSD): People can experience PTSD after undergoing a traumatic life experience, such as war, natural disaster, or any other serious incident

Chronic Stress: People are in a state of uncomfortable emotional stress—accompanied by predictable biochemical, physiological, and behavioral changes—that is constant and persists over an extended period of time

Evidence shows that mental health disorders such as depression, anxiety, and PTSD can develop after cardiac events, including heart failure, stroke, and heart attack. These disorders can be brought on after an acute heart disease event from factors including pain, fear of death or disability, and financial problems associated with the event (08 September 2022, Heart disease and mental health), (National Center for Chronic Disease Prevention and Health , May 6, 2020), (De Hert M, Detraux J, Vancampfort D)

Mental health disorders such as anxiety and depression may increase the chance of adopting behaviors such as smoking, inactive lifestyle, or failure to take prescribed medications. This is because people experiencing a mental health disorder may have fewer healthy coping strategies for stressful situations, making it difficult for them to make healthy lifestyle choices to reduce their risk for heart disease

The research is to predict the Cardiovascular Diseases Risk among the people who has these types of mental illnesses

1.2 Project Aim and Objectives

The aim and objectives of this study is to examine the relationship between mental health disorders and CVD, including the common risk factors contributing to the development of both conditions and to build a system to predict the risk factor of having Cardiovascular Diseases when having mental illnesses considering a data set and training them using machine learning methods

The specific objectives are as follows:

1. To critically review the literature on the relationship between mental health disorders and CVD
2. To identify the common risk factors contributing to the development of mental health disorders and CVD
3. To investigate the prevalence of CVD risk among individuals with mental health disorders
4. To determine the impact of mental health disorders on the management and prognosis of CVD

5. To develop a decision support web based system to predict the CVD risk factors when having mental illnesses

1.3 Scope

1.3.1 In Scope

- Collect data related to people having several mental illnesses and Cardiovascular Diseases.
- Build a prediction algorithm by training a dataset to get the predictions about Cardiovascular Diseases risk factor.
- Provide a system to enter the details about the patients and their mental health history to get a CVD risk factor as an output.
- Perform exploratory data analysis to identify patterns and correlations between mental health disorders and CVD.
- Ensure data preprocessing, including handling missing values and normalizing data, to prepare for effective machine learning model training.
- Implement various machine learning models and evaluate their performance to select the best predictive model for CVD risk assessment.
- Develop a user-friendly web-based interface for healthcare professionals to input patient data and obtain CVD risk predictions.
- Regularly update the system with new data to improve the accuracy and reliability of the predictive model.

1.3.2 Out Scope

- The system does not provide any medical consultancy services or any other medical or predictive health solutions.
- The system does not offer any diagnostic services or confirm any medical conditions.
- Preventive measures for cardiovascular diseases are not provided by the system.
- The system does not offer personalized treatment plans or recommendations for managing mental health disorders or cardiovascular diseases.
- The system does not provide interpretations or explanations of the predictive model’s results in clinical terms.

CHAPTER 2
LITERATURE REVIEW

2.1 Related Work/ Background Study

The purpose of this research is to find the possibilities of getting Cardiovascular Diseases when people have mental illnesses and how much the risk factors could be. This research can also be used to get predictions about how the mental illnesses lead to CVD and to get treated for them to avoid CVD risk which is one of the leading causes of death in the world now.

The research is going to develop through the literature searches of previous researches, studies, and using a standardized questionnaire. The questionnaire will include questions on demographics, medical history, lifestyle factors, and mental health status.

Following is similar work done in the area of relationship between mental illnesses and cardiovascular diseases.

- Study: "Association Between Mental Illness and Cardiovascular Disease: A Systematic Review" by Osborn DP, et al. (2017) - This study explores the links between mental illnesses and cardiovascular disease, highlighting the increased risk and potential mechanisms underlying this association
- Study: "Mental Illness and Risk of Cardiovascular Disease Mortality" by Druss BG, et al. (2018) - Investigates the impact of mental illness on cardiovascular disease mortality, emphasizing the need for integrated care to address the cardiovascular health of individuals with mental illnesses
- Study: "Psychiatric Disorders and Cardiovascular Disease: A Review of the Evidence" by Lichtman JH, et al. (2019) - Provides a comprehensive overview of the evidence linking psychiatric disorders to cardiovascular disease, emphasizing the importance of collaborative care in managing both conditions simultaneously
- Study: "Cardiovascular Disease and Mental Health: A Review of the Bidirectional Relationship" by Walker ER, et al. (2020) - Explores the bidirectional relationship between cardiovascular disease and mental health, highlighting the potential shared risk factors and pathways that contribute to the increased risk in this population
- Study: "Management of Cardiovascular Risk in Patients with Severe Mental Illnesses" by Carvalho AF, et al. (2021) - Discusses the challenges and strategies for managing

cardiovascular risk in patients with severe mental illnesses, emphasizing the importance of integrated and collaborative care approaches

- Study: “Cardiovascular Risk Prediction Models for People With Severe Mental Illness” by Osborn DPJ, Hardoon S, Omar RZ, et al. (2014) - A prospective study of anonymous data collected between January 1, 1995, and December 31, 2010, Conducted to develop and validate a 10-year risk score for predicting newly recorded cardiovascular events in people with SMI

- Study: “A novel approach for heart disease prediction using strength scores with significant predictors” by Yazdani, A., Varathan, K.D., Chiam, Y.K. et al. (2021) - This study managed to provide a significant contribution in computing the strength scores with significant predictors in heart disease prediction

- Study: “Cardiovascular disease prediction by machine learning incorporation with deep learning” by Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K and Rohini K. (2023) - Finding the patients at high risk for CVD identified by prediction models that use risk stratification

- Study: “Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation” by NCBI. (2020) - Based on a large population of patients with hypertension in Shenzhen, China, the study is aimed at establishing a high-precision coronary heart disease (CHD) prediction model through big data and machine-learning

Comparison between the related research and proposed project: Table 2.1.1 shows the comparison between the similar work related to the cardiovascular diseases prediction and proposed project.

Related Project Name	Related Project Description	Proposed Project
“Association Between Mental Illness and Cardiovascular Disease: A Systematic Review" by Osborn DP, et al. (2017)	This study explores the links between mental illnesses and cardiovascular disease, highlighting the increased risk and potential mechanisms underlying this association	The proposed project predict the mental statuses and their association with the CVD by also considering other factors such as BMI, Alcohol Consumption, Physical

"Mental Illness and Risk of Cardiovascular Disease Mortality" by Druss BG, et al. (2018)	Investigates the impact of mental illness on cardiovascular disease mortality, emphasizing the need for integrated care to address the cardiovascular health of individuals with mental illnesses	Health, Mental health, Difficulty Walking, Physical Activity, General Health, Sleep Time
"Psychiatric Disorders and Cardiovascular Disease: A Review of the Evidence" by Lichtman JH, et al. (2019)	Provides a comprehensive overview of the evidence linking psychiatric disorders to cardiovascular disease, emphasizing the importance of collaborative care in managing both conditions simultaneously	
"Cardiovascular Disease and Mental Health: A Review of the Bidirectional Relationship" by Walker ER, et al. (2020)	Explores the bidirectional relationship between cardiovascular disease and mental health, highlighting the potential shared risk factors and pathways that contribute to the increased risk in this population	
"Management of Cardiovascular Risk in Patients with Severe Mental Illnesses" by Carvalho AF, et al. (2021)	Discusses the challenges and strategies for managing cardiovascular risk in patients with severe mental illnesses, emphasizing the importance of integrated and collaborative care approaches	
"Cardiovascular Risk Prediction Models for People With Severe Mental Illness" by Osborn DPJ, Hardoon S, Omar RZ, et al. (2014)	A prospective study of anonymous data collected between January 1, 1995, and December 31, 2010, Conducted to develop and validate a 10-year risk score for predicting newly recorded cardiovascular events in people with SMI	
"A novel approach for heart disease prediction using strength scores with significant predictors" by Yazdani, A.,	This study managed to provide a significant contribution in computing the strength scores with	

Varathan, K.D., Chiam, Y.K. et al. (2021)	significant predictors in heart disease prediction	
“Cardiovascular diseases prediction by machine learning incorporation with deep learning” by Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K and Rohini K. (2023)	Finding the patients at high risk for CVD identified by prediction models that use risk stratification	
“Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation” by NCBI. (2020)	Based on a large population of patients with hypertension in Shenzhen, China, the study is aimed at establishing a high-precision coronary heart disease (CHD) prediction model through big data and machine-learning	

Table 2.1.1 - Comparison between the similar work related to the cardiovascular diseases prediction and proposed project

2.2 Novelty of the Research and Expected Research Contribution to the Field

There are several other researches related to Cardiovascular Disease risk among the people who have severe mental illnesses (SMI) such as schizophrenia and bipolar disorders. This research is intended to use the previous data sets along with additionally collected data and to consider other mental illnesses such as depression, anxiety, PTSD (Post-traumatic stress disorder) for the prediction of CVD Risk factors. Also predict mental statuses and their association with the CVD by considering other factors such as BMI, Alcohol Consumption, Physical Health, Mental health, Difficulty Walking, Physical Activity, General Health, Sleep Time.

CHAPTER 3
METHODOLOGY

3.1 Research Philosophy

The research philosophy guiding this study is pragmatism, as it allows for a flexible and practical approach to address the complex relationship between mental health disorders and cardiovascular diseases. Pragmatism acknowledges the importance of combining both quantitative and qualitative methods to gain a comprehensive understanding of the research problem. This approach aligns with the study's aim to develop a predictive model (quantitative) while considering the contextual nuances and experiences of individuals (qualitative).

3.2 Research Design

This research adopts a mixed-methods research design. The integration of qualitative and quantitative data will provide a more comprehensive understanding of the relationship between mental health disorders and cardiovascular diseases. The quantitative aspect involves the development of a machine learning model to predict cardiovascular disease risk, while the qualitative aspect encompasses exploring the experiences of individuals through interviews and surveys.

The project is carried out in three main stages: collecting data and creating a database, building the machine learning model and implementing the web based system to use the prediction algorithm.

3.3 Dataset

As an initial task, I analyzed some data sets about the people who have mental illnesses and have a risk of getting CVD. Collected and added more data related to that using a questionnaire. This will include questions on demographics, medical history, lifestyle factors, and mental health status.

- Quantitative Data: A comprehensive dataset will be collected, including demographic details, medical history, mental health diagnosis, lifestyle factors, and cardiovascular outcomes. This data will be obtained from electronic health records, surveys, and relevant databases

- Qualitative Data: Semi-structured interviews and surveys will be conducted to explore the lived experiences of individuals with mental illnesses and cardiovascular diseases. Open-ended questions will be used to capture rich narratives

Mental health issues can be predicted using factors such as

1. Family history and genetics: People with a family history of mental health issues may be at higher risk of developing mental health problems themselves. Genetic testing and analysis can also provide insights into potential risk factors
2. Early life experiences: Traumatic experiences or stress during childhood, such as abuse or neglect, can increase the likelihood of developing mental health issues later in life
3. Behavioral changes: Changes in behavior, such as withdrawal from social activities, increased substance use, or changes in sleep or eating habits, may be early warning signs of mental health issues.
4. Cognitive changes: Changes in thinking patterns, such as persistent negative thoughts or difficulty concentrating, may also be early warning signs of mental health issues
5. Psychological assessments: Standardized assessments such as the Beck Depression Inventory (BDI) or the Generalized Anxiety Disorder (GAD-7) questionnaire can help identify individuals who may be experiencing mental health issues
6. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world (About BRFSS." CDC, 2023).

The below datasets have been used to prepare and develop a prediction algorithm to filter the risk factor of CVD considering several factors using a machine learning approach.

Table 3.3.1 shows the fields and details of the BRFSS in the dataset used to get the data for the system.

Field of data	Data Type	Description
Poor Health	0-30	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
Health Care Access	Yes/No	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
Health Care Access	Yes/No	Do you have one person you think of as your personal doctor or health care provider?
Chronic Health Conditions	Yes/No	About the heart diseases having
Demographics	Yes/No	Are you limited in any way in any activities because of physical, mental, or emotional problems?
Demographics	Yes/No	Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone?
Demographics	Yes/No	Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?
Demographics	Yes/No	Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor's office or shopping?
Tobacco Use	Yes/No	Have you smoked at least 100 cigarettes in your entire life?

Table 3.3.1 - Fields and details of the BRFSS dataset

(National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health, May 16, 2014)

The Personal Key Indicators of Heart Disease dataset can be used in the research to consider factors associated with both mental health and CVD. The dataset is a part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

Table 3.3.2 shows the fields and details of The Personal Key Indicators of Heart Disease in the dataset. Appendix A contains a snapshot of the data obtained from the Personal Key Indicators of Heart Disease dataset.

Appendix B shows the questionnaire used to collect the data from the people and the responses to the questionnaire.

Field of data	Data Type	Description
Chronic Health Conditions	Yes/No	About the heart diseases having
BMI	Number	Body Mass Index
Smoking	Yes/No	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
Alcohol Consumption	Yes/No	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
Stroke	Yes/No	About the heart diseases having
Physical Health	0-30	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
Mental Health	0-30	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
Difficulty Walking	Yes/No	Do you have serious difficulty walking or climbing stairs?
Sex	Male/Female/Other	Male, Female other

Age Category	Number	Fourteen-level age category
Race	White/Black	White or Black
Diabetic	Yes/No	Yes or No
Physical Activity	0-30	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
General Health	0-30	Would you say that in general your health is?
Sleep Time	Number	Hours per day
Asthma	Yes/No	Yes or No
Kidney Diseases	Yes/No	Yes or No
Skin Cancer	Yes/No	Yes or No

Table 3.3.2 - Fields and details of the The Personal Key Indicators of Heart Disease dataset

(Kamil Pytlak, Personal Key Indicators of Heart Disease)

3.4 Machine Learning Model

Machine learning technologies can analyze large amounts of data, to identify patterns. By giving the factors associated with the mental health conditions, an ML model can be created to predict the risk factors of the CVD risk

For this, data is selected from the dataset and preprocessed by removing and changing the data types. After that, machine learning is used to analyze the data to get the expected output.

The Figure 3.4.1 shows the High Level System Architecture to be designed to use the prediction model to predict CVD risk when given the data about mental health

The system uses a database to store the dataset data and use them to process the data given to the system by users.

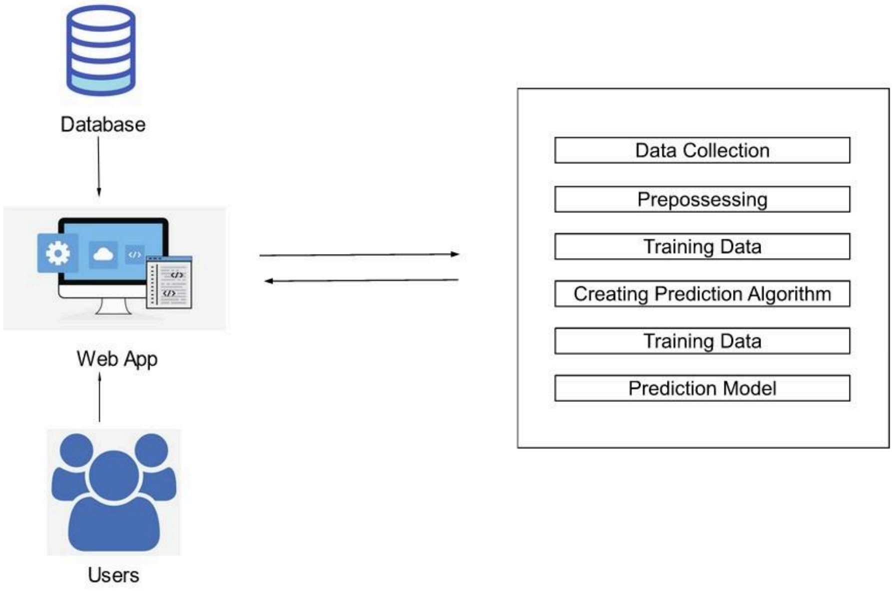


Figure 3.4.1 - High Level System Architecture

The project has used supervised machine learning, it is a ML technique where models are trained on labeled data. The models find the mapping function to map input variables with the output variable or the labels.

In this project, R is used to build the machine learning model and the predictions. R is one of the major languages for data science. It provides excellent visualization features, which are essential to explore the data before submitting it to any automated learning, as well as assessing the results of the learning algorithm. Many R packages for machine learning are available off the shelf and many modern methods in statistical learning are implemented in R as part of their development.

Below are the steps of creating the machine learning model using R,

Import the dataset and view a summary to get an idea about the statistical distribution of the data. Figure 3.4.2 shows the code used to import and visualize the dataset.

```
#importing libraries
library(tidyverse)
library(caret)
library(randomForest)

#Read Data
dataset<-read.csv("D:\\Nidu\\project\\dataset.csv")
#Describing Data
glimpse(dataset)

summary(dataset)
```

Figure 3.4.2 - Code used to import and visualize the dataset

Figure 3.4.3 shows a section of the imported dataset

```
Rows: 319,795
Columns: 18
$ HeartDisease     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No"…
$ BMI              <dbl> 16.60, 20.34, 26.58, 24.21, 23.71, 28.87, 21.63, 31.6…
$ Smoking          <chr> "Yes", "No", "Yes", "No", "No", "Yes", "No", "Yes", "…
$ AlcoholDrinking  <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No",…
$ Stroke           <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "No"…
$ PhysicalHealth   <int> 3, 0, 20, 0, 28, 6, 15, 5, 0, 0, 30, 0, 0, 7, 0, 1, 5…
$ MentalHealth     <int> 30, 0, 30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 30, 0, 2,…
$ DiffWalking      <chr> "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "N…
$ Sex              <chr> "Female", "Female", "Male", "Female", "Female", "Fema…
$ AgeCategory      <chr> "55-59", "80 or older", "65-69", "75-79", "40-44", "7…
$ Race             <chr> "White", "White", "White", "White", "White", "Black",…
$ Diabetic         <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "Yes", "N…
$ PhysicalActivity <chr> "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", …
$ GenHealth        <chr> "Very good", "Very good", "Fair", "Good", "Very good"…
$ SleepTime        <int> 5, 7, 8, 6, 8, 12, 4, 9, 5, 10, 15, 5, 8, 7, 5, 6, 10…
$ Asthma           <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "Yes", "…
$ KidneyDisease    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "Yes"…
$ SkinCancer       <chr> "Yes", "No", "No", "Yes", "No", "No", "Yes", "No", "N…
```

Figure 3.4.3 - Section of the imported dataset

Figure 3.4.4 shows the summary of the dataset

```
 HeartDisease           BMI          Smoking          AlcoholDrinking   
 Length:319795      Min.   :12.02   Length:319795      Length:319795     
 Class :character   1st Qu.:24.03   Class :character   Class :character  
 Mode  :character   Median :27.34   Mode  :character   Mode  :character  
                    Mean   :28.33                                        
                    3rd Qu.:31.42                                        
                    Max.   :94.85                                        
    Stroke          PhysicalHealth    MentalHealth    DiffWalking       
 Length:319795      Min.   : 0.000   Min.   : 0.000   Length:319795     
 Class :character   1st Qu.: 0.000   1st Qu.: 0.000   Class :character  
 Mode  :character   Median : 0.000   Median : 0.000   Mode  :character  
                    Mean   : 3.372   Mean   : 3.898                     
                    3rd Qu.: 2.000   3rd Qu.: 3.000                     
                    Max.   :30.000   Max.   :30.000                     
     Sex            AgeCategory            Race             Diabetic        
 Length:319795      Length:319795      Length:319795      Length:319795     
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  


 PhysicalActivity    GenHealth           SleepTime         Asthma         
 Length:319795      Length:319795      Min.   : 1.000   Length:319795     
 Class :character   Class :character   1st Qu.: 6.000   Class :character  
 Mode  :character   Mode  :character   Median : 7.000   Mode  :character  
                                       Mean   : 7.097                     
                                       3rd Qu.: 8.000                     
                                       Max.   :24.000                     
 KidneyDisease       SkinCancer       
 Length:319795      Length:319795     
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Figure 3.4.4 - Summary of the dataset

Here Age Category, Sex, Race, Asthma, Kidney Disease, and Skin Cancer fields will be dropped from the dataset as they are not used for the prediction model.

Figure 3.4.5 shows the code to drop the above columns from the dataset and Figure 3.4.6 shows the modified dataset.

```
# Remove specified columns
dataset <- dataset %>%
  select(-Sex, -AgeCategory, -Diabetic, -Asthma, -KidneyDisease, -SkinCancer)
```

Figure 3.4.5 - Code to drop the columns from the dataset

```
Rows: 319,795
Columns: 12
$ HeartDisease     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No…
$ BMI              <dbl> 16.60, 20.34, 26.58, 24.21, 23.71, 28.87, 21.63, 31.64, 26…
$ Smoking          <chr> "Yes", "No", "Yes", "No", "No", "Yes", "No", "Yes", "No", …
$ AlcoholDrinking  <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…
$ Stroke           <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "No", "No…
$ PhysicalHealth   <int> 3, 0, 20, 0, 28, 6, 15, 5, 0, 0, 30, 0, 0, 7, 0, 1, 5, 0, …
$ MentalHealth     <int> 30, 0, 30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 30, 0, 2, 30, …
$ DiffWalking      <chr> "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "No", "…
$ Race             <chr> "White", "White", "White", "White", "White", "Black", "Whi…
$ PhysicalActivity <chr> "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", "No",…
$ GenHealth        <chr> "Very good", "Very good", "Fair", "Good", "Very good", "Fa…
$ SleepTime        <int> 5, 7, 8, 6, 8, 12, 4, 9, 5, 10, 15, 5, 8, 7, 5, 6, 10, 8, …
```

Figure 3.4.6 - Modified dataset

The next step is to convert all the character columns to factors. Figure 3.4.7 shows the code that has converted the data and Figure 3.4.8 shows the converted dataset with the columns Heart Diseases, Stroke, Smoking, Alcohol Drinking, Difficulty Walking, Mental Health, Physical Activity, General Health, Heart Disease Risk, Stroke Risk, Mental Diseases Risk

```
# Convert categorical variables to factors
dataset$HeartDisease <- as.factor(dataset$HeartDisease)
dataset$Stroke <- as.factor(dataset$Stroke)
dataset$Smoking <- as.factor(dataset$Smoking)
dataset$AlcoholDrinking <- as.factor(dataset$AlcoholDrinking)
dataset$DiffWalking <- as.factor(dataset$DiffWalking)
dataset$PhysicalActivity <- as.factor(dataset$PhysicalActivity)
dataset$GenHealth <- as.factor(dataset$GenHealth)
```

Figure 3.4.7 - Converting data to factors

```
Rows: 319,795
Columns: 12
$ HeartDisease      <fct> No, No, No, No, No, Yes, No, No, No, No, Yes, No, No, No, …
$ BMI               <dbl> 16.60, 20.34, 26.58, 24.21, 23.71, 28.87, 21.63, 31.64, 26…
$ Smoking           <fct> Yes, No, Yes, No, No, Yes, No, Yes, No, No, Yes, Yes, Yes,…
$ AlcoholDrinking   <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, No…
$ Stroke            <fct> No, Yes, No, No, No, No, No, No, No, No, No, No, No, No, N…
$ PhysicalHealth    <int> 3, 0, 20, 0, 28, 6, 15, 5, 0, 0, 30, 0, 0, 7, 0, 1, 5, 0, …
$ MentalHealth      <int> 30, 0, 30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 30, 0, 2, 30, …
$ DiffWalking       <fct> No, No, No, No, Yes, Yes, No, Yes, No, Yes, Yes, No, Yes, …
$ Race              <chr> "White", "White", "White", "White", "White", "Black", "Whi…
$ PhysicalActivity  <fct> Yes, Yes, Yes, No, Yes, No, Yes, No, No, Yes, No, Yes, Yes…
$ GenHealth         <fct> Very good, Very good, Fair, Good, Very good, Fair, Fair, G…
$ SleepTime         <int> 5, 7, 8, 6, 8, 12, 4, 9, 5, 10, 15, 5, 8, 7, 5, 6, 10, 8, …
```

Figure 3.4.8 - Converted data

The following charts shows how the

- Difficulty Walking affects Mental Health
- Physical Health affects the Mental Health
- Physical Activity affects the Mental Health
- General Health affects Mental Health
- Sleep Time affects Mental Health

Figure 3.4.9 shows the code used to plot the bar graphs to show the above relationships among the data fields and Figure 3.4.10 - Figure 3.4.14 shows the plots generated for each of the above relationships.

```
# Visualizations of Difficulty Walking, Physical Health, Physical Activity, General Health, Sleep Time vs Mental_Diseases_Risk
# Bar plot for Difficulty Walking affects  Mental Health
ggplot(dataset, aes(x = DiffWalking, fill = Mental_Diseases_Risk)) +
  geom_bar(position = "fill") +
  labs(title = "Mental Health affects Difficulty Walking", x = "Difficulty Walking", y = "Mental Diseases Risk Proportion")

# Bar plot for Physical Health affects the Mental Health
ggplot(dataset, aes(x = Mental_Diseases_Risk, fill = PhysicalHealth)) +
  geom_bar(position = "fill") +
  labs(title = "Physical Health affects the Mental Health", x = "Mental Diseases Risk", y = "Physical Health Proportion")

# Bar plot for Physical Activity affects the Mental Health
ggplot(dataset, aes(x = Mental_Diseases_Risk, fill = PhysicalActivity)) +
  geom_bar(position = "fill") +
  labs(title = "Physical Activity affects the Mental Health", x = "Mental Diseases Risk", y = "Physical Activity Proportion")

# Bar plot for General Health affects Mental Health
ggplot(dataset, aes(x = GenHealth, fill = Mental_Diseases_Risk)) +
  geom_bar(position = "fill") +
  labs(title = "General Health affects Mental Health", x = "General Health", y = "Mental Diseases Risk Proportion")

# Bar plot for Sleep Time affects Mental Health
ggplot(dataset, aes(x = SleepTime, fill = Mental_Diseases_Risk)) +
  geom_bar(position = "fill") +
  labs(title = "Sleep Time affects Mental Health", x = "Sleep Time", y = "Mental Diseases Risk Proportion")
```

Figure 3.4.9 - Code used to plot the bar graphs

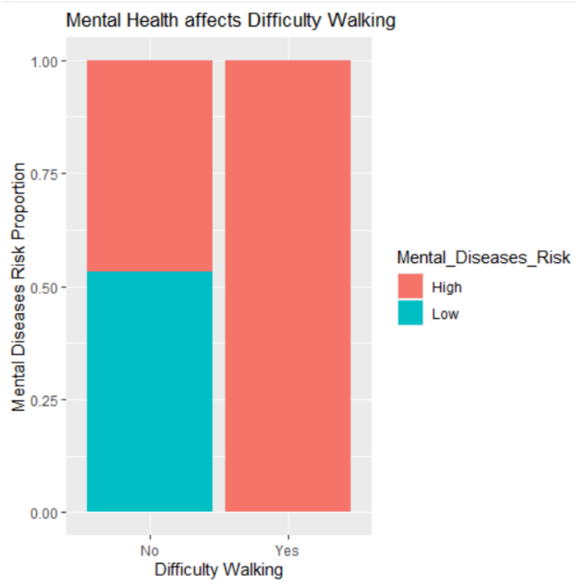


Figure 3.4.10 - Plot for Difficulty Walking

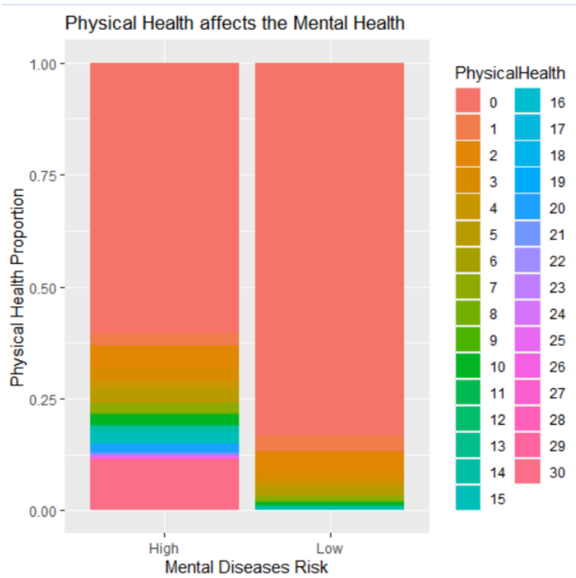


Figure 3.4.11 - Plot for Physical Health affects the Mental Health

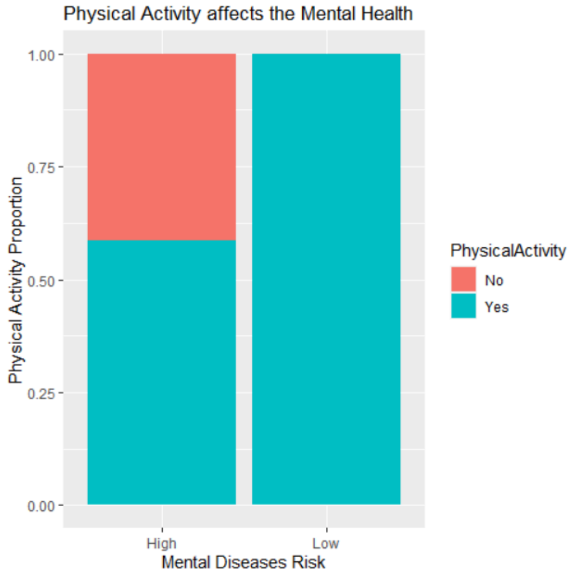


Figure 3.4.12 - Plot for Physical Activity affects the Mental Health

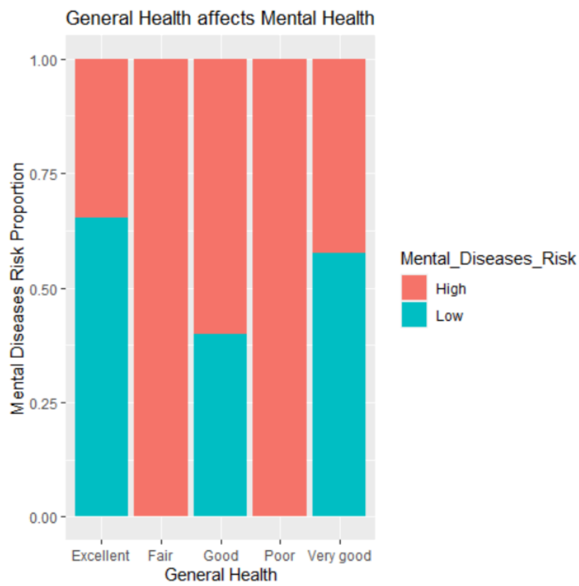


Figure 3.4.13 - Plot for General Health affects Mental Health

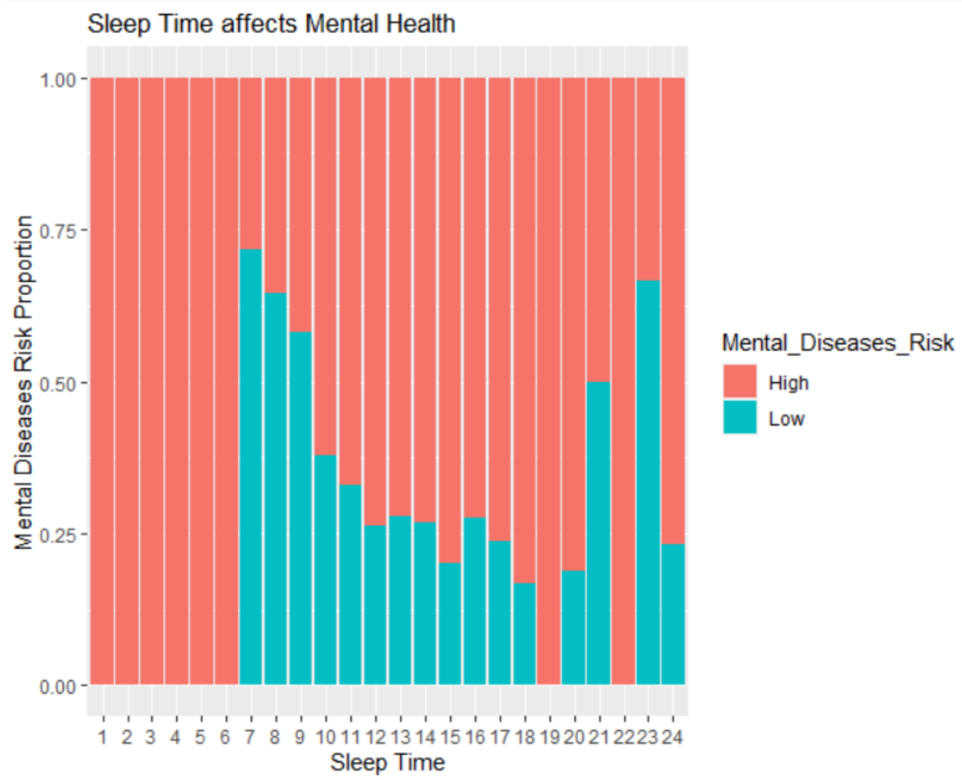


Figure 3.4.14 - Plot for Sleep Time affects Mental Health

- The bar plot in Figure 3.4.10 indicates that there's an association between mental disease risk levels and the proportion of individuals experiencing difficulty walking. While the specifics vary based on the risk level, the overall trend is that as mental disease risk increases, the proportion of individuals experiencing difficulty walking tends to increase as well. However, even within the high-risk group, there's a portion of individuals who do not experience difficulty walking

- The bar plot in Figure 3.4.11 indicates that when there is an increased number of days of physical health not good, there is a high risk of mental health diseases

- The bar plot in Figure 3.4.12 shows that a proportion of not doing any physical activity affects the high mental health risk and when engaged in physical activities, the risk of having mental health risk is low

- The plot in Figure 3.4.13 shows that the proportions of poor and fair general health conditions contributes to a higher risk of mental health diseases and the contribution of excellent, very good and good general health levels for a higher mental health disease risk is low

- The plot in Figure 3.4.14 indicates for each sleep time how the risk of mental health disease increases. According to the bar plots, there is a high risk of mental health diseases when the sleep time is low

The data fields, BMI, Alcohol Consumption, Physical Health, Mental health, Difficulty Walking, Physical Activity, General Health, Sleep Time, Heart Disease are used to predict the Heart Diseases risk

The data fields, BMI, Alcohol Consumption, Physical Health, Mental health, Difficulty Walking, Physical Activity, General Health, Sleep Time, Stroke are used to predict the Stroke risk

The data fields, Physical Health, Mental health, Difficulty Walking, Physical Activity, General Health, Sleep Time are used to predict the Mental Health risk

Figure 3.4.15 and Figure 3.4.16 show the 2 graphs for the relationship between the Mental Health Risk vs Heart Disease Risk and Mental Health vs Stroke and Figure 3.4.17 shows the code used to generate the graphs.

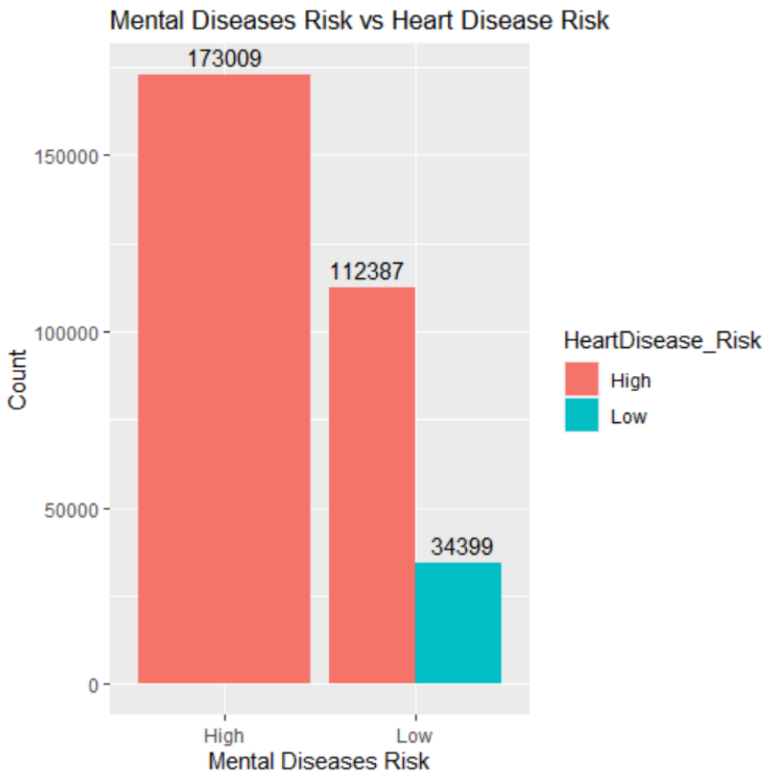


Figure 3.4.15 - Relationship between the Mental Health Risk vs Heart Disease Risk

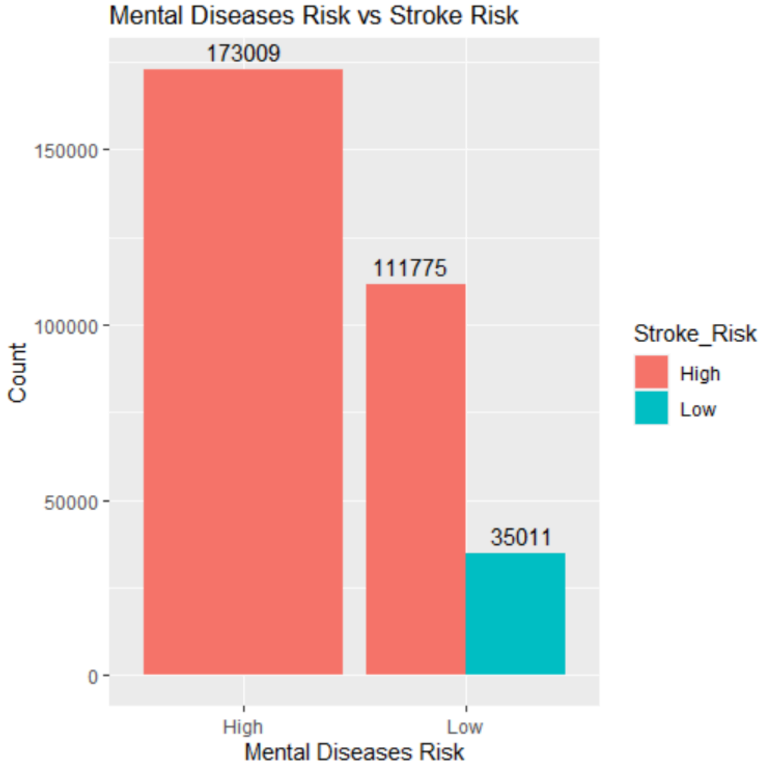


Figure 3.4.16 - Relationship between the Mental Health Risk vs Stroke Risk

```
# Visualizations of the relationship between the Mental Health Risk vs Heart Disease Risk and Stroke Risk
# Bar plot for Mental Diseases Risk vs Heart Disease_Risk
ggplot(dataset, aes(x = Mental_Diseases_Risk, fill = HeartDisease_Risk)) +
  geom_bar(position = "fill") +
  labs(title = "Mental Diseases Risk vs Heart Disease Risk",
      x = "Mental Diseases Risk", y = "Heart Disease Risk Proportion")

# Bar plot for Mental Diseases Risk vs Stroke_Risk
ggplot(dataset, aes(x = Mental_Diseases_Risk, fill = Stroke_Risk)) +
  geom_bar(position = "fill") +
  labs(title = "Mental Diseases Risk vs Stroke Risk",
      x = "Mental Diseases Risk", y = "Stroke Risk Proportion")
```

Figure 3.4.17 - Code used to generate the relationship between the Mental Health Risk vs Heart Disease Risk and Mental Health Risk vs Stroke Risk

Plotting 2 graphs of the relationship between Mental Health Risk vs Heart Disease Risk and Mental Health Risk vs Stroke Risk shows that higher proportion of mental health risk contributes for a high risk of heart disease or stroke.

- High Mental Health Risk: For individuals with a high risk of mental health issues, the entire range of the heart attack risk and stroke proportion (from 0.00 to 1.00) is associated with a high risk of heart attacks and Strokes. This indicates that within the group classified as having a high risk of mental health issues, all proportions of individuals face a high risk of heart attacks or Stroke.

- Low Mental Health Risk: Conversely, for individuals with a low risk of mental health issues, the heart attack or stroke risk proportion is split into two ranges:
    - From 0.00 to 0.25: Within this range, individuals have a low risk of heart attacks or stroke.
    - From 0.25 to 1.00: Within this range, individuals have a high risk of heart attacks or stroke.

Individuals with a high mental health risk are uniformly associated with a high risk of heart attacks or stroke, while individuals with a low mental health risk have a transition from low to high heart attack risk or stroke risk as the proportion shifts from 0.00 to 0.25 to 0.25 to 1.00.

To compare the observed results with the expected results, a chi-square test has been used. The Chi-Square test can be used for categorical data to check if the variables have an association. Figure 3.4.18 shows the code used to create new columns for Heart Diseases Risk, Stroke Risk and Mental Health Risk considering the factors BMI, Smoking, Alcohol Drinking, Physical Health, Difficulty Walking, Physical Activity, General Health, Sleep Time. Creating a contingency table that counts the frequency of occurrences for each combination of levels between the "Mental Diseases Risk" vs “Heart Diseases Risk” and "Mental Diseases Risk" vs "Stroke Risk" variables. And finally perform the Chi-square Test.

```r
# Create a new column for heart disease risk
dataset$HeartDisease_Risk <- ifelse(dataset$BMI >= 25 | dataset$Smoking == "Yes" | dataset$AlcoholDrinking == "Yes" |
                                    dataset$PhysicalHealth > 15 | dataset$MentalHealth > 15 | dataset$DiffWalking == "Yes" |
                                    dataset$PhysicalActivity == "No" | dataset$GenHealth %in% c("Fair", "Poor") |
                                    dataset$SleepTime < 7 | dataset$HeartDisease == "Yes", "High", "Low")

# Create a new column for stroke risk
dataset$Stroke_Risk <- ifelse(dataset$BMI >= 25 | dataset$Smoking == "Yes" | dataset$AlcoholDrinking == "Yes" |
                              dataset$PhysicalHealth > 15 | dataset$MentalHealth > 15 | dataset$DiffWalking == "Yes" |
                              dataset$PhysicalActivity == "No" | dataset$GenHealth %in% c("Fair", "Poor") |
                              dataset$SleepTime < 7 | dataset$Stroke == "Yes", "High", "Low")

# Create a new column for mental diseases risk
dataset$Mental_Diseases_Risk <- ifelse(dataset$MentalHealth > 15 | dataset$DiffWalking == "Yes" |
                                       dataset$PhysicalHealth > 15 | dataset$PhysicalActivity == "No" |
                                       dataset$GenHealth %in% c("Fair", "Poor") | dataset$SleepTime < 7, "High", "Low")

# Create a contingency tables
contingency_table_for_heart_diseases <- table(dataset$Mental_Diseases_Risk, dataset$HeartDisease_Risk)
contingency_table_for_stroke <- table(dataset$Mental_Diseases_Risk, dataset$Stroke_Risk)

# Perform Chi-square test
chi_square_test_for_heart_diseases <- chisq.test(contingency_table_for_heart_diseases)
chi_square_test_for_stroke <- chisq.test(contingency_table_for_stroke)

# Print the results
print(chi_square_test_for_heart_diseases)
print(chi_square_test_for_stroke)

# Convert the contingency table to a data frame for ggplot
contingency_df_for_heart_diseases <- as.data.frame(contingency_table_for_heart_diseases)
contingency_df_for_stroke <- as.data.frame(contingency_table_for_stroke)
names(contingency_df_for_heart_diseases) <- c("Mental_Diseases_Risk", "HeartDisease_Risk", "Count")
names(contingency_df_for_stroke) <- c("Mental_Diseases_Risk", "Stroke_Risk", "Count")
```

Figure 3.4.18 - Code for creating new columns for Heart Diseases Risk, Stroke Risk and Mental Health Risk

Figure 3.4.19 gives the Chi-square Test result that contains a test statistic (X-squared) of 45429, with 1 degree of freedom (df), and a p-value less than 2.2e-16 for Mental Diseases Risk vs Heart Diseases Risk.

P-value shows that there is a statistically significant association between the variables represented in the contingency table ,as the p - value is smaller than 0.05.

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table_for_heart_diseases
X-squared = 45429, df = 1, p-value < 2.2e-16
```

Figure 3.4.19 - Chi-square Test mental diseases risk vs heart disease risk

Figure 3.4.20 gives the Chi-square Test result that contains a test statistic (X-squared) of 46336, with 1 degree of freedom (df), and a p-value less than 2.2e-16 for Mental Diseases Risk vs Heart Diseases Risk.

P-value shows that there is a statistically significant association between the variables represented in the contingency table ,as the p - value is smaller than 0.05.

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table_for_stroke
X-squared = 46336, df = 1, p-value < 2.2e-16
```

Figure 3.4.20 - Chi-square Test mental diseases risk vs stroke risk

Figure 3.4.21 shows the code used to generate the heatmap by creating a data frame using the contingency table. Figure 3.4.22 shows the heatmap that shows the association of Mental Diseases Risk vs Heart Diseases Risk and Figure 3.4.23 represents the association of Mental Diseases Risk vs Stroke Risk

```
# Create the heatmaps
ggplot(contingency_df, aes(x = Mental_Diseases_Risk, y = HeartDisease_Risk, fill = Count)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(title = "Heatmap of Mental Diseases Risk vs Heart Diseases Risk",
       x = "Mental Diseases Risk",
       y = "Heart Diseases Risk",
       fill = "Count") +
  theme_minimal()

ggplot(contingency_df, aes(x = Mental_Diseases_Risk, y = Stroke_Risk, fill = Count)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(title = "Heatmap of Mental Diseases Risk vs Stroke Risk",
       x = "Mental Diseases Risk",
       y = "Stroke Risk",
       fill = "Count") +
  theme_minimal()
```

Figure 3.4.21 - Code to generate the heatmaps for mental diseases risk vs heart disease risk and stroke risk

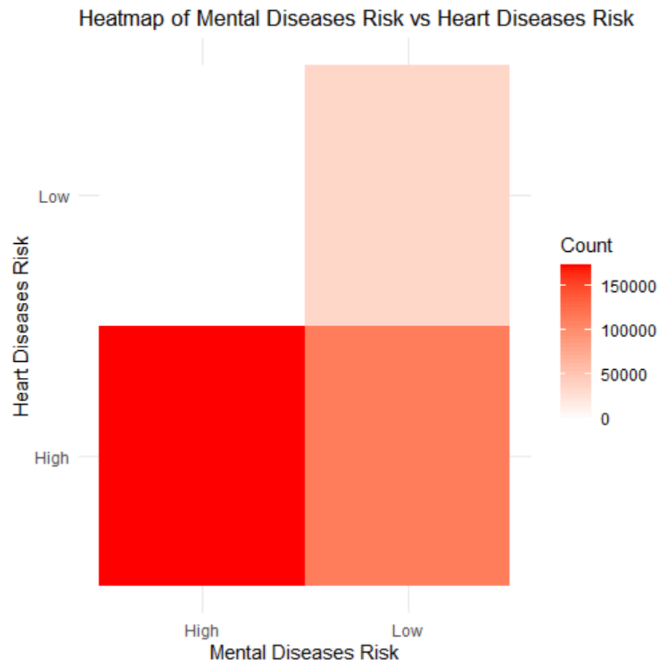


Figure 3.4.22 - Heatmap of mental diseases risk vs heart disease risk

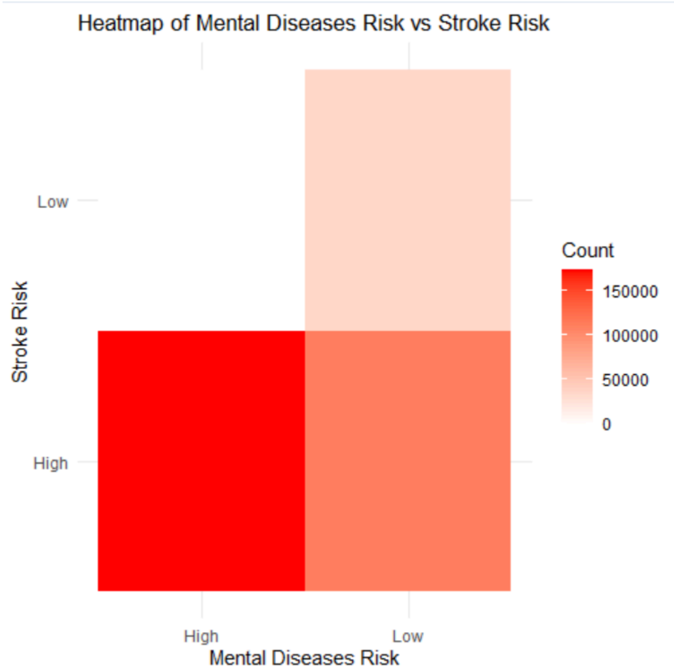


Figure 3.4.23 - Heatmap of mental diseases risk vs stroke

In summary, the visualizations suggest that there's a relationship between mental health risk levels and the risk of heart attacks or stroke.

After the above exploratory data analysis, the data is split into training and test datasets. Here, the Random Forest algorithm is used to train the data. Setting up the model is done using ‘set.seed’ to select a random seed and make the model reproducible. Next, call the randomForest classifier and point it to the ‘HeartDisease_Risk’ and ‘Stroke_Risk’ column for the outcome and provide the ‘train’ set as input. Figure 3.4.18 and Figure 3.4.19 shows the code and Figure 3.4.24 to Figure 3.4.29 shows the output.

```
# Set seed for reproducibility
set.seed(123)

# Split the dataset into training and test sets (e.g., 70% training, 30% test)
train_index <- createDataPartition(dataset$HeartDisease_Risk, p = 0.7, list = FALSE)
train_data <- dataset[train_index, ]
test_data <- dataset[-train_index, ]


# Define predictors (features) and outcome variable
predictors <- c("Mental_Diseases_Risk","HeartDisease_Risk")

# Train the Random Forest model
model <- randomForest(HeartDisease_Risk ~ ., data = train_data[predictors], ntree = 500)

# Make predictions on the test set
predictions <- predict(model, newdata = test_data[predictors])

# Evaluate the accuracy of the model
accuracy <- table(predictions, test_data$HeartDisease_Risk)
print(accuracy)
model

# Create the confusion matrix
confusion_matrix <- confusionMatrix(data = as.factor(predictions), reference = as.factor(test_data$HeartDisease_Risk))
print(confusion_matrix)
```

Figure 3.4.24 - Code for training mental diseases risk vs heart disease risk data

```
# Set seed for reproducibility
set.seed(123)

# Split the dataset into training and test sets (e.g., 70% training, 30% test)
train_index <- createDataPartition(dataset$Stroke_Risk, p = 0.7, list = FALSE)
train_data <- dataset[train_index, ]
test_data <- dataset[-train_index, ]


# Define predictors (features) and outcome variable
predictors <- c("Mental_Diseases_Risk","Stroke_Risk")

# Train the Random Forest model
model <- randomForest(Stroke_Risk ~ ., data = train_data[predictors], ntree = 500)

# Make predictions on the test set
predictions <- predict(model, newdata = test_data[predictors])

# Evaluate the accuracy of the model
accuracy <- table(predictions, test_data$Stroke_Risk)
print(accuracy)
model

# Create the confusion matrix
confusion_matrix <- confusionMatrix(data = as.factor(predictions), reference = as.factor(test_data$Stroke_Risk))
print(confusion_matrix)
```

Figure 3.4.25 - Code for training mental diseases risk vs stroke risk data

```
predictions  High   Low
       High 85618 10319
       Low      0     0
> model

Call:
 randomForest(formula = HeartDisease_Risk ~ ., data = train_data[predictors],      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 10.76%
Confusion matrix:
       High Low class.error
High 199778   0           0
Low   24080   0           1
```

Figure 3.4.26 - Output of the training data of heart disease risk

Estimate of error rate (10.76%), the number of trees (500), the variables at each split (1), and the function used to build the classifier (randomForest).

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85618 10319
      Low      0     0

               Accuracy : 0.8924
                 95% CI : (0.8905, 0.8944)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8924
         Neg Pred Value :    NaN
             Prevalence : 0.8924
         Detection Rate : 0.8924
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High
```

Figure 3.4.27 - Output of the confusion matrix of heart disease risk

```
predictions   High    Low
       High 85435 10503
       Low       0      0
> model

Call:
 randomForest(formula = Stroke_Risk ~ ., data = train_data[predictors],      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 10.95%
Confusion matrix:
       High Low class.error
High 199349   0           0
Low   24508   0           1
```

Figure 3.4.28 - Output of the training data of stroke risk

Estimate of error rate (10.95%), the number of trees (500), the variables at each split (1), and the function used to build the classifier (randomForest).

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85435 10503
      Low      0     0

               Accuracy : 0.8905
                 95% CI : (0.8885, 0.8925)
    No Information Rate : 0.8905
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8905
         Neg Pred Value :    NaN
             Prevalence : 0.8905
         Detection Rate : 0.8905
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High
```

Figure 3.4.29 - Output of the confusion matrix of stroke risk

Random Forest Tree algorithm is used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

As the final step of building the machine learning model, four functions have been created to get probabilities of Heart Diseases Risk, Stroke Risk, Mental Diseases Risk and Heart Diseases/Stroke Risk when having a Mental Diseases Risk as percentages.

Figure 3.4.30 shows the code used to get the estimations of the predictor variables of the heart diseases risk probability as a percentage based on BMI, Smoking, Alcohol Drinking, Physical Health, Diff Walking, Physical Activity, General Health, SleepTime fields taken as the independent variables and Heart Diseases field as the dependant variable. This prediction has been done as a logistic regression. Figure 3.4.31 shows the summary of the model that includes estimated coefficients and their statistical significance of the given fields. The significance of each predictor variable can help in understanding their relative importance in predicting the outcome variable. The coefficient estimates show the effect of each predictor variable on the log-odds of the dependent variable. The z-value and p-value provide information about the statistical significance of each coefficient estimate.

Figure 3.4.32 shows the code for the function that used the above generated estimations that calculate the probability of Heart Diseases Risk.

```
# Fit logistic regression model
model <- glm(HeartDisease ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime,
             data = dataset,
             family = binomial)

# Display summary of the model
summary(model)

# Create the formula
formula <- formula(HeartDisease ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime)

# Fit logistic regression model
model <- glm(formula, data = dataset, family = binomial)

# Display summary of the model
summary(model)
```

Figure 3.4.30 - Code used to get the estimations of the predictor variables of the heart diseases risk

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -4.1624732  0.0499377 -83.353  < 2e-16 ***
BMI                  -0.0035544  0.0009696  -3.666 0.000246 ***
SmokingYes            0.5021451  0.0135581  37.036  < 2e-16 ***
AlcoholDrinkingYes   -0.5490177  0.0321442 -17.080  < 2e-16 ***
PhysicalHealth        0.0015145  0.0008072   1.876 0.060619 .  
DiffWalkingYes        0.6540042  0.0170215  38.422  < 2e-16 ***
PhysicalActivityYes -0.0547622  0.0151523  -3.614 0.000301 ***
GenHealthFair         1.9985704  0.0314914  63.464  < 2e-16 ***
GenHealthGood         1.4209660  0.0287775  49.378  < 2e-16 ***
GenHealthPoor         2.4721658  0.0385363  64.152  < 2e-16 ***
GenHealthVery good    0.7084722  0.0297315  23.829  < 2e-16 ***
SleepTime             0.0505009  0.0039082  12.922  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 186906  on 319794  degrees of freedom
Residual deviance: 165263  on 319783  degrees of freedom
AIC: 165287

Number of Fisher Scoring iterations: 6
```

Figure 3.4.31 - Summary of the heart diseases risk model

```
# Define a function to predict the likelihood of heart disease
predict_heart_disease <- function(BMI, SmokingYes, AlcoholDrinkingYes, PhysicalHealth, DiffWalkingYes, PhysicalActivityYes,
GenHealthFair, GenHealthGood, GenHealthPoor, GenHealthVeryGood, SleepTime) {

  # Coefficients from the logistic regression model
  intercept <- -4.1259703
  coef_BMI <- -0.0035544
  coef_SmokingYes <- 0.5021451
  coef_AlcoholDrinkingYes <- -0.5490177
  coef_PhysicalHealth <- 0.0015145
  coef_DiffWalkingYes <- 0.6540042
  coef_PhysicalActivityYes <- -0.0547622
  coef_GenHealthFair <- 1.9985704
  coef_GenHealthGood <- 1.4209660
  coef_GenHealthPoor <- 2.4721658
  coef_GenHealthVeryGood <- 0.7084722
  coef_SleepTime <- 0.0505009

  # Calculate the log odds based on the coefficients and predictor values
  log_odds <- intercept +
              BMI * coef_BMI +
              SmokingYes * coef_SmokingYes +
              AlcoholDrinkingYes * coef_AlcoholDrinkingYes +
              PhysicalHealth * coef_PhysicalHealth +
              DiffWalkingYes * coef_DiffWalkingYes +
              PhysicalActivityYes * coef_PhysicalActivityYes +
              GenHealthFair * coef_GenHealthFair +
              GenHealthGood * coef_GenHealthGood +
              GenHealthPoor * coef_GenHealthPoor +
              GenHealthVeryGood * coef_GenHealthVeryGood +
              SleepTime * coef_SleepTime

  # Calculate the probability of heart disease using the logistic function
  probability <- 1 / (1 + exp(-log_odds))
  return(probability)
}
```

Figure 3.4.32 - Code to calculate the probability of heart diseases risk

Similarly, Figure 3.4.33 shows the code used to get the estimations of the predictor variables of the stroke risk. Figure 3.4.34 shows the summary of the stroke risk model and Figure 3.4.35 is for the code to calculate the probability of stroke risk.

```
# Fit logistic regression model
model <- glm(Stroke ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime,
              data = dataset,
              family = binomial)

# Display summary of the model
summary(model)

# Create the formula
formula <- formula(Stroke ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime)

# Fit logistic regression model
model <- glm(formula, data = stroke_yes_data, family = binomial)

# Display summary of the model
summary(model)
```

Figure 3.4.33 - Code used to get the estimations of the predictor variables of the stroke risk

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -4.540990   0.070472 -64.437  < 2e-16 ***
BMI                -0.017658   0.001417 -12.465  < 2e-16 ***
SmokingYes          0.307094   0.019643  15.634  < 2e-16 ***
AlcoholDrinkingYes -0.426032   0.046539  -9.154  < 2e-16 ***
PhysicalHealth      0.004774   0.001094   4.362 1.29e-05 ***
DiffWalkingYes      0.970274   0.023748  40.857  < 2e-16 ***
PhysicalActivityYes -0.107860  0.021461  -5.026 5.01e-07 ***
GenHealthFair       1.760029   0.045897  38.347  < 2e-16 ***
GenHealthGood       1.218661   0.042462  28.700  < 2e-16 ***
GenHealthPoor       2.093943   0.053943  38.817  < 2e-16 ***
GenHealthVery good  0.618886   0.043869  14.108  < 2e-16 ***
SleepTime           0.057761   0.005256  10.989  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102778  on 319794  degrees of freedom
Residual deviance:  91733  on 319783  degrees of freedom
AIC: 91757

Number of Fisher Scoring iterations: 7
```

Figure 3.4.34 - Summary of the stroke risk model

```
# Define a function to predict the likelihood of heart disease
predict_stroke <- function(BMI, SmokingYes, AlcoholDrinkingYes, PhysicalHealth, DiffWalkingYes, PhysicalActivityYes,
 GenHealthFair, GenHealthGood, GenHealthPoor, GenHealthVeryGood, SleepTime) {

  # Coefficients from the logistic regression model
  intercept <- -4.540990
  coef_BMI <- -0.017658
  coef_SmokingYes <- 0.307094
  coef_AlcoholDrinkingYes <- -0.426032
  coef_PhysicalHealth <- 0.004774
  coef_DiffWalkingYes <- 0.970274
  coef_PhysicalActivityYes <- -0.107860
  coef_GenHealthFair <- 1.760029
  coef_GenHealthGood <- 1.218661
  coef_GenHealthPoor <- 2.093943
  coef_GenHealthVeryGood <- 0.618886
  coef_SleepTime <- 0.057761

  # Calculate the log odds based on the coefficients and predictor values
  log_odds <- intercept +
              BMI * coef_BMI +
              SmokingYes * coef_SmokingYes +
              AlcoholDrinkingYes * coef_AlcoholDrinkingYes +
              PhysicalHealth * coef_PhysicalHealth +
              DiffWalkingYes * coef_DiffWalkingYes +
              PhysicalActivityYes * coef_PhysicalActivityYes +
              GenHealthFair * coef_GenHealthFair +
              GenHealthGood * coef_GenHealthGood +
              GenHealthPoor * coef_GenHealthPoor +
              GenHealthVeryGood * coef_GenHealthVeryGood +
              SleepTime * coef_SleepTime

  # Calculate the probability of stroke using the logistic function
  probability <- 1 / (1 + exp(-log_odds))
  return(probability)
}
```

Figure 3.4.35 - Code to calculate the probability of stroke risk

Figure 3.4.36 shows the code used to get the estimations of the predictor variables of the mental health risk. Figure 3.4.37 shows the summary of the mental health risk model and Figure 3.4.38 is for the code to calculate the probability of mental health risk. Figure 3.4.39 shows the code used to calculate the probability of heart diseases and stroke risk based on the mental diseases risk probability.

```
# Fit logistic regression model
model <- glm(MentalHealth ~ PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime,
              data = dataset,
              family = binomial)

# Display summary of the model
summary(model)

# Create the formula
formula <- formula(MentalHealth ~ PhysicalHealth + DiffWalking + PhysicalActivity + GenHealth + SleepTime)

# Fit logistic regression model
model <- glm(formula, data = dataset, family = binomial)

# Display summary of the model
summary(model)
```

Figure 3.4.36 - Code used to get the estimations of the predictor variables of the mental health risk

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.6059489  0.0340256 -47.198  < 2e-16 ***
PhysicalHealth       0.0445812  0.0006862  64.964  < 2e-16 ***
DiffWalkingYes      -0.0247972  0.0167970  -1.476     0.14    
PhysicalActivityYes -0.1011481  0.0136970  -7.385 1.53e-13 ***
GenHealthFair        1.1601883  0.0235283  49.310  < 2e-16 ***
GenHealthGood        0.6975336  0.0201643  34.593  < 2e-16 ***
GenHealthPoor        1.3900905  0.0313437  44.350  < 2e-16 ***
GenHealthVery good   0.3203412  0.0204049  15.699  < 2e-16 ***
SleepTime           -0.1689265  0.0038363 -44.033  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228199  on 319794  degrees of freedom
Residual deviance: 205942  on 319786  degrees of freedom
AIC: 205960
```

Figure 3.4.37 - Summary of the mental health risk model

```
# Define a function to predict the likelihood of Mental Health disease
predict_mental_diseases <- function(PhysicalHealth, DiffWalkingYes, PhysicalActivityYes, GenHealthFair,
GenHealthGood, GenHealthPoor, GenHealthVeryGood, SleepTime) {

  # Coefficients from the logistic regression model
  intercept <- -1.6059489
  coef_PhysicalHealth <- 0.0445812
  coef_DiffWalkingYes <- -0.0247972
  coef_PhysicalActivityYes <- -0.1011481
  coef_GenHealthFair <- 1.1601883
  coef_GenHealthGood <- 0.6975336
  coef_GenHealthPoor <- 1.3900905
  coef_GenHealthVeryGood <- 0.3203412
  coef_SleepTime <- -0.1689265

  # Calculate the log odds based on the coefficients and predictor values
  log_odds <- intercept +
              PhysicalHealth * coef_PhysicalHealth +
              DiffWalkingYes * coef_DiffWalkingYes +
              PhysicalActivityYes * coef_PhysicalActivityYes +
              GenHealthFair * coef_GenHealthFair +
              GenHealthGood * coef_GenHealthGood +
              GenHealthPoor * coef_GenHealthPoor +
              GenHealthVeryGood * coef_GenHealthVeryGood +
              SleepTime * coef_SleepTime

  # Calculate the probability of heart disease using the logistic function
  probability <- 1 / (1 + exp(-log_odds))
  return(probability)
}
```

Figure 3.4.38 - Code to calculate the probability of mental health risk

```
# Calculate combined probability of heart disease or stroke given mental health
combined_probability <- heart_disease_probability * stroke_probability * mental_health_probability
```

Figure 3.4.39 - Code to calculate the heart diseases and stroke risk based on the mental health risk probability

3.5 Web Based System

As the final task, building the web based system is developed using the prediction algorithm developed. There is a home page, factors page that shows some factors that considers heart diseases and mental diseases risks and an About page.

The system has the following functionalities,

1. Registration and Login to the system
2. Enter user details
3. Fill the questionnaire
4. Analyze data and receive the risk factor of CVD - The four functions built in the machine learning model will be used to get the probabilities of Heart Diseases Risk, Stroke Risk, Mental Diseases Risk and Heart Diseases/Stroke Risk when having a Mental Diseases Risk as percentages

Appendix C contains the Images of the web based system

3.6 Ethical Considerations

The system ensures transparency and user privacy. When submitting data, users will receive a confirmation prompt to verify their consent. Access to view data is restricted to authenticated users, safeguarding sensitive information from unauthorized access. Additionally, a disclaimer informs users that the analysis relies on predictive models, acknowledging the variability in such assessments.

3.7 Limitations

Limitations may include potential biases in the collected data, external factors influencing mental and cardiovascular health, and the generalizability of findings to diverse populations.

CHAPTER 4
EVALUATION AND RESULTS

4.1 Evaluation Approach of Machine Learning Model

The project has used supervised machine learning, it is a ML technique where models are trained on labeled data. The models find the mapping function to map input variables with the output variable or the labels.

Predicting cardiovascular disease risk in people with mental illnesses based on the provided data involves assigning individuals to different categories (e.g., at risk or not at risk). Therefore, it is a classification problem.

In a classification model, the goal is to categorize input data into discrete classes or labels. In this case, the classes are "The people having mental illnesses are at risk for cardiovascular disease" and "The people having mental illnesses are not at risk for cardiovascular disease." Classification models are suitable when the target variable is categorical, as opposed to continuous.

Following are the steps used in evaluating cardiovascular disease risk in people with mental illnesses

4.1.1 Data Preparation

In the data set, Age Category, Sex, Race, Asthma, Kidney Disease, and Skin Cancer fields are dropped as they are not used for the prediction model. Heart Diseases, BMI, Stroke, Heart Diseases, Smoking, Physical Health, Mental Health, General Health, Sleep Time, Difficulty Walking, Physical Activity fields are used for the prediction model. As the next step, all the character columns are converted to factors.

4.1.2 Data Splitting

The data is split into training and test datasets

4.1.3 Model Selection

Choosing a classification algorithm suitable for the problem. The following classification models are considered

KNN: The k Nearest Neighbors technique involves grouping the closest objects in a dataset and finding the most frequent or average characteristics among the objects.

Decision Trees: Decision trees are classifiers that are used to determine what category an input falls into by traversing the leaf's and nodes of a tree

Random Forest: Random forest is a collection of many decision trees from random subsets of the data, resulting in a combination of trees that may be more accurate in prediction than a single decision tree

4.1.4 Confusion Matrix

Confusion Matrix is used to analyze and understand true positives, true negatives, false positives, and false negatives. Table 4.1.1 shows the confusion matrix for cardiovascular disease risk among people with mental illnesses.

	Predicted Class		
		Has Cardiovascular Disease Risk	**No Cardiovascular Disease Risk**
Actual Class	**Has Cardiovascular Disease Risk**	TP	FN
	No Cardiovascular Disease Risk	FP	TN

Table 4.1.1 - Confusion matrix for cardiovascular disease risk among people with mental illnesses

- True Positives (TP): The instances that are correctly predicted as positive (actual positive, predicted positive)
- True Negatives (TN): The instances that are correctly predicted as negative (actual negative, predicted negative)
- False Positives (FP): The instances that are incorrectly predicted as positive (actual negative, predicted positive)
- False Negatives (FN): The instances that are incorrectly predicted as negative (actual positive, predicted negative)

4.1.5 Evaluation Metrics

- Accuracy: Overall correctness of the model

Accuracy = (TP + FP) / (P + N)

- Error Rate: Percentage of our prediction are wrong =

Error Rate = (1 − Accuracy)

- Precision: Proportion of true positives among predicted positives

Precision = TP / (TP + FP)

- Recall (Sensitivity): Proportion of true positives among actual positives

Recall = TP / (TP + FN)

- F1-score: Harmonic mean of precision and recall, balancing the two

F1 Score = (2 x Precision x Recall) / (Precision + Recall)

- Specificity: Also known as the true negative rate

Specificity = TN / (TN + FP)

- Area Under the Receiver Operating Characteristic (ROC-AUC): Measures the model's ability to discriminate between positive and negative instances

4.1.6 Training the Model

Train the model using the training set

- **Random Forest Tree Model**

Figure 4.1.1, Figure 4.1.2 and Figure 4.1.3 shows the confusion matrix and metrics for the training data of heart disease risk prediction when having a mental diseases risk using Random Forest tree model

```
predictions  High   Low
       High 85618 10319
       Low      0     0
> model

Call:
 randomForest(formula = HeartDisease_Risk ~ ., data = train_data[predictors],      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 10.76%
```

Figure 4.1.1 - Output of the training data of heart disease risk using Random Forest Trees

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85618 10319
      Low      0     0

               Accuracy : 0.8924
                 95% CI : (0.8905, 0.8944)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8924
         Neg Pred Value :    NaN
             Prevalence : 0.8924
         Detection Rate : 0.8924
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High
```

Figure 4.1.2 - Output of the confusion matrix of heart disease risk using Random Forest Trees

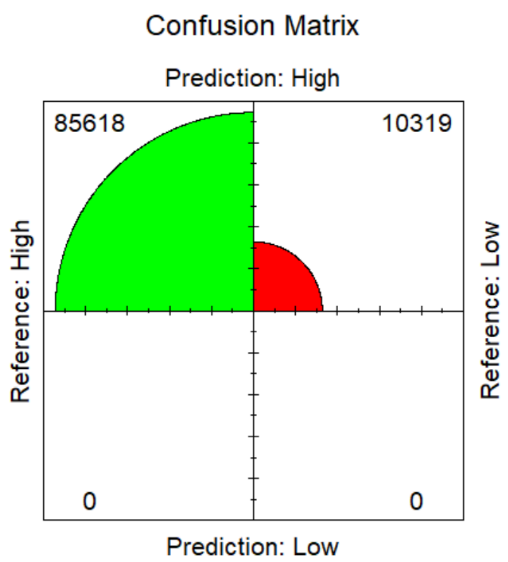


Figure 4.1.3 -Confusion matrix of heart disease risk using Random Forest Trees

- Accuracy: 0.8924
- Sensitivity: 1.0000
- Specificity: 0.0000
- Error Rate: 10.76%

85618 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."
10319 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."
0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.
0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

Figure 4.1.4, Figure 4.1.5 and Figure 4.1.6 shows the confusion matrix and metrics for the training data of stroke risk prediction when having a mental diseases risk using Random Forest tree model

```
predictions  High   Low
       High 85435 10503
       Low      0     0
> model

Call:
 randomForest(formula = Stroke_Risk ~ ., data = train_data[predictors],      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 10.95%
```

Figure 4.1.4 - Output of the training data of stroke risk using Random Forest Trees

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85435 10503
      Low      0     0

               Accuracy : 0.8905
                 95% CI : (0.8885, 0.8925)
    No Information Rate : 0.8905
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8905
         Neg Pred Value :    NaN
             Prevalence : 0.8905
         Detection Rate : 0.8905
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High
```

Figure 4.1.5 - Output of the confusion matrix of stroke risk using Random Forest Trees

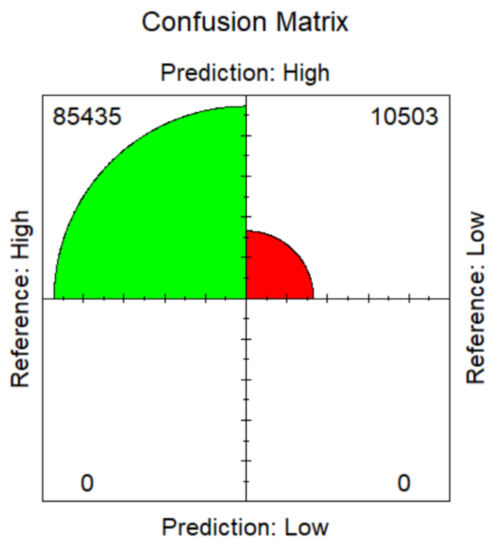


Figure 4.1.6 - Confusion matrix of stroke risk using Random Forest Trees

- Accuracy: 0.8905
- Sensitivity: 1.0000
- Specificity: 0.0000
- Error Rate: 10.95%

85435 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."

10503 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."

0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.

0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

- **Decision Trees Model**

Figure 4.1.7 and Figure 4.1.8 shows the confusion matrix and metrics for the training data of heart diseases risk prediction when having a mental diseases risk using Decision Trees

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85618 10319
      Low      0     0

               Accuracy : 0.8924
                 95% CI : (0.8905, 0.8944)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8924
         Neg Pred Value :    NaN
             Prevalence : 0.8924
         Detection Rate : 0.8924
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High

> print(paste("Error Rate:", round(error_rate_percentage, 2), "%"))
[1] "Error Rate: 10.76 %"
```

Figure 4.1.7 - Output of the confusion matrix of heart diseases risk using Decision Trees

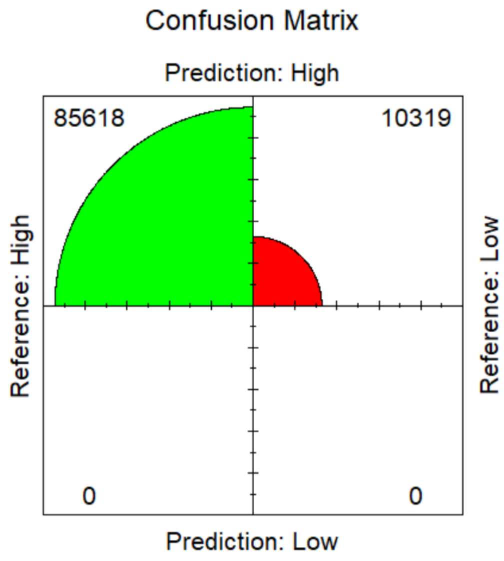


Figure 4.1.8 - Confusion matrix of heart diseases risk using Decision Trees

- Accuracy: 0.8924
- Sensitivity: 1.0000
- Specificity: 0.000
- Error Rate - 10.76%

85618 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."
10319 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."
0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.
0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

Figure 4.1.9 and Figure 4.1.10 shows the confusion matrix and metrics for the training data of stroke risk prediction when having a mental diseases risk using Decision Trees

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85435 10503
      Low      0     0

               Accuracy : 0.8905
                 95% CI : (0.8885, 0.8925)
    No Information Rate : 0.8905
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8905
         Neg Pred Value :    NaN
             Prevalence : 0.8905
         Detection Rate : 0.8905
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High

> print(paste("Error Rate:", round(error_rate_percentage, 2), "%"))
[1] "Error Rate: 10.95 %"
```

Figure 4.1.9 - Output of the confusion matrix of stroke risk using Decision Trees

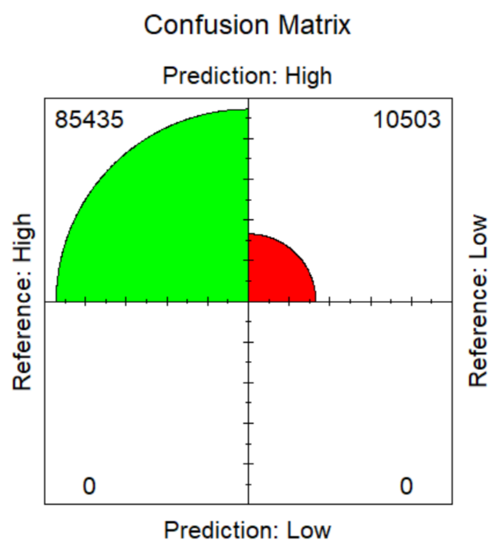


Figure 4.1.10 - Confusion matrix of stroke risk using Decision Trees

- Accuracy: 0.8905
- Sensitivity: 1.0000
- Specificity: 0.000
- Error Rate - 10.95%

85435 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."

10503 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."

0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.

0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

- **K Nearest Neighbors Technique (KNN)**

Figure 4.1.11 and Figure 4.1.12 shows the confusion matrix and metrics for the training data of heart diseases risk prediction when having a mental diseases risk using KNN

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85618 10319
      Low      0     0

               Accuracy : 0.8924
                 95% CI : (0.8905, 0.8944)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8924
         Neg Pred Value :    NaN
             Prevalence : 0.8924
         Detection Rate : 0.8924
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High

> print(paste("Error Rate:", round(error_rate_percentage, 2), "%"))
[1] "Error Rate: 10.76 %"
```

Figure 4.1.11 - Output of the confusion matrix of heart diseases risk using KNN

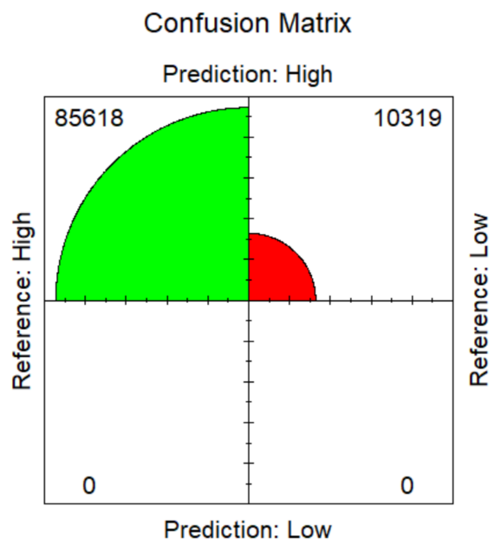


Figure 4.1.12 - Confusion matrix of heart diseases risk using KNN

- Accuracy: 0.8924
- Sensitivity: 1.0000
- Specificity: 0.000
- Error Rate - 10.76%

85618 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."
10319 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."
0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.
0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

Figure 4.1.13 and Figure 4.1.14 shows the confusion matrix and metrics for the training data of stroke risk prediction when having a mental diseases risk using KNN

```
Confusion Matrix and Statistics

          Reference
Prediction  High   Low
      High 85435 10503
      Low      0     0

               Accuracy : 0.8905
                 95% CI : (0.8885, 0.8925)
    No Information Rate : 0.8905
    P-Value [Acc > NIR] : 0.5026

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8905
         Neg Pred Value :    NaN
             Prevalence : 0.8905
         Detection Rate : 0.8905
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : High

> print(paste("Error Rate:", round(error_rate_percentage, 2), "%"))
[1] "Error Rate: 10.95 %"
```

Figure 4.1.13 - Output of the confusion matrix of stroke risk using KNN

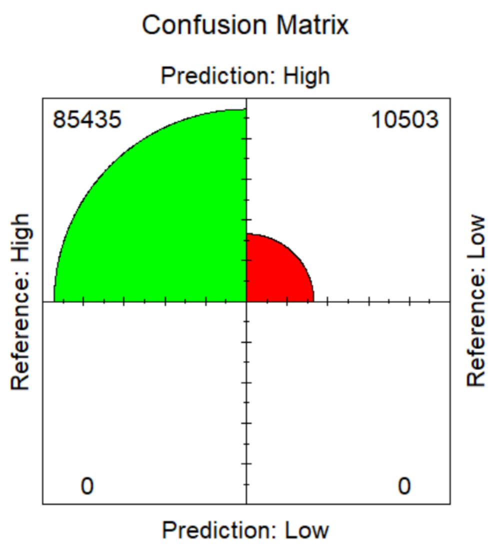


Figure 4.1.14 - Confusion matrix of stroke risk using KNN

- Accuracy: 0.8905
- Sensitivity: 1.0000
- Specificity: 0.000
- Error Rate - 10.95%

85435 (True Positives): The number of instances where the model correctly predicted the class as "High Risk of CVD" and the actual class was also "High Risk of CVD."

10503 (False Positives): The number of instances where the model predicted the class as "High Risk of CVD" but the actual class was "Low Risk of CVD." This indicates misclassification where the model falsely labeled "Low Risk of CVD" instances as "High Risk of CVD."

0 (True Negatives): The number of instances where the model correctly predicted the class as "Low Risk of CVD" and the actual class was also "Low Risk of CVD." In this case, there are no true negatives.

0 (False Negatives): The number of instances where the model predicted the class as "Low Risk of CVD" but the actual class was "High Risk of CVD." In this case, there are no false negatives.

In this project, the Random Forest Tree algorithm is used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

4.2 Web Based System Validation and Testing

Assess the development process and validation techniques used for the decision support web-based system to predict CVD risk factors in individuals with mental illnesses. Evaluate the accuracy, reliability, and usability of the system through validation studies and user feedback

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

In summary, this report suggests that there is a connection between mental health issues and heart problems. People with conditions like depression or anxiety tend to have higher risks of developing heart diseases like heart attacks or strokes. Our initial analysis using R showed that these individuals often struggle with maintaining a healthy weight, have poorer physical health, and experience disturbed sleep patterns. This, in turn, seems to increase their chances of developing heart issues.

Our early machine learning models seem promising in predicting the likelihood of heart problems in people with mental health conditions. However, we still need to fine-tune these models to ensure their accuracy. Our future plans include delving deeper into the specific factors linking mental health and heart diseases, using more comprehensive data and advanced techniques. We are also working on a user-friendly online tool that can help individuals assess their risk of heart issues based on their mental health status.

Overall, this study aims to contribute to better preventive strategies and tailored interventions for people dealing with both mental health and heart issues. We are working on refining our models further and conducting more thorough tests to ensure the reliability of our proposed system.

5.2 Future Work

Future work includes identifying which risk factors impact the most for the developing cardiovascular diseases in the context of mental illness and giving suggestions to mitigate these risk factors by lifestyle modifications, stress management techniques, targeted medication regimes, and early detection and treatment of mental health issues

5.3 Deliverables

Develop a prediction algorithm to predict the risk factor of CVD for people having mental illnesses

A web based system to enter the details about the health conditions and other related data about the people and predict the risk of getting CVD considering the mental illnesses if exists in a person

REFERENCES

"Cardiovascular diseases (CVDs)." World Health Organization (WHO), 11 June 2021, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 30 October 2023.

Nielsen, R.E., Banner, J. & Jensen, S.E. Cardiovascular disease in patients with severe mental illness. Nat Rev Cardiol 18, 136–145 (2021). [Online] URL: https://www.nature.com/articles/s41569-020-00463-7

Osborn DPJ, Hardoon S, Omar RZ, et al. Cardiovascular Risk Prediction Models for People With Severe Mental Illness: Results From the Prediction and Management of Cardiovascular Risk in People With Severe Mental Illnesses (PRIMROSE) Research Program. JAMA Psychiatry. 2015;72(2):143–151. doi:10.1001/jamapsychiatry.2014.2133, [Online] URL: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2043173

"Cardiovascular disease - Illnesses & conditions." NHS inform, 18 November 2022, https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/cardiovascular-disease. Accessed 30 October 2023.

"Heart disease and mental health." Better Health Channel, https://www.betterhealth.vic.gov.au/health/healthyliving/heart-disease-and-mental-health. Accessed 30 October 2023.

"Cardiovascular Disease dataset." Kaggle, https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. Accessed 30 October 2023.

"Heart Disease and Mental Health Disorders | cdc.gov." CDC, https://www.cdc.gov/heartdisease/mentalhealth.htm. Accessed 30 October 2023.

"The intriguing relationship between coronary heart disease and mental disorders." NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6016051/. Accessed 30 October 2023.

"About BRFSS." CDC, https://www.cdc.gov/brfss/about/index.htm. Accessed 30 October 2023.

"Indicators of Heart Disease (2022 UPDATE)." Kaggle, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. Accessed 30 October 2023.

Madhugiri, Devashree. "Build a Step-by-step Machine Learning Model Using R." Analytics Vidhya, 26 July 2022, https://www.analyticsvidhya.com/blog/2022/06/build-a-step-by-step-machine-learning-model-using-r/. Accessed 30 October 2023.

“A novel approach for heart disease prediction using strength scores with significant predictors - BMC Medical Informatics and Decision Making.” BMC Medical Informatics and Decision Making, 21 June 2021, https://doi.org/10.1186/s12911-021-01527-5. Accessed 30 October 2023.

“Cardiovascular diseases prediction by machine learning incorporation with deep learning.” Frontiers, 17 April 2023, https://www.frontiersin.org/articles/10.3389/fmed.2023.1150933/full. Accessed 30 October 2023.

Ding, Eric, and Sai Veeranki. “Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation.” NCBI, 6 July 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7381262/. Accessed 30 October 2023.

APPENDICES

APPENDIX A: Data obtained from the Personal Key Indicators of HeartDiseases dataset

This dataset comprises essential indicators related to heart disease risk factors. The following fields and details are included:

- BMI (Body Mass Index):
  - Type: Number
  - Definition: Body Mass Index, a measure calculated from a person's weight and height, used to assess body fatness and potential health risks associated with weight.
- Smoking:
  - Type: Yes/No
  - Definition: Indicates whether the individual has smoked at least 100 cigarettes in their lifetime. Note: 5 packs of cigarettes equate to 100 cigarettes.
- Alcohol Consumption:
  - Type: Yes/No
  - Definition: Indicates whether the individual is a heavy drinker, with adult men consuming more than 14 drinks per week and adult women consuming more than 7 drinks per week.
- Physical Health:
  - Type: Number (0-30)
  - Definition: The number of days in the past 30 days during which the individual's physical health was not good.
- Mental Health:
  - Type: Number (0-30)
  - Definition: The number of days in the past 30 days during which the individual's mental health was not good.
- Difficulty Walking:
  - Type: Yes/No
  - Definition: Indicates whether the individual experiences serious difficulty walking or climbing stairs.
- Sex:
  - Type: Male/Female/Other

- Definition: Gender identification of the individual.

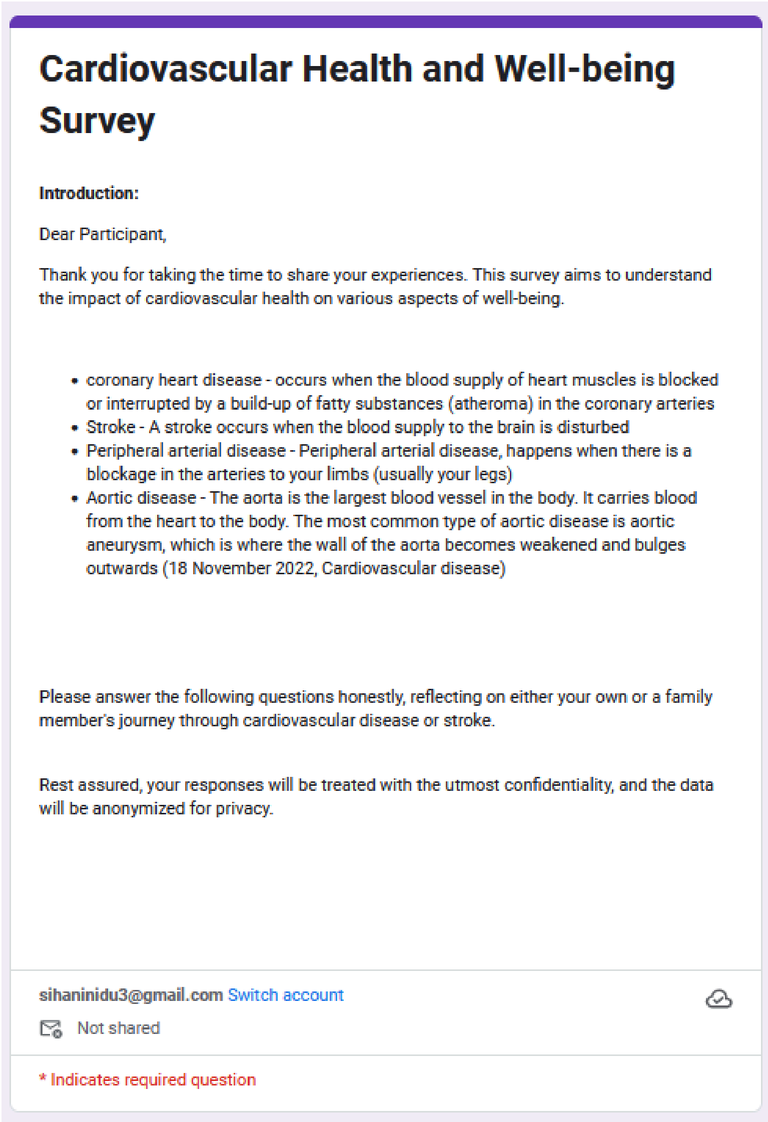
- Age Category:
    - Type: Number
    - Definition: Fourteen-level categorization of age.
- Physical Activity:
    - Type: Number (0-30)
    - Definition: Indicates the number of days during the past 30 days in which the individual engaged in physical activity or exercise other than their regular job.
- General Health:
    - Type: Poor, Fair, Good, Very Good, Excellent
    - Definition: Subjective assessment of the individual's general health status.
- Sleep Time:
    - Type: Number
    - Definition: Hours per day of sleep.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
2	No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
3	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
4	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
5	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
6	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
7	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
8	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
9	No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No	Good	9	Yes	No	No
10	No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, borderline	No	Fair	5	No	Yes	No
11	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No
12	Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No
13	No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes	Very good	5	No	No	No
14	No	28.37	Yes	No	No	0	0	Yes	Male	75-79	White	Yes	Yes	Very good	8	No	No	No
15	No	28.15	No	No	No	7	0	Yes	Female	80 or older	White	No	No	Good	7	No	No	No
16	No	29.29	Yes	No	No	0	30	Yes	Female	60-64	White	No	No	Good	5	No	No	No
17	No	29.18	No	No	No	1	0	No	Female	50-54	White	No	Yes	Very good	6	No	No	No
18	No	26.26	No	No	No	5	2	No	Female	70-74	White	No	No	Very good	10	No	No	No
19	No	22.59	Yes	No	No	0	30	Yes	Male	70-74	White	No, borderline	Yes	Good	8	No	No	No
20	No	29.86	Yes	No	No	0	0	Yes	Female	75-79	Black	Yes	No	Fair	5	No	Yes	No
21	No	18.13	No	No	No	0	0	No	Male	80 or older	White	No	Yes	Excellent	8	No	No	Yes
22	No	21.16	No	No	No	0	0	No	Female	80 or older	Black	No, borderline	No	Good	8	No	No	No
23	No	28.9	No	No	No	2	5	No	Female	70-74	White	Yes	No	Very good	7	No	No	No

Figure A - 1 - Personal Key Indicators of HeartDiseases dataset

APPENDIX B: Data obtained from the Questionnaire

The questionnaire utilized to collect data from individuals included the following inquiries and corresponding responses. These questions aim to gather information related to key indicators of heart disease risk factors, lifestyle choices, and general health status.



Cardiovascular Health and Well-being Survey

Introduction:

Dear Participant,

Thank you for taking the time to share your experiences. This survey aims to understand the impact of cardiovascular health on various aspects of well-being.

- coronary heart disease - occurs when the blood supply of heart muscles is blocked or interrupted by a build-up of fatty substances (atheroma) in the coronary arteries
- Stroke - A stroke occurs when the blood supply to the brain is disturbed
- Peripheral arterial disease - Peripheral arterial disease, happens when there is a blockage in the arteries to your limbs (usually your legs)
- Aortic disease - The aorta is the largest blood vessel in the body. It carries blood from the heart to the body. The most common type of aortic disease is aortic aneurysm, which is where the wall of the aorta becomes weakened and bulges outwards (18 November 2022, Cardiovascular disease)

Please answer the following questions honestly, reflecting on either your own or a family member's journey through cardiovascular disease or stroke.

Rest assured, your responses will be treated with the utmost confidentiality, and the data will be anonymized for privacy.

sihaninidu3@gmail.com Switch account

Not shared

* Indicates required question

Figure B - 1 - Questionnaire

Section 1: Participant Details *

1.1. Are you responding based on your own experience or that of a family member?

- ◯ My own experience
- ◯ Family member's experience

1.2. If responding for a family member, please specify the relationship (e.g., parent, sibling, etc.):

Your answer

Section 2: BMI (Body Mass Index) *

2.1. What is/was the BMI (Body Mass Index) of the individual?

(BMI = kg/m2 where kg is a person's weight in kilograms and m2 is their height in metres squared)

Your answer

Section 3: Mental Health in the Past 30 Days *

3.1. Have there been any noticeable changes in mental well-being during this period?

- ◯ Yes
- ◯ No
- ◯ Not Sure

Figure B - 2 - Questionnaire

3.2 For how many days during the past 30 days the mental health of the individual was not good? *

Your answer

Section 4: Physical Health in the Past 30 Days *

4.1. Have there been any specific challenges or improvements in physical health during this period?

- ◯ Challenges
- ◯ Improvements
- ◯ No significant changes

4.2 For how many days during the past 30 days the physical health of the individual was not good? *

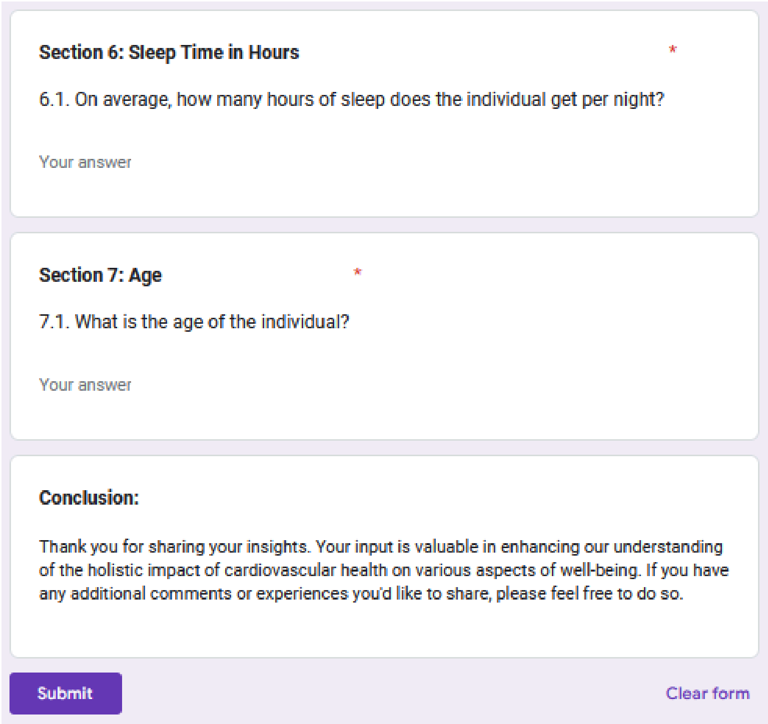
Your answer

Section 5: General Health *

5.1. How would you describe the overall general health of the individual?

- ◯ Poor
- ◯ Fair
- ◯ Good
- ◯ Very Good
- ◯ Excellent

Figure B - 3 - Questionnaire



Section 6: Sleep Time in Hours *

6.1. On average, how many hours of sleep does the individual get per night?

Your answer

Section 7: Age *

7.1. What is the age of the individual?

Your answer

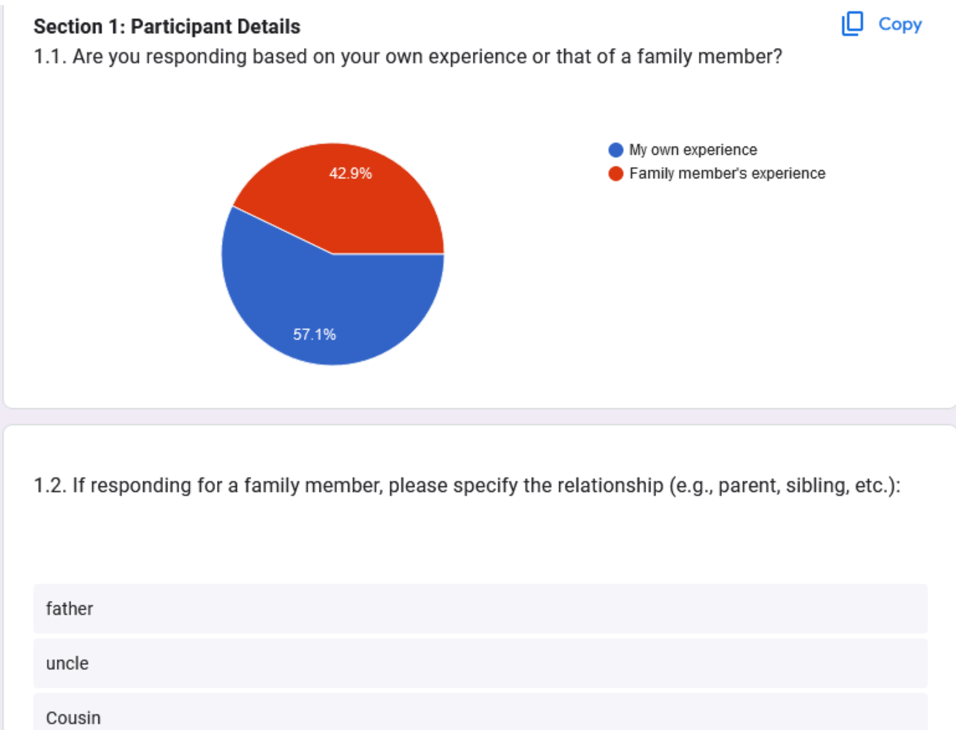
Conclusion:

Thank you for sharing your insights. Your input is valuable in enhancing our understanding of the holistic impact of cardiovascular health on various aspects of well-being. If you have any additional comments or experiences you'd like to share, please feel free to do so.

Submit

Clear form

Figure B - 4 - Questionnaire



Section 1: Participant Details

Copy

1.1. Are you responding based on your own experience or that of a family member?

1.2. If responding for a family member, please specify the relationship (e.g., parent, sibling, etc.):

father

uncle

Cousin

Figure B - 5 - Responses

Section 2: BMI (Body Mass Index)

2.1. What is/was the BMI (Body Mass Index) of the individual?

18

25

20.9

27.3

22

25.7

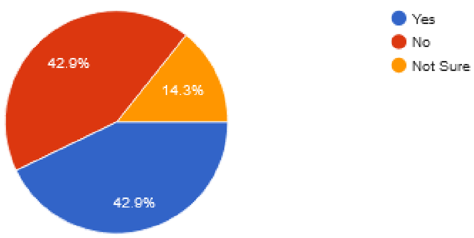
20.2

Figure B - 6 - Responses

Copy

Section 3: Mental Health in the Past 30 Days

3.1. Have there been any noticeable changes in mental well-being during this period?



3.2 For how many days during the past 30 days the mental health of the individual was not good?

20

10

-

30 Days. Last few months he was under lot of mental pressure

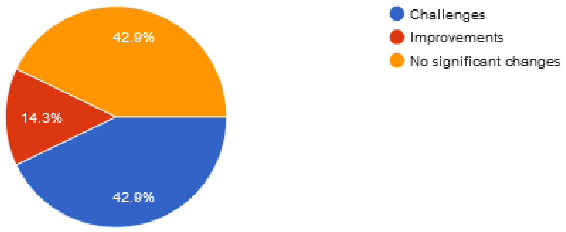
some days at work

Figure B - 7 - Responses

Copy

Section 4: Physical Health in the Past 30 Days

4.1. Have there been any specific challenges or improvements in physical health during this period?



4.2 For how many days during the past 30 days the physical health of the individual was not good?

25

5-6

-

He seemed usual.

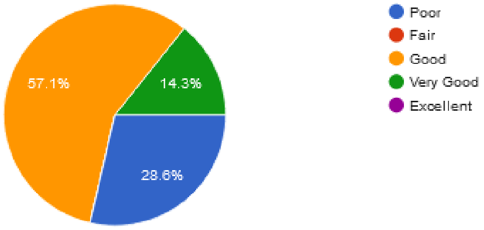
0

Figure B - 8 - Responses

Copy

Section 5: General Health

5.1. How would you describe the overall general health of the individual?



Section 6: Sleep Time in Hours

6.1. On average, how many hours of sleep does the individual get per night?

4

6-7

5 hours

5 Hours

Figure B - 9 - Responses

APPENDIX C: UI of the Web Based System

The system consists of following main UIs

- Registration and Login to the system
- Enter user details
- Fill the questionnaire
- Analyze data and receive the risk factor of CVD - The four functions built in the machine learning model will be used to get the probabilities of Heart Diseases Risk, Stroke Risk, Mental Diseases Risk and Heart Diseases/Stroke Risk when having a Mental Diseases Risk as percentages

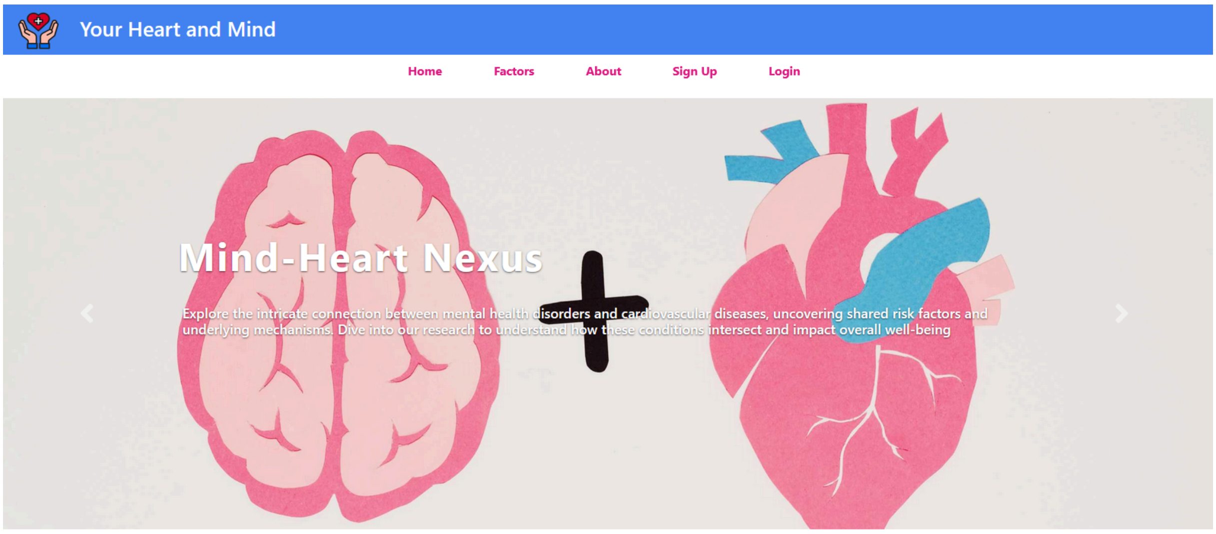


Figure C - 1 - UI of the Home Page

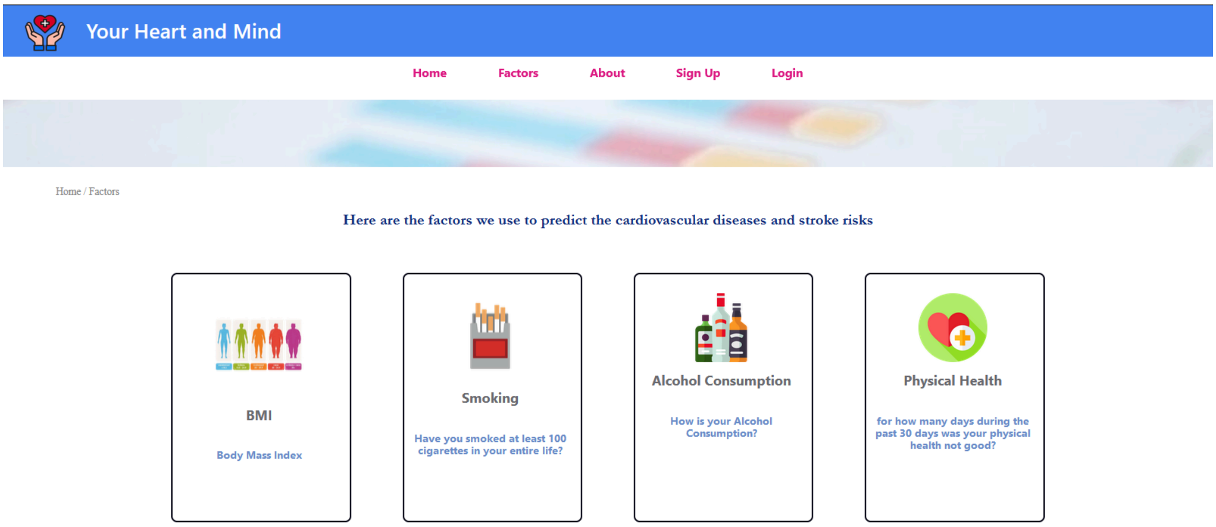


Figure C - 2 - UI of the Factors Page

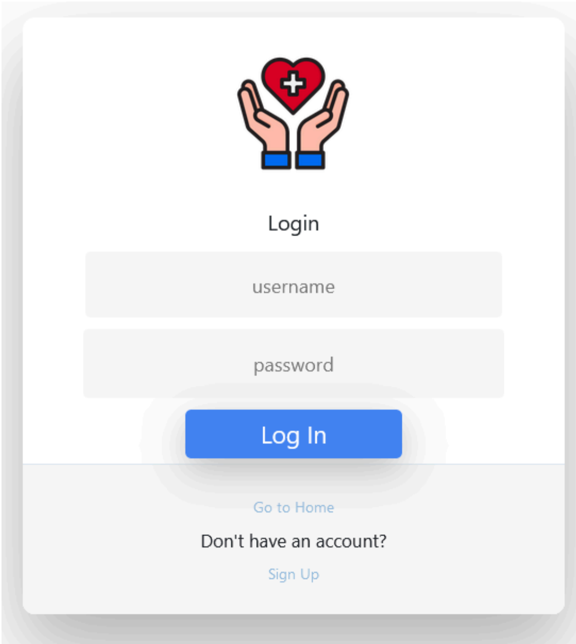


Figure C - 3 - Login Page

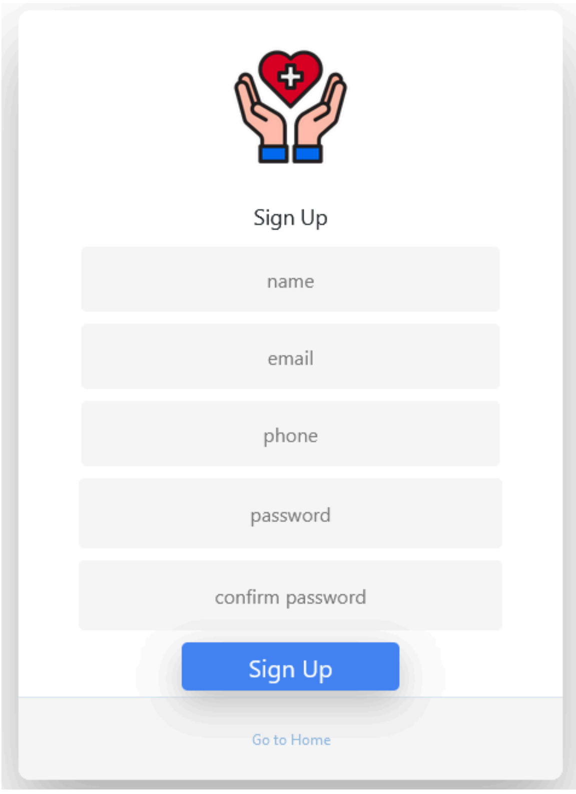


Figure C - 4 - Registration Page

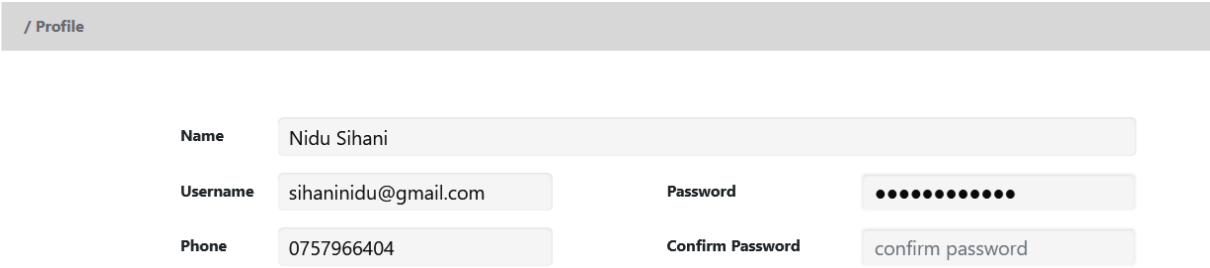


Figure C - 5 - Data entry screen to enter the user details

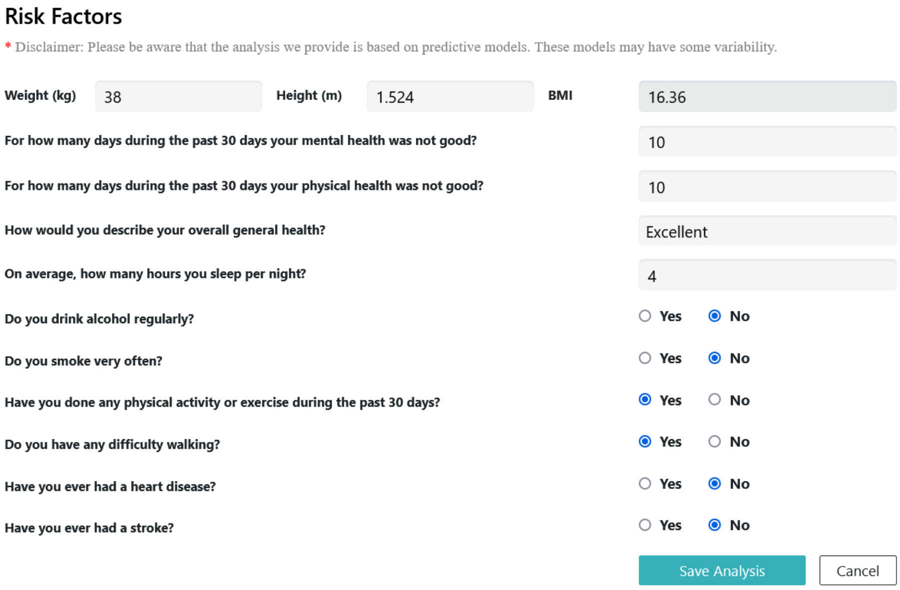


Figure C - 6 - Data entry screen to fill the questionnaire

Figure C - 7 - Risk Analysis screen (Heart Diseases and Stroke Risk Percentages)

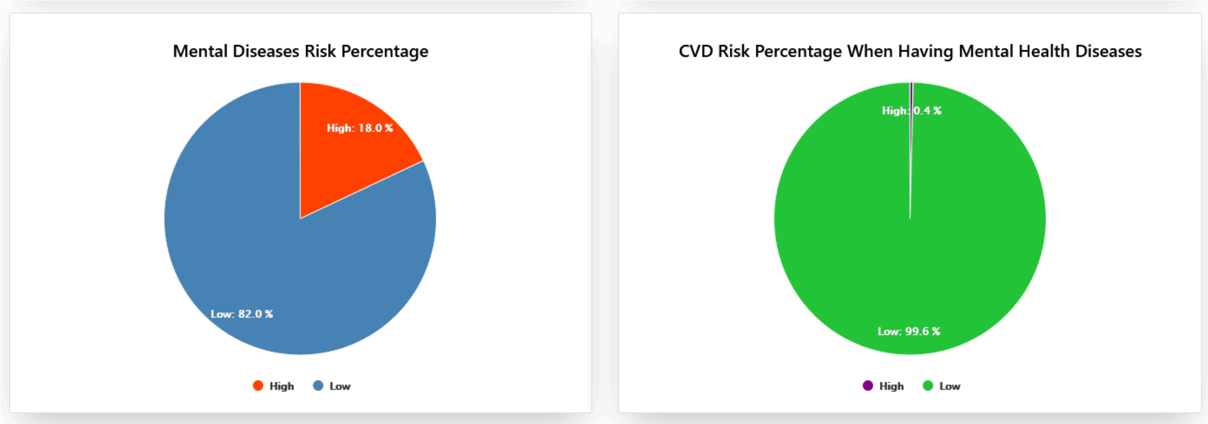
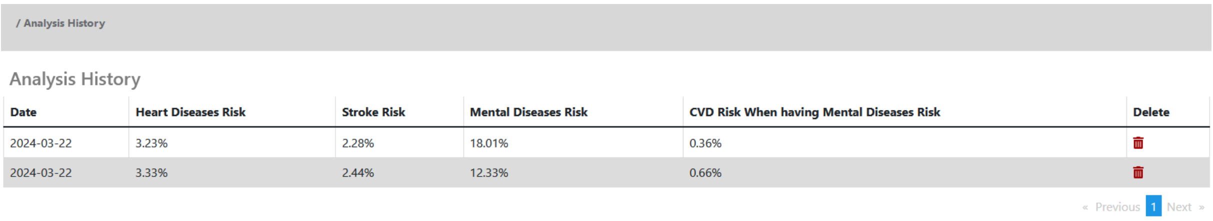


Figure C - 8 - Risk Analysis screen (Mental Diseases Risk and CVD Risk Percentages)

Figure C - 9 - Risk Analysis screen (Association of Each Risk Factor in Heart Diseases and Stroke)


/ Analysis History

Analysis History

Date	Heart Diseases Risk	Stroke Risk	Mental Diseases Risk	CVD Risk When having Mental Diseases Risk	Delete
2024-03-22	3.23%	2.28%	18.01%	0.36%	
2024-03-22	3.33%	2.44%	12.33%	0.66%	

« Previous 1 Next »

Figure C - 10 - Risk Analysis History