BC2407 - Analytics II: Advanced Predictive Techniques

Semester 2, AY2023/2024

Instructor: Professor Neumann Chew

Seminar Group: S01

Group: 5

Name	**Matriculation Number**
Sally Ngui Yu Ying	U2222782A
Tenia Xu Yuan	U2210808E
Lum Shi Zhen	U2210198H
Poh Lee Tin	U2111981A
Teo Man Ru Joleen	U2111561F

Table of Contents

1. Executive Summary

Cardiovascular disease (CVD) is the leading cause of death globally, taking an estimated 17.9 million lives each year (World Health Organization, n.d.). It has become a global health concern, and Singapore is no exception. Singapore faces unique challenges as a thriving urban centre due to its rapid economic growth, changing lifestyles, and cultural influences.

In 2023, NHCS was ranked 57th in Newsweek's list of the World's Best Specialised Hospitals for Cardiology (Cooper, 2023). Our team believes that our analysis will greatly benefit NHCS in their research and services, ultimately achieving greater operational efficiency.

Our project, Project HeartBeat, seeks to investigate the factors driving CVD by developing and evaluating predictive models. The models in this report aim to predict the risk of individuals developing CVD and identify the key contributing factors of CVD. By accurately predicting the likelihood of individuals developing CVD, NHCS can identify individuals at higher risk of developing CVD and prioritise them for further diagnosis testing and preventive interventions. This proactive approach can lead to early detection, timely treatment, and better management of CVD, ultimately reducing the burden on healthcare systems and improving patient outcomes.

The dataset we sourced is 2022 Behavioural Risk Factor Surveillance System Survey Data, obtained from the Centers for Disease Control and Prevention (CDC) due to its wide variety of health-related factors and accessibility.

Firstly, we explored the different variables using exploratory data analysis (EDA) to identify general patterns in our dataset, including patterns and features of the data that might be unexpected. These insights allow us to examine the relations between the variables and CVD, identifying key contributing factors of CVD and helping patients reduce their likelihood of heart disease through education.

Secondly, we utilised four different models: Logistic Regression, Classification and Regression Tree (CART), Neural Network, and Random Forest. After comparing the accuracy of the four models, we finalised that Logistic Regression is the most optimal model for detecting the presence of heart disease.

Thirdly, the final model, Logistic Regression, will be implemented in the suggested application Cardiac Health Monitoring and Prediction (CHAMP). This empowers individuals to self-check their risk of CVD conveniently and receive personalised risk predictions. For those at risk, individuals will be recommended on subsequent steps, such as an appointment with the doctor, while everyone benefits from educational resources related to CVD and quarterly reassessments.

Overall, the team consolidated the findings and discussed the limitations of the dataset and models for users to take note of. Further recommendations were also proposed to enhance the existing business implementation. We believe that the insights from the predictive models and analysis can help NHCS gain credibility by predicting heart diseases more accurately. It will also transform NHCS into a more cost-efficient organisation by reducing resources spent on delayed diagnosis of CVD.

2. Business Understanding

2.1 Introduction

The high mortality rate attributed to CVD underscores the critical need for early detection and preventive measures (Appendix A, Figure 7.1.1) (Dattani et al., 2023). Despite advancements in medical science, CVD remains a leading cause of death globally, affecting individuals, healthcare systems, and society at large. To effectively combat this pervasive health issue, it is imperative to address several key challenges and barriers patients and healthcare providers face.

2.2 Business Problem

Despite the significant strides made in medical science, the persistently high mortality rate attributed to CVD underscores a critical gap in early detection and preventive measures. The process of undergoing diagnostic tests for CVD often proves to be time-consuming, due to long waiting times, therefore delaying crucial treatment initiation and worsening patient outcomes.

2.2.1 Hospitals

In addition to the challenges posed by time-consuming diagnostic procedures, healthcare providers face many obstacles, including manpower shortages and the high costs associated with both diagnosis and treatment. Furthermore, the cost of treating CVD at advanced stages far exceeds the expenses incurred for early detection and prevention, highlighting the urgency for innovative solutions to enhance efficiency and reduce costs in patient care and healthcare operations (Brouwer et al., 2015).

2.2.2 Individuals

Compounding these challenges is the reluctance exhibited by individuals to undergo regular checkups due to various factors, including perceived high costs (ranging from $70 to $1,299, as shown in Appendix A, Figure 7.1.2), fear of diagnosis, and apprehension about potential health outcomes. This reluctance hinders timely intervention and preventive measures, contributing to the progression of CVD and worsening patient outcomes.

2.3 Opportunities provided by analytics

NHCS has a prime opportunity to enhance its operational efficiency and cost-effectiveness through the application of analytics. Data analytics offers invaluable insights and solutions to enhance clinical capabilities and improve patient care and resource allocation. As such, NHCS can gain valuable insights into patient demographics, risk factors, and disease trends, facilitating more targeted preventive interventions and optimising resource allocation to improve overall operational efficiency.

Research conducted in collaboration with Boston-area hospitals illustrates the potential impact of predictive analytics in healthcare. Predictive models have successfully forecasted hospitalizations related to chronic illnesses such as heart disease and diabetes up to a year in advance with an impressive accuracy rate of up to 82% (Paschalidis, 2017). By leveraging machine-learning algorithms, healthcare providers can intervene earlier, preventing costly hospitalizations and improving patient outcomes.

Moreover, analytics has shown promise in significantly reducing unnecessary hospitalizations, thereby cutting costs and improving healthcare outcomes. For instance, by preventing hospitalizations related to just two widespread chronic illnesses – heart disease and diabetes – the United States could save billions of dollars annually (Paschalidis, 2017).

As healthcare systems increasingly adopt value-based care models and assume more financial risks, analytics and technology are becoming integral components of hospital operations and revolutionising healthcare delivery. By leveraging data-driven insights and predictive analytics, NHCS can solidify its position as a leader in cardiovascular care, driving improvements in patient outcomes while achieving greater cost-efficiency and sustainability in healthcare delivery.

2.4 Opportunity Statement

The overarching problem is the lack of effective strategies for early detection and prevention of CVD, which is exacerbated by patient reluctance and constraints within the healthcare system. We seek to overcome these challenges by harnessing the power of analytics to develop a comprehensive predictive model and analytical framework for anticipating CVD.

Firstly, the key focus is on leveraging data-driven insights to enhance clinical screening processes, facilitate early detection of CVD, and ultimately improve patient outcomes. Additionally, NHCS will stand out by extending the application of these predictive models to other disease frameworks, contributing to operational excellence in the broader medical sector.

2.4.1 Predictive

NHCS aims to develop a predictive model to identify individuals at risk of developing CVD. The approach begins with thoroughly analysing diverse datasets encompassing patient demographics, medical history, lifestyle factors, and relevant clinical parameters. By leveraging data analytics techniques, NHCS can identify trends and correlations within the dataset that may indicate heightened risk factors for CVD development.

Leveraging machine learning algorithms, our predictive model will analyse these features to assign personalised risk scores, accurately anticipating CVD onset. Considering factors such as age, lifestyle habits, and medical history, our model will predict the likelihood of an individual developing CVD within a specified timeframe.

2.4.2 Preventive

Simultaneously, NHCS employs data-driven insights from the predictive model to recommend tailored preventive measures to reduce patients' risk of CVD at an early stage. These preventive measures include lifestyle modifications, dietary changes, exercise regimens, and personalised medication management strategies tailored to each patient's risk profile. By proactively addressing key risk factors such as hypertension, obesity, smoking, and sedentary behaviour, NHCS aims to empower patients to make healthier lifestyle choices and mitigate their risk of developing CVD.

2.5 Objectives

The objective of this project is to develop a predictive model for CVD based on non-medical data to make it accessible to the general public. By leveraging these factors, the aim is to create a tool that can generate a CVD risk prediction anytime, anywhere, and by anyone. This model seeks to democratise access to CVD risk prediction, empowering individuals to take proactive steps towards their cardiovascular health. Through this approach, we aim to improve early detection and prevention of heart disease, ultimately leading to better health outcomes for all.

3. Data and Methodology

3.1 Data Sourcing

Project HeartBeat aims to build models to predict whether people are at risk for CVD. Therefore, we require a dataset with a wide variety of factors that are correlated with CVD and might lead to the onset of CVD.

Although it would be ideal to have a local dataset or a dataset from NHCS, these datasets are not publicly available. Therefore, our next best alternative was to find a similar governmental dataset from other countries.

Our dataset is the 2022 Behavioural Risk Factor Surveillance System Survey Data, obtained from the Centers for Disease Control and Prevention (CDC), the national public health agency of the United States (Centers for Disease Control and Prevention, 2023). It collects data, such as health-related risk behaviours and chronic conditions, via telephone surveys with over 400,000 adults from 50 states. As a governmental dataset, it is well-documented and less susceptible to biases.

Our dataset includes factors from various categories related to CVD, allowing for in-depth exploration and a comprehensive analysis of the factors that might increase the likelihood of CVD. A data dictionary for the dataset can be found in Appendix B.

1. Personal Particulars
2. Physical Characteristics
3. General Health Status
4. Health Habits & Behaviours
5. Health Issues & Illness History

3.2 Data Cleaning

The BRFSS dataset contains 445132 rows and 328 columns. However, not all columns were relevant to our project, and some columns contained inconsistent or missing values. Therefore, data was cleaned to select relevant columns and ensure the data can be used for further exploration and modelling. Since the original dataset is in XPT format (.xpt), the dataset was converted to a comma-separated values file (.csv).

3.2.1 Selection of Relevant Columns

Since this is a U.S. dataset, we selected variables that are not only relevant to CVD, but are also universal and applicable to Singapore's context.

3.2.2 Recoding Values

Based on the BRFSS codebook, many variables such as GENERAL_HEALTH represent "Don't know/Not Sure" with 7, "Refused" with 8 and "Not asked or Missing" with 9. To prevent misinterpretation of the data, we recoded these values first. Since all these labels mean data was not collected, we converted these values to null values. This is consistent with how CDC handled the data for some columns.

Furthermore, since the original dataset was in XPT format, each column consists of numerical values, including those intended to represent categorical variables. These numerical representations were converted back into their respective categorical formats to facilitate exploration and modelling.

3.2.3 Handling Null Values

Based on context, we used three different methods to handle the null values in the dataset:

1. Derivation of Values from Other Columns

Some columns are derived from other columns. Hence, we can replace the null values based on other columns. For example, if HEIGHT and WEIGHT values are not missing, BMI should not be null. Therefore, calculation was done to replace the null values with the actual BMI values.

2. Imputation with Mean

Some columns are related to other columns, but their numerical values cannot be derived from these columns. One example would be that when SMOKING is 'Yes', NO_OF_PACKS smoked should be more than 0 and not null. Imputation with mean was done to replace the null values since it cannot be derived from any column in the dataset.

3. Placeholder

For columns such as INCOME, there might be various reasons why the data was omitted, including the fact that it is a sensitive topic that individuals might be uncomfortable revealing. A placeholder such as 'Missing' was used to replace the null values due to a significant number of missing rows and the difficulty of estimating the values. These rows can then be filtered out later.

If the methods above were not applicable and the estimation of values was unreliable, we decided to remove the rows with null values completely. After checking, these rows were a small proportion of the dataset, and the missing values appeared to be random, allowing the rows to be safely removed from the dataset.

3.2.4 Further Data Processing

HEIGHT, WEIGHT and BMI columns were converted to follow the SI units of height, mass and BMI for easier understanding. NO_OF_PACKS were also converted to CIGARETTES_PER_DAY to get integer values. BMI, BMI_CATEGORY and OVERWEIGHT_OR_OBESE were recalculated and recategorised according to Singapore's standards.

3.2.5 Removing Duplicates

All 281 rows of duplicates were removed to prevent unnecessary redundancy and the misguiding of our models. This allows the dataset to be more streamlined.

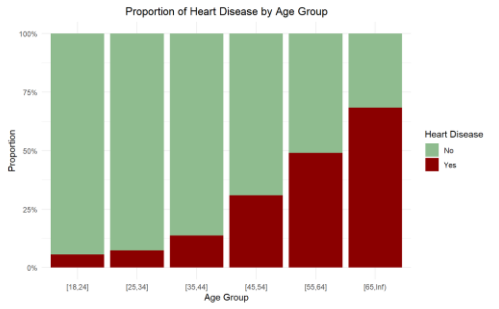
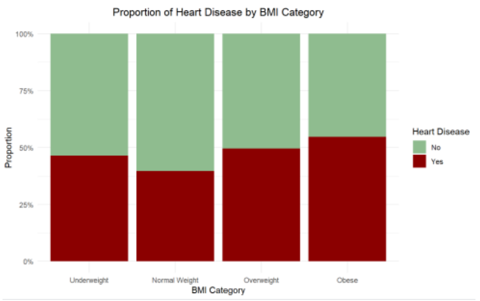
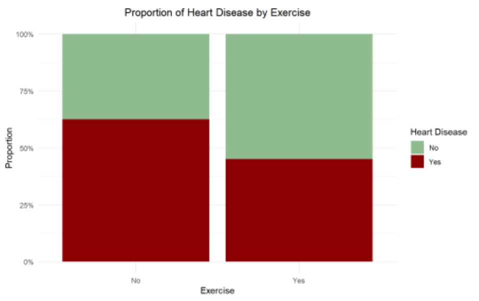
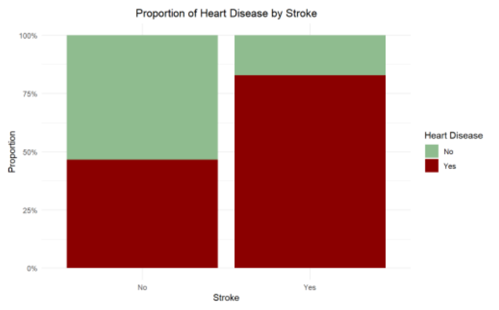
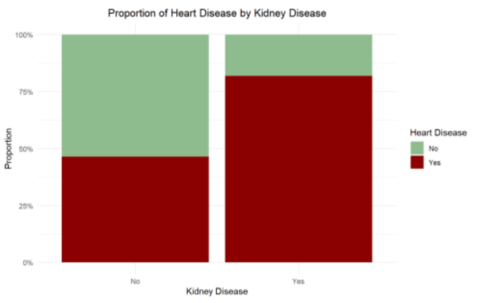
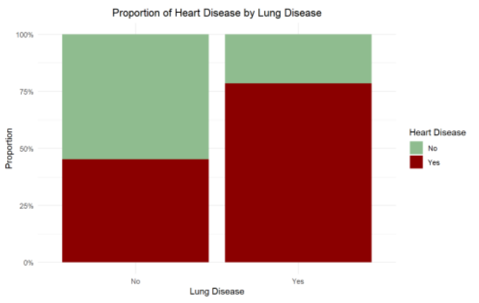
3.2.6 Results after Data Cleaning

Our final dataset, after cleaning, consists of 333374 rows and 31 columns.

3.3 Exploratory Data Analysis

After data cleaning, we conducted comprehensive data exploration to observe patterns and obtain data-driven insights.

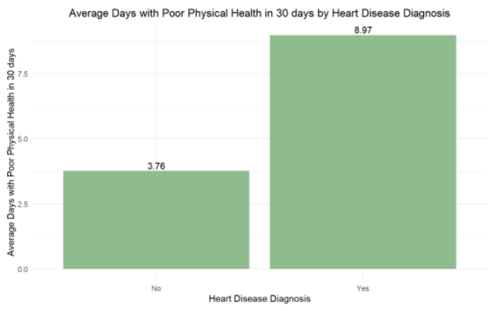
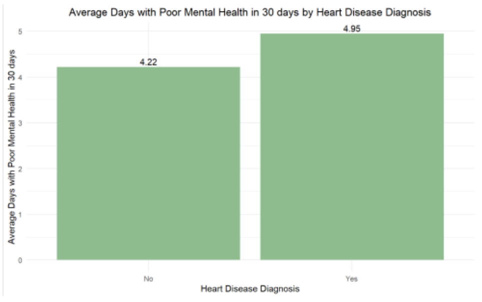
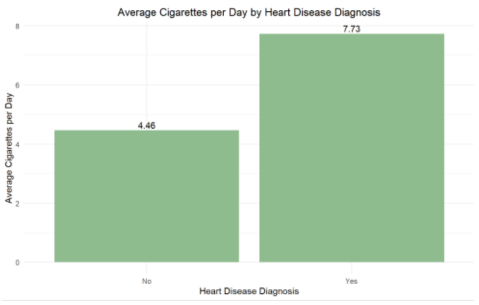
3.3.1 Bivariate EDA

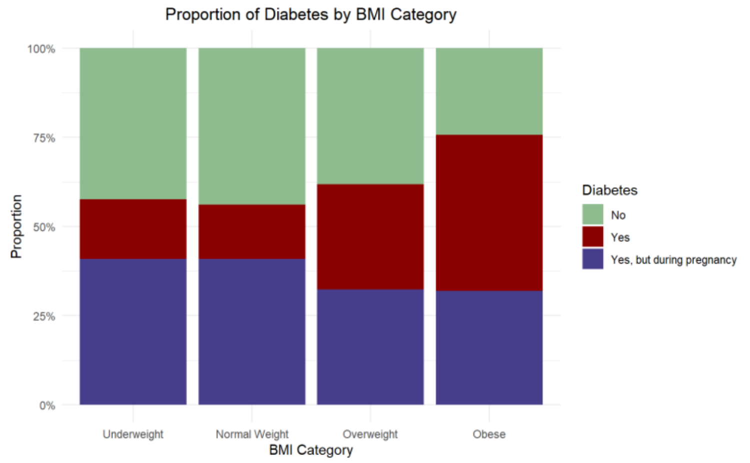
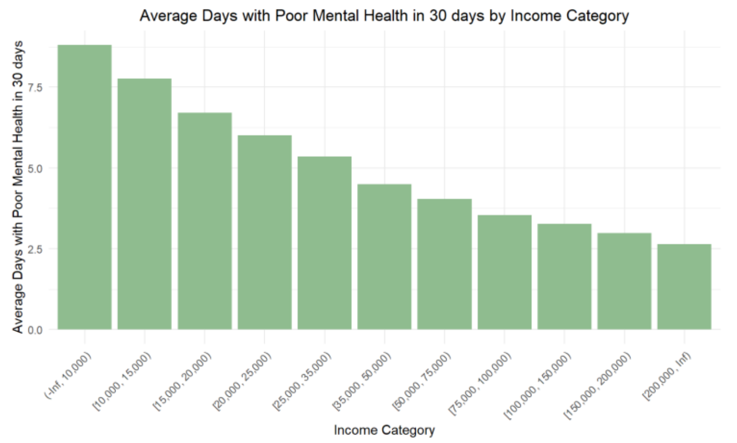
<table>
<tr><th colspan="3">HEART DISEASE VS CATEGORICAL PREDICTOR VARIABLES</th></tr>
<tr><td colspan="3">The variables are split into different categories: Personal Particulars, Physical Characteristics, Health Habits & Behaviours, Health Issues & Illness History. The other factors in certain categories follow a similar trend (Appendix C).</td></tr>
<tr><th>HEART DISEASE VS PERSONAL FACTORS</th><th>HEART DISEASE VS PHYSICAL CHARACTERISTICS</th><th>HEART DISEASE VS HEALTH HABITS & BEHAVIOURS</th></tr>
<tr><th>AGE GROUP</th><th>BMI CATEGORY</th><th>EXERCISE</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>The graph reveals a clear upward trajectory in the occurrence of heart disease with advancing age. This observation is substantiated by research indicating age is a significant risk factor that compromises the optimal functioning of the cardiovascular system (Rodgers et al., 2019). The elderly population faces a heightened risk of developing heart disease due to myriad physiological changes such as increased oxidative stress and inflammation, emphasising the importance of age-related factors in assessing cardiovascular health (Ciumărnean et al., 2021).</td>
<td>The graph highlights those who are overweight and obese are more likely to be diagnosed with heart disease. This finding aligns with research indicating that obesity contributes to various physiological mechanisms that increase the risk of cardiovascular events (Volpe & Gallo, 2023). The graph also shows that those underweight are more likely to get CVD. While there might not be a direct correlation between the two factors, research has shown that a haemoglobin deficiency, a cause of heart disease, is more prevalent among underweight individuals (Sehat, 2016).</td>
<td>The graph shows that individuals who do not engage in physical activity outside of their regular job have a higher risk of heart disease. Physical inactivity can contribute to other risk factors for heart disease, such as obesity, diabetes, and high blood cholesterol, increasing the risk of being diagnosed with heart disease (Centers for Disease Control and Prevention, 2022b).</td>
</tr>
<tr><th colspan="3">HEART DISEASE VS HEALTH ISSUES & ILLNESS HISTORY</th></tr>
<tr><th>STROKE</th><th>KIDNEY DISEASE</th><th>LUNG DISEASE</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
</table>

The graph underscores a notable connection between stroke and heart disease, indicating that individuals who have experienced a stroke may face an increased risk of developing heart disease (Centers for Disease Control and Prevention, 2022). This association is often observed because both heart disease and stroke share several key risk factors, such as high blood pressure, diabetes, and obesity.	The graph indicates a significant correlation between heart disease and kidney disease. Extensive research confirms that individuals with chronic kidney disease (CKD) face an elevated risk of developing heart disease (Jankowski et al., 2021). This heightened risk stems from the strain placed on the heart due to kidney dysfunction, which requires increased effort to circulate blood effectively to the kidneys (Mark et al., 2023) . The concurrent presence of high blood pressure exacerbates this strain on the heart and increases the likelihood of developing heart disease.	The graph shows that heart diseases are often diagnosed in Chronic Obstructive Pulmonary Disease (COPD) (Papaporfyriou et al., 2023). Evidence suggests that heart disease is the most common comorbidity in COPD patients, and the association between COPD and CVD is intricate. Systemic inflammation is an underlying cause that affects the cardiovascular system (Chen et al., 2023).

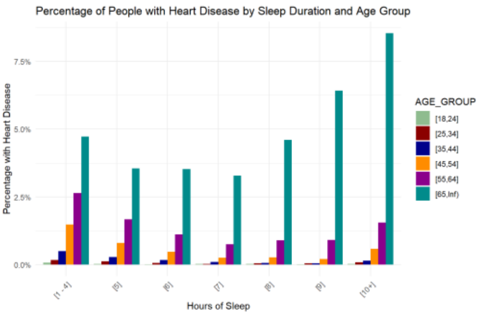
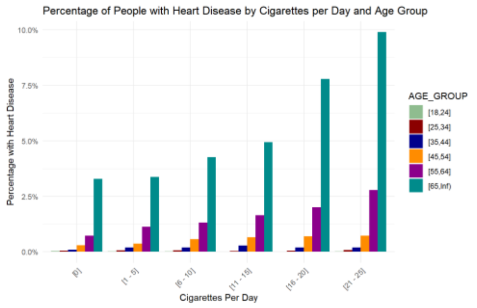
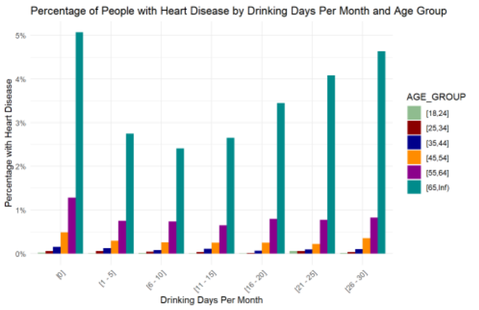
HEART DISEASE VS CONTINUOUS PREDICTOR VARIABLES

We are analysing heart disease against continuous predictor variables to observe the correlation between one another. For the variable CIGARETTES_PER_DAY, the extreme values are removed from the dataset to avoid skewness.

AVERAGE DAYS WITH POOR PHYSICAL HEALTH	AVERAGE DAYS WITH POOR MENTAL HEALTH	AVERAGE CIGARETTES PER DAY
		
The graph shows that individuals diagnosed with heart disease have an average of 9 out of 30 days of poor physical health. Research studies have shown that heart disease compromises physical health, which can lead to symptoms such as chest pain, shortness of breath, fatigue, and reduced exercise tolerance (Interior Community Health Center, 2024).	The graph indicates that individuals diagnosed with heart disease experience a higher average of 5 out of 30 days of poor mental health. Research findings suggest that symptoms of mental health disorders may emerge following an acute heart disease event (Centers for Disease Control and Prevention, 2020) . The stress of heart disease management can contribute to feelings of anxiety, depression, and overall psychological distress (Ryder, 2024).	The graph illustrates that individuals who smoke a higher average number of cigarettes per day have a higher risk of developing heart disease. This relationship is reinforced by studies showing that the compounds present in cigarette smoke promote the accumulation of plaque in blood vessels, leading to their constriction and thickening (Centers for Disease Control and Prevention, 2024). Harmful substances in cigarettes can elevate heart rate and induce inflammation (Parmar et al., 2023). These factors collectively heighten the risk of developing cardiovascular conditions.

<table>
<tr><th colspan="2">BIVARIATE ANALYSIS BETWEEN 2 PREDICTOR VARIABLES</th></tr>
<tr><th>BMI CATEGORY BY DIABETES</th><th>AVERAGE DAYS WITH POOR MENTAL HEALTH BY INCOME</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>The graph illustrates a significant association between obesity and the frequency of diabetes. Excessive body fat accumulation is a key contributor to the development of type 2 diabetes, with the risk escalating proportionally with higher body mass index levels. Consequently, the global rise in obesity rates has corresponded with a parallel surge in the incidence of type 2 diabetes. (Klein et al., 2022)</td><td>The graph shows that when the level of income earned increases, the average number of days with poor mental health in 30 days decreases. This aligns with our research that higher income allows for increased life satisfaction, meeting of personal needs and access to healthcare (Li et al., 2022). This reduces the stress that comes with financial difficulties, enhancing their mental health. This also explains the generally decreasing trend of CVD risk as income increases (Appendix C).</td></tr>
</table>

3.3.2 Multivariate EDA

<table>
<tr><th>HEART DISEASE BY SLEEP DURATION AND AGE GROUP</th><th>HEART DISEASE BY CIGARETTES PER DAY AND AGE GROUP</th><th>HEART DISEASE BY DRINKING DAYS PER MONTH AND AGE GROUP</th></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Based on the graph, those with less than 7 hours of sleep represent a higher percentage of heart disease compared to those who sleep more than 7 hours. This trend aligns with our research as a lack of</td><td>The figure shows an increasing trend of those with heart disease as age and number of cigarettes smoked per day increases. This trend is consistent with our findings on how harmful chemicals</td><td>In general, the percentage of people with heart disease increases as alcohol consumption and age increases. As excessive consumption of alcohol can increase one's risk of high blood</td></tr>
</table>

sufficient sleep can disrupt hunger hormones, causing one to indulge in more high-fat food, as well as increase blood pressure and blood sugar. (Corliss, 2022) Concluding that those sleeping less than 6 hours a day have a 20% higher incidence of heart attack. It is also notable that despite having more than 10 hours of sleep, the highest probability of suffering from heart attack is amongst those above 65 years old. A possible reason could be due to underlying health issues coupled with old age, causing one to sleep longer. Hence, those who sleep longer have an increased risk of getting heart disease by 41%. (Avramova, 2018)	from smoking promote plaque to form in blood vessels and increases one’s chance of suffering from cardiovascular disease. Furthermore, as one ages, it increases one’s chance of developing atherosclerosis; buildup of cholesterol and fats in the wall of the arteries, causing it to harden and restrict blood flow (U.S. Department of Health and Human Services, 2022). Thus, smoking coupled with old age advances plaque buildup and significantly increases one’s chance of developing cardiovascular-related disease.	pressure and affect heart health. Over time, as blood is pumped more forcefully through the arteries, it can put an immense strain over the heart and cause heart issues (Miller, 2024). However, those who do not drink have the highest percentage of people with heart disease. This phenomenon is prevalent amongst those in the age groups above 55 years old. Even though most researchers state that alcohol has a negative impact on heart health, new studies argue that moderate consumption of alcohol can reduce stress over time and lower one’s risk of cardiovascular disease (Chase, 2023). This indicates that alcohol consumption is less likely to cause heart disease compared to age and is dependent on the frequency and amount of alcohol consumed.

4. Modelling

Overall Approach

Our models aim to predict whether patients are at risk for CVD. This is done so by having HEART_DISEASE as our dependent variable with 2 categories – 'Yes' and 'No'. We then selected the best model based on their metrics.

Pre-Processing of Data

Unlike the other numerical variables, CIGARETTES_PER_DAY does not have an upper or lower limit. Therefore, extreme values were removed at the EDA stage to prevent them from skewing the results of our data analysis and hampering model performance (Jadhav, 2023).

A Z-score approach was adopted by eliminating the Z-scores of observations exceeding a predefined threshold of 3. This resulted in the removal of 8095 rows with extreme values, reducing the dataset size from 333,374 rows to 325,279 rows.

Train-Test Split

Our dataset is imbalanced as the Yes:No ratio for our dependent variable HEART_DISEASE is 1:15. This can lead to the models forming a bias towards the larger class (No), which can cause a high false negative rate (Gandler, 2020).

This is especially undesirable and serious as it might falsely diagnose a patient who is at risk for CVD as having no risk, which can be fatal. Therefore, an undersampling method was used to balance the train set and test set after the 70-30 train-test split to a Yes:No ratio of 1:2. This allows a more balanced train set and test set with a Yes:No ratio of 13068:26136 and 5601:11202 respectively, improving our models' ability to predict the risk of CVD and reducing false negative rates.

4.1 Modelling

We utilised 4 different models to predict CVD and evaluated the best model.

4.1.1 Logistic Regression Model

Logistic Regression is a statistical method that models the relationship between a categorical dependent variable and one or more independent variables by predicting the probability of an event's occurrence through a logistic curve, ensuring predictions fall within the range of 0 to 1, making it suitable for classification tasks.

Firstly, the model was built based on insights derived from EDA which identifies relevant variables showing strong associations with the target variable, ensuring the retention of only impactful predictors (Appendix D, Figure 7.4.1.1).

Next, backward stepwise selection was employed to iteratively remove predictors, on a step-by-step basis, based on statistical criteria like p-values or information gain until an optimal subset of variables is obtained (Appendix D, Figure 7.4.1.2).

Following stepwise selection, variance inflation factor (VIF) analysis was conducted to detect multicollinearity, with variables possessing high VIF values indicating high variable correlation and potential issues; however, no modification was necessary in this instance as no multicollinearity was observed in our model (Appendix D, Figure 7.4.1.3).

Lastly, variables with high p-values, namely DRINKING and EXERCISE, were removed from the model due to their lack of statistically significant impact on the target variable, resulting in the optimal Logistic Regression model (Appendix D, Figure 7.4.1.4).

4.1.2 CART – Classification Tree

Classification and Regression Tree (CART) is a predictive modelling technique that uses a decision tree structure to analyse datasets based on a specified target variable. CART is valued for its interpretability, as it produces a visual representation of the decision-making process in the form of a tree diagram.

Firstly, the classification tree is grown to its maximum depth. Next, the complexity parameter is tuned for pruning the tree using cross-validation (Appendix D, Figure 7.4.2.2) and the optimal complexity parameter is determined based on the cross-validation results. Finally, the tree is pruned using this optimal complexity parameter of 0.00118497, resulting in the optimal CART model. (Appendix D, Figure 7.4.2.3)

4.1.3 Random Forest

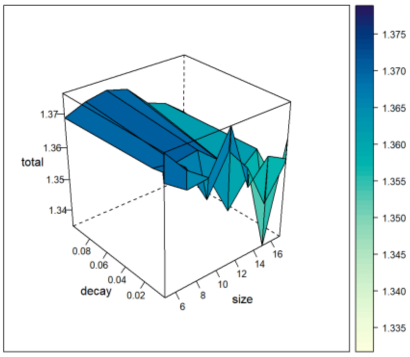
Random forest is a machine learning algorithm that constructs multiple decision trees through bagging during the training phase and gives a final prediction by calculating the mode of the classes of individual trees.

At each tree split, a randomly chosen subset of the relevant predictor variables is selected to determine the best split. This ensures stability and prevents any dominant predictor variable from dominating the model.

The optimal model was determined by minimising and stabilising the Out-of-Bag (OOB) error, which we have observed to stabilise after 100 to 150 trees (Appendix D, Figure 7.4.3.2). The OOB error was found to be 23.92%.

4.1.4 Neural Network

A neural network is a machine-learning algorithm inspired by the human brain. It consists of layers of interconnected nodes, or artificial neurons, organised in input, hidden, and output layers. The first layer receives raw input, which is processed by multiple hidden layers, and the last layer produces the result. Through training, neural networks can learn complex patterns and relationships, making them a powerful tool.



Hyperparameter tuning is done to find the optimal number of nodes within the single hidden layer and weight decay. Finding the optimal number of nodes is crucial to ensure that there are enough nodes to capture patterns, but not too many to prevent overfitting.

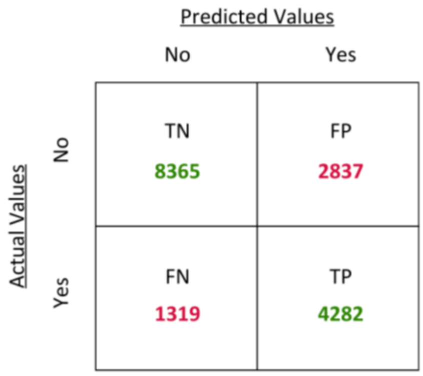
Through grid search, the optimal size of 17 nodes and weight decay of 1e-05, gives the best model accuracy and the optimal Neural Network model. (Appendix D, Figure 7.4.4.1).

4.2 Model Evaluation

We used the confusion matrix to determine the performance of our classification models, showcasing the proportion of correct and incorrect predictions. From the metrics, the accuracy level was further calculated.

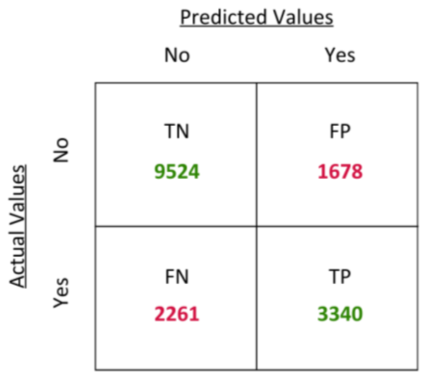
For random forest, we evaluated its performance through two key metrics: the OOB error rate and the accuracy of the test set. Since both metrics provided similar results and accuracy, we decided to use the OOB confusion matrix to compare against the other models.

Logistic Regression



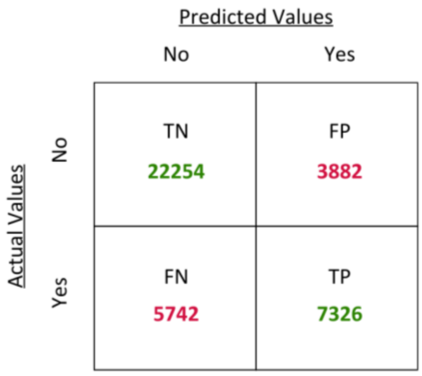
		Predicted Values	
		No	Yes
Actual Values	No	TN 8365	FP 2837
	Yes	FN 1319	TP 4282

CART



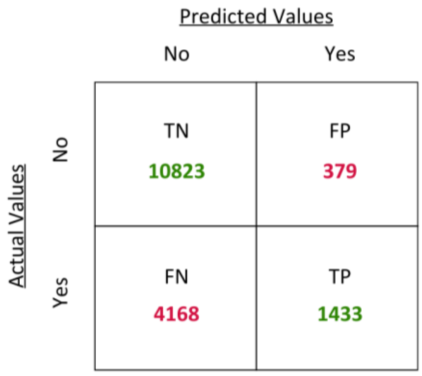
		Predicted Values	
		No	Yes
Actual Values	No	TN 9524	FP 1678
	Yes	FN 2261	TP 3340

Random Forest



		Predicted Values	
		No	Yes
Actual Values	No	TN 22254	FP 3882
	Yes	FN 5742	TP 7326

Neural Network



		Predicted Values	
		No	Yes
Actual Values	No	TN 10823	FP 379
	Yes	FN 4168	TP 1433

Model/Metrics	TPR (%)	FNR (%)	FPR (%)	TNR (%)	Accuracy (%)
Logistic Regression	76.5	23.5	25.3	74.7	75.3
CART	59.6	40.4	15.0	85.0	76.6
Random Forest	56.1	43.9	14.9	85.1	75.5
Neural Network	25.6	74.4	3.4	96.6	72.9

To determine the best model, we need to recognise that in the context of predicting CVD risk, a higher false positive rate (FPR) is deemed acceptable to make sure that no potential cases of CVD were missed. This approach of maximising true positive rate (TPR) and minimising false negative rate (FNR) prioritises safety and precaution, allowing for early intervention measures even if they are not actually at risk.

From the performance metrics of the four models, while they have similar accuracy levels of approximately 75%, the Logistic Regression model produces the most optimal results of having the highest TPR of 76.5% – it has a 76.5% accuracy in correctly identifying individuals at risk for CVD, allowing them to take measures promptly.

4.3 Business Insights from Model

4.3.1 Variable Importance

Upon analysing the Logistic Regression model, certain variables emerge as significant predictors of CVD risk (Appendix D, Figure 7.4.1.6). Variables such as AGE_GROUP, GENERAL_HEALTH and STROKE, appear to have notable impacts on the likelihood of developing CVD.

Understanding the importance of these variables can guide healthcare professionals in prioritising risk factors during patient assessments and interventions, enabling more targeted and effective preventive strategies.

4.3.2 Odds Ratio for Logistic Regression

Calculating the odd ratios associated with each predictor allows for a deeper understanding of their individual contributions to higher risk of CVD (Appendix D, Figure 7.4.1.5).

For example, a higher odds ratio (14.729) in AGE_GROUP for individuals aged above 65 signifies a heightened risk of CVD—when an individual is aged above 65, the odds of them developing CVD multiply by 14.729. This emphasises the importance of age-specific screening and intervention strategies to address this demographic's unique healthcare needs.

Another example, a high odds ratio (7.600) for poor GENERAL_HEALTH signifies a strong correlation with increased CVD risk, highlighting the critical importance of addressing overall health status in CVD prevention by prioritising interventions aimed at improving general health, such as lifestyle modifications, chronic disease management, and access to primary care services, to effectively reduce the risk of CVD.

Leveraging these insights empowers healthcare providers to implement targeted interventions and resource allocations, ultimately improving patient outcomes.

5. Proposed Business Implementation and Value Proposition

5.1 NHCS' Current Measures

Before we propose a possible business implementation, it is important to examine NHCS' current measures to analyse their effectiveness and identify areas for improvement.

Currently, NHCS employs a comprehensive approach to treating CVD. They use a range of cardiac tests, including the Echocardiogram (ECG), Cardiac Computed Tomography (CT) Scan and Exercise Stress Test (Treadmill Exercise), to create accurate diagnoses and assessments (Lau , 2019).

In addition to medical diagnostic tests, in recent years NHCS has also integrated an Artificial Intelligence (AI) tool with a 98.5% accuracy rate (Jahmunah et al., 2021). The tool uses the Gabor-Convolutional Neural Network (Gabor-CNN) algorithm to train it to recognise ECG patterns, increasing the efficiency and accuracy of CVD diagnosis (NANYANG TECHNOLOGICAL UNIVERSITY, 2023) (Appendix E, Figure 7.5.1).

However, several issues have been identified in the current approach:

1. **Limited Improvement in Efficiency and Cost:** Despite the increased efficiency the AI tool provides, it still requires the ECG test to be conducted, which remains costly and time-consuming (Section 2.2).
2. **Lack of Early Prevention Tools and Measures:** The diagnostic tests and the AI tool focus on diagnosing CVD. However, they do not cater to individuals who may not yet have CVD but are considered at risk. This causes delayed diagnoses for such individuals, often only when they present symptoms of CVD.
3. **Limited Predictive Variables:** The AI tool relies solely on ECG signals and patterns to predict CVD. However, non-medical factors, such as lifestyle and habits, also play a crucial role in determining an individual's risk for CVD.

While the cardiac tests and the AI tool are rather comprehensive in diagnosing CVD and providing post-treatment monitoring, there is a lack of early risk assessment tools that are accessible and affordable for the general public.

5.2 Proposed Business Implementation

To address the limitations of NHCS' current measures and achieve the aim of providing predictive and preventive measures for CVD, as mentioned in Sections 2.4 and 2.5, we propose to introduce the Cardiac Health Monitoring and Prediction (CHAMP) App. This app allows the general public to fill in the risk assessment and predict whether they are at risk for CVD. This provides accessibility to the public with non-medical knowledge, allowing them to understand their potential risk of CVD so that they can take early preventive measures. It will be used alongside the current reactive measures to create a comprehensive strategy for tackling CVD. The flowchart that illustrates how the app can be implemented and how the public can utilise it is shown below (Figure 5.2.1).

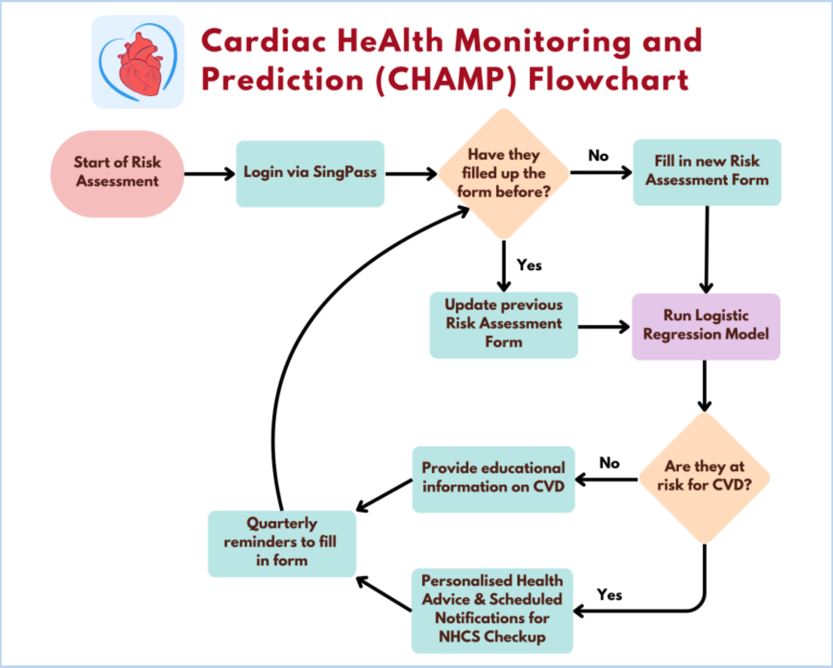

Cardiac HeAlth Monitoring and Prediction (CHAMP) Flowchart
Start of Risk Assessment
Login via SingPass
Have they filled up the form before?
No
Fill in new Risk Assessment Form
Yes
Update previous Risk Assessment Form
Run Logistic Regression Model
Are they at risk for CVD?
No
Provide educational information on CVD
Yes
Personalised Health Advice & Scheduled Notifications for NHCS Checkup
Quarterly reminders to fill in form

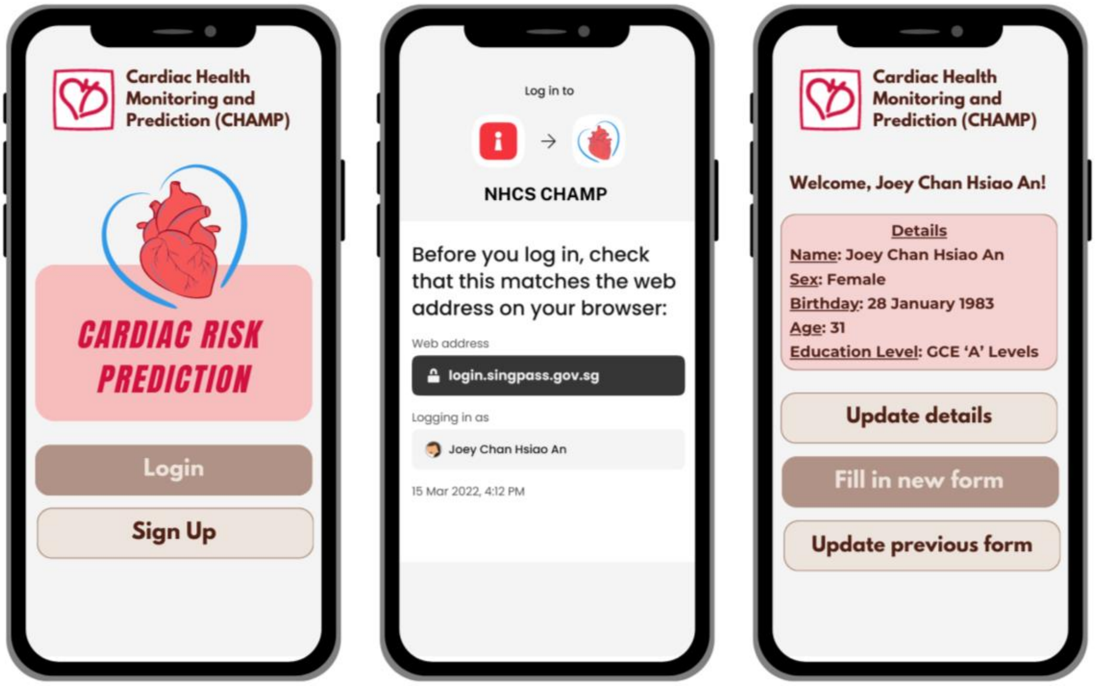
Figure 5.2.1: Flowchart for CHAMP Application

The comprehensive features in the CHAMP App are split into two types – predictive and preventive.

5.2.1 Predictive Features – Using Logistic Regression for Risk Prediction

Feature 1: Automatic Retrieval of Details

Upon login, users can link their SingPass to automatically retrieve their personal and health-related information required for risk prediction (Figure 5.2.2). This simplifies the setup and form-filling process in Feature 2.



Before you log in, check that this matches the web address on your browser:

Web address

login.singpass.gov.sg

Logging in as

Joey Chan Hsiao An

15 Mar 2022, 4:12 PM

Welcome, Joey Chan Hsiao An!

Details

Name: Joey Chan Hsiao An

Sex: Female

Birthday: 28 January 1983

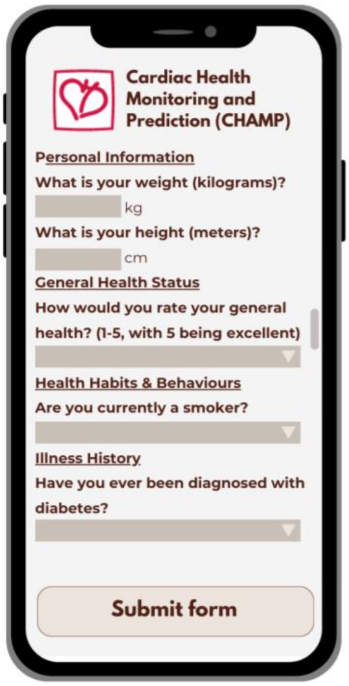
Age: 31

Education Level: GCE 'A' Levels

Figure 5.2.2: Login Page and Automatic Retrieval of Details with SingPass

Feature 2: Risk Assessment Form

Users will have to complete a comprehensive risk assessment form consisting of risk factors from various categories, such as their general health status and illness history (Figure 5.2.3).



Cardiac Health Monitoring and Prediction (CHAMP)

Personal Information

What is your weight (kilograms)?

kg

What is your height (meters)?

cm

General Health Status

How would you rate your general health? (1-5, with 5 being excellent)

Health Habits & Behaviours

Are you currently a smoker?

Illness History

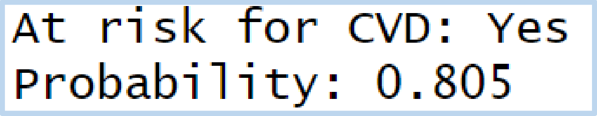
Have you ever been diagnosed with diabetes?

Figure 5.2.3: Snippet of Risk Assessment Form in CHAMP

Feature 3: Utilisation of Logistic Regression Model

The factors from the risk assessment form will be treated as input variables for our logistic regression model. After running the model, it will predict whether the individual is at risk for CVD.

This is an example of the input factors that an individual can key in, allowing the model to generate a prediction of their risk for CVD (Figure 5.2.4).

<table>
<tr><td colspan="2">
At risk for CVD: Yes
Probability: 0.805
<i>Figure 5.2.4: Sample Output of CVD Risk Prediction</i></td></tr>
<tr><td><b>Personal Particulars</b></td><td>AGE_GROUP: [55,64]
EDUCATION: Graduated College
UNABLE_TO_AFFORD_MED: Yes</td></tr>
<tr><td><b>Physical Characteristics</b></td><td>BMI_CATEGORY: Overweight</td></tr>
<tr><td><b>General Health Status</b></td><td>GENERAL_HEALTH: Poor
PHYSICAL_HEALTH: 5</td></tr>
<tr><td><b>Health Habits & Behaviours</b></td><td>CIGARETTES_PER_DAY: 10</td></tr>
</table>

Health Issues & Illness History	STROKE: Yes SKIN_CANCER: No OTHER_CANCER: No LUNG_DISEASE: No KIDNEY_DISEASE: No ARTHRITIS: No DIABETES: Yes

5.2.2 Preventive Features for Scenario 1 – Individual predicted to be at risk for CVD

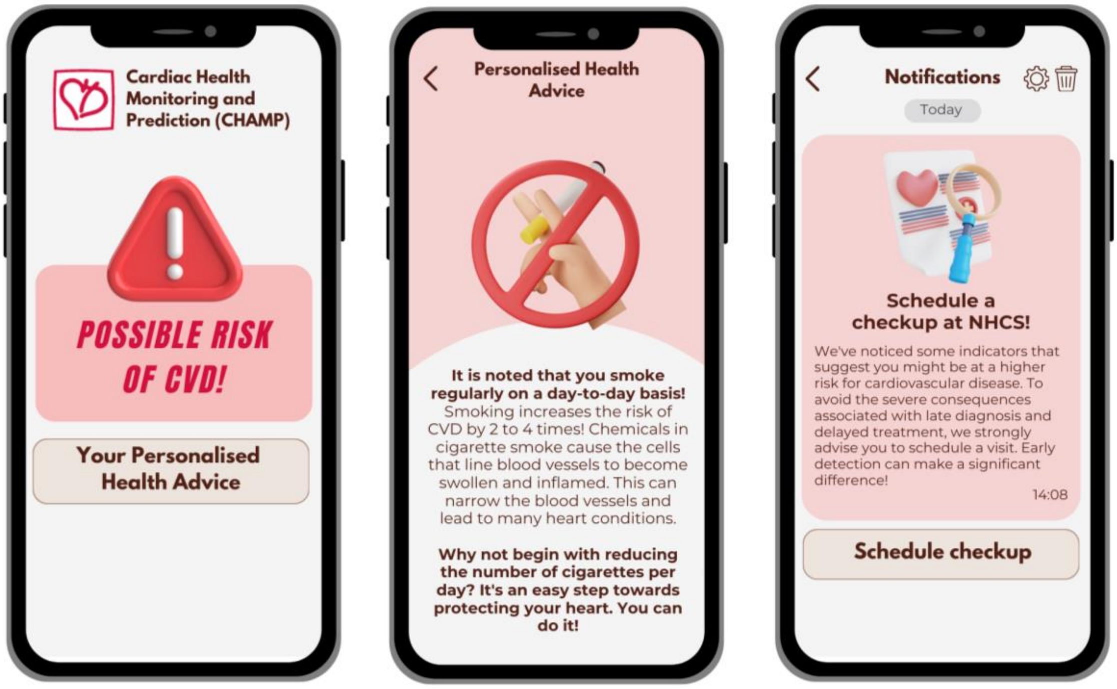
Feature 1: Personalised Health Advice

Users identified as at risk for CVD will receive personalised recommendations for preventive measures, such as lifestyle modifications. They will be provided with concrete measures and encouraged to take small steps, as changing their habits takes time. This aims to reduce individual risk factors and their overall risk of CVD.

For example, with reference to the example in Figure 5.2.4, since they recorded that they smoke 10 cigarettes per day, the system will recommend measures to reduce the effects of smoking, such as by gradually reducing the number of cigarettes smoked every day.

Feature 2: Scheduled Check-up Notifications

While lifestyle and behavioural changes are important in reducing the risk of CVD, it is still crucial for individuals to receive help from a professional. At-risk individuals will receive notifications regularly to schedule a heart check-up at NHCS. This ensures they take tangible steps to seek medical help and receive early treatment, preventing serious consequences of the further progression of CVD. The implementation of Features 1 to 2 is shown in Figure 5.2.5.


Cardiac Health Monitoring and Prediction (CHAMP)
POSSIBLE RISK OF CVD!
Your Personalised Health Advice
Personalised Health Advice
Notifications
Today
Schedule a checkup at NHCS!
14:08
Schedule checkup

It is noted that you smoke regularly on a day-to-day basis! Smoking increases the risk of CVD by 2 to 4 times! Chemicals in cigarette smoke cause the cells that line blood vessels to become swollen and inflamed. This can narrow the blood vessels and lead to many heart conditions.

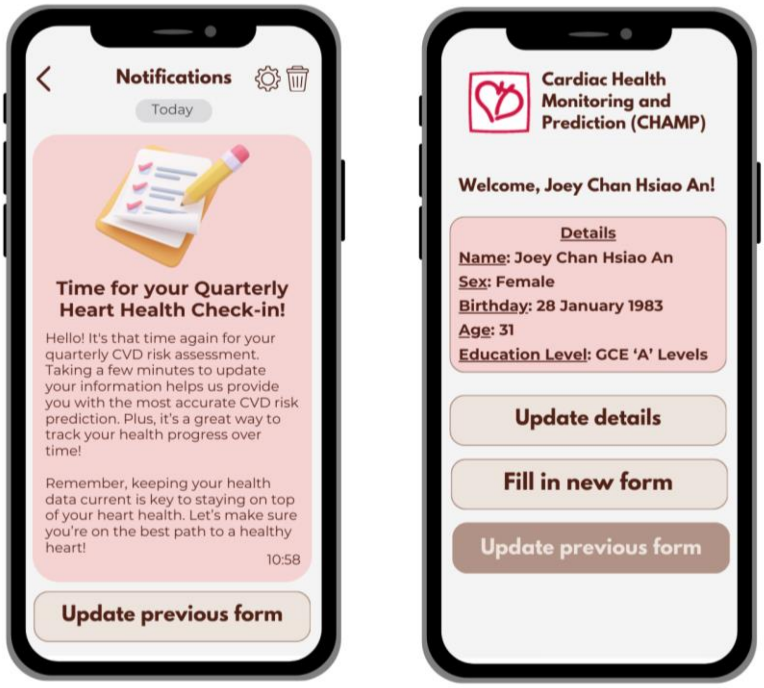
Why not begin with reducing the number of cigarettes per day? It's an easy step towards protecting your heart. You can do it!

We've noticed some indicators that suggest you might be at a higher risk for cardiovascular disease. To avoid the severe consequences associated with late diagnosis and delayed treatment, we strongly advise you to schedule a visit. Early detection can make a significant difference!

Figure 5.2.5: Scenario 1 – Predicted at risk for CVD, Personalised Health Advice and Notifications to Schedule Checkup at NHCS

Feature 3: Quarterly Reminders for Updated Risk Assessment

To monitor their progress and keep them up to date with their risk assessment, the CHAMP App will send quarterly notifications to urge users to update their risk assessment form to reassess their CVD risk (Figure 5.2.6). This continuous monitoring ensures dynamic adjustments in personalised health advice and motivates at-risk individuals to improve their heart health.



Time for your Quarterly Heart Health Check-in!

Hello! It's that time again for your quarterly CVD risk assessment. Taking a few minutes to update your information helps us provide you with the most accurate CVD risk prediction. Plus, it's a great way to track your health progress over time!

Remember, keeping your health data current is key to staying on top of your heart health. Let's make sure you're on the best path to a healthy heart!

10:58

Welcome, Joey Chan Hsiao An!

Details

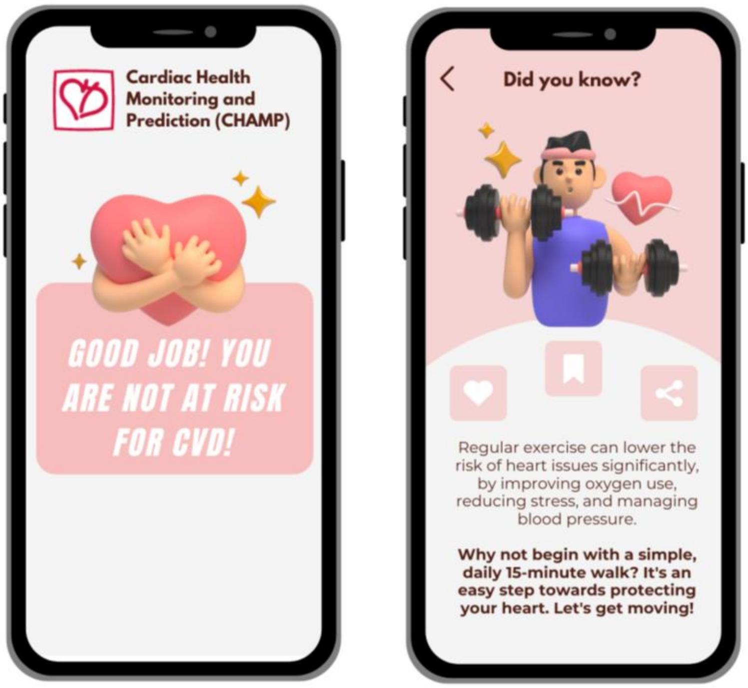
Name: Joey Chan Hsiao An
Sex: Female
Birthday: 28 January 1983
Age: 31
Education Level: GCE 'A' Levels

Figure 5.2.6: Quarterly Notifications to update the Risk Assessment Form

5.2.3 Preventive Features for Scenario 2 – Individual predicted to be not at risk for CVD

Feature 1: Educational Information about CVD

Users will receive educational information on CVD risk factors and preventive strategies (Figure 5.2.7). This will raise awareness and encourage individuals to be more proactive in managing their heart health, even if they are not currently at risk for CVD.



Cardiac Health Monitoring and Prediction (CHAMP)

GOOD JOB! YOU ARE NOT AT RISK FOR CVD!

Did you know?

Regular exercise can lower the risk of heart issues significantly, by improving oxygen use, reducing stress, and managing blood pressure.

Why not begin with a simple, daily 15-minute walk? It's an easy step towards protecting your heart. Let's get moving!

Figure 5.2.7: Scenario 2 – Predicted not at risk for CVD, Educational Information about Factors affecting CVD

<ins>Feature 2: Quarterly Reminders for Updated Risk Assessment</ins>

This feature will be extended to all individuals, regardless of whether they are identified as at risk. While they might not be at risk currently, they should continually monitor their CVD risk status.

5.3 Value Proposition

The CHAMP App interface is designed to be user-friendly for various age groups and intuitively structured for ease of use. The majority of details are retrieved from SingPass, which greatly simplifies the registration process and enhances accessibility. CHAMP will be freely available across various app stores, ensuring widespread availability for users without financial barriers.

The CHAMP App distinguishes itself from NHCS' traditional reactive methods like diagnosis and treatment by using predictive analytics to provide data-driven preventive measures. This effectively ensures that they visit a doctor after being predicted as at risk, allowing the diagnosis of CVD at an early stage. With early diagnosis, we can significantly reduce the need for time-consuming processes and the cost of potential advanced-stage treatment, meeting the aims of improving cost and operational efficiency for NHCS. Most importantly, this aligns with NHCS' goal of being a leader in heart care.

What sets our app apart is how CHAMP integrates lifestyle and behavioural factors in the analytics process to create a proactive risk management strategy. It raises awareness and empowers individuals to make more informed lifestyle decisions, reducing the work done by NHCS and minimising the demand for NHCS' resources.

5.3.1 Feasibility

In the short run, CHAMP will incur initial investment costs, such as software development and integration of the system with SingPass. However, it is freely available for users across all app stores, eliminating the financial barrier for users and ensuring ease of accessibility without imposing direct costs on individuals. By removing the financial barrier associated with regular health check-ups for preventive measures, CHAMP empowers individuals to proactively take preventive measures to safeguard their health without being burdened by financial constraints.

Early detection means that those at risk of CVD can promptly go for a checkup and, if necessary, receive treatment at the early stages of CVD, which contributes to a lower overall CVD mortality rate. In the long run, successful early detection of CVD by using the app can help to significantly reduce one's healthcare costs, minimise the burden on healthcare systems, and potentially save lives. These long-term benefits outweigh the short-term costs associated with building the app.

To measure CHAMP's success, we will track three key metrics. First, the CVD mortality rate can help us identify how early risk detection of cardiovascular disease can save lives. Second, by calculating the difference in the average cost of treatment for CHAMP users and non-users, we can assess the app's effectiveness in reducing users' healthcare expenses. Lastly, we will monitor the changes in average waiting time for appointments to evaluate improvements in healthcare accessibility facilitated by the app.

5.3.2 Limitations & Improvements

1. Inaccurate Information

The accuracy of the predictions heavily depends on the accuracy and completeness of the information provided by users. Without verification from medical professionals, there may be inaccuracies in the data, leading to less reliable predictions. To address this issue, NHCS can consider integrating the app with wearable devices capable of monitoring health metrics (e.g., heart rate, activity levels) to provide real-time data and enhance the accuracy of risk predictions. Furthermore, NHCS can schedule periodic health checkups for all users, regardless of their predicted risk level. Annual checkups can serve as opportunities to validate the information provided by users, identify any discrepancies or changes in health status, and offer personalised recommendations for maintaining or improving cardiovascular health.

2. Subjectivity of Self-Assessment

Some of the input factors rely on user-provided data and subjective assessments. One example would be rating general health from 1 to 5, with 1 being poor. This introduces variability as different individuals might have different perceptions and understanding of their health standards, lowering our predictions' accuracy. One possible solution would be to guide users by providing more structured prompts for classification. For example, providing users with specific prompts about their daily health – such as their frequency of falling sick, level of fatigue, and level of motivation for daily activities – can allow them to have a clearer perception of their general health status. This reduces subjectivity and uncertainty during classification, improving CHAMP's accuracy.

3. Limited Comprehensiveness of Risk Factors

This current dataset from the CDC lacks important factors, such as dietary habits, which significantly influence heart health. This limits CHAMP's ability to fully assess whether an individual is at risk for CVD. With access to local datasets containing a broader range of factors and with further utilisation of user data, our app's predictive ability and accuracy can be enhanced.

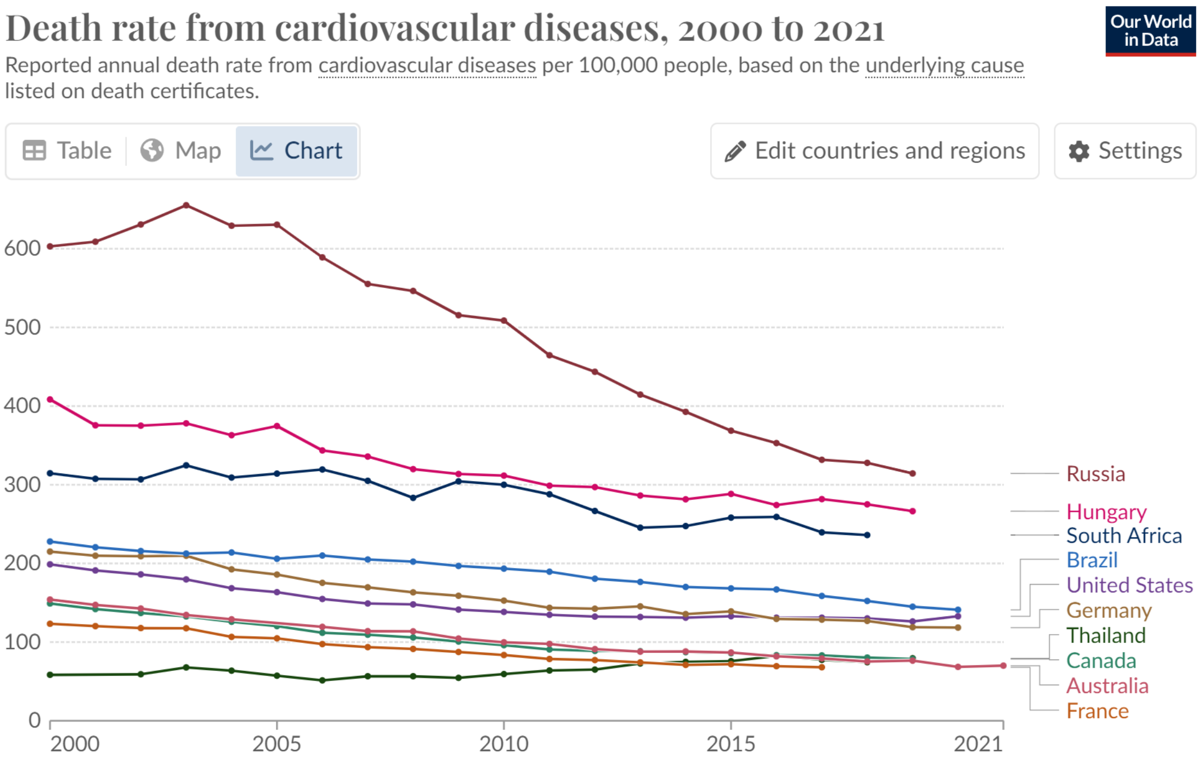
6. Conclusion

In conclusion, the report has effectively highlighted the considerable advantages attainable to the National Heart Centre Singapore (NHCS) through the strategic utilisation of analytics to enhance its clinical capabilities. Through the implementation of predictive models aimed at precisely evaluating an individual's risk of developing cardiovascular disease (CVD) in its early stages, NHCS stands poised to reinforce its status as a premier institution in cardiovascular care.

Furthermore, adopting such predictive analytics ensures the provision of optimal patient care and facilitates greater cost-efficiency and sustainability in the delivery of healthcare services. By embracing these innovative approaches, NHCS can position itself as a leader in cardiovascular care, demonstrating a commitment to excellence in patient outcomes and the prudent management of resources.

7. Appendix

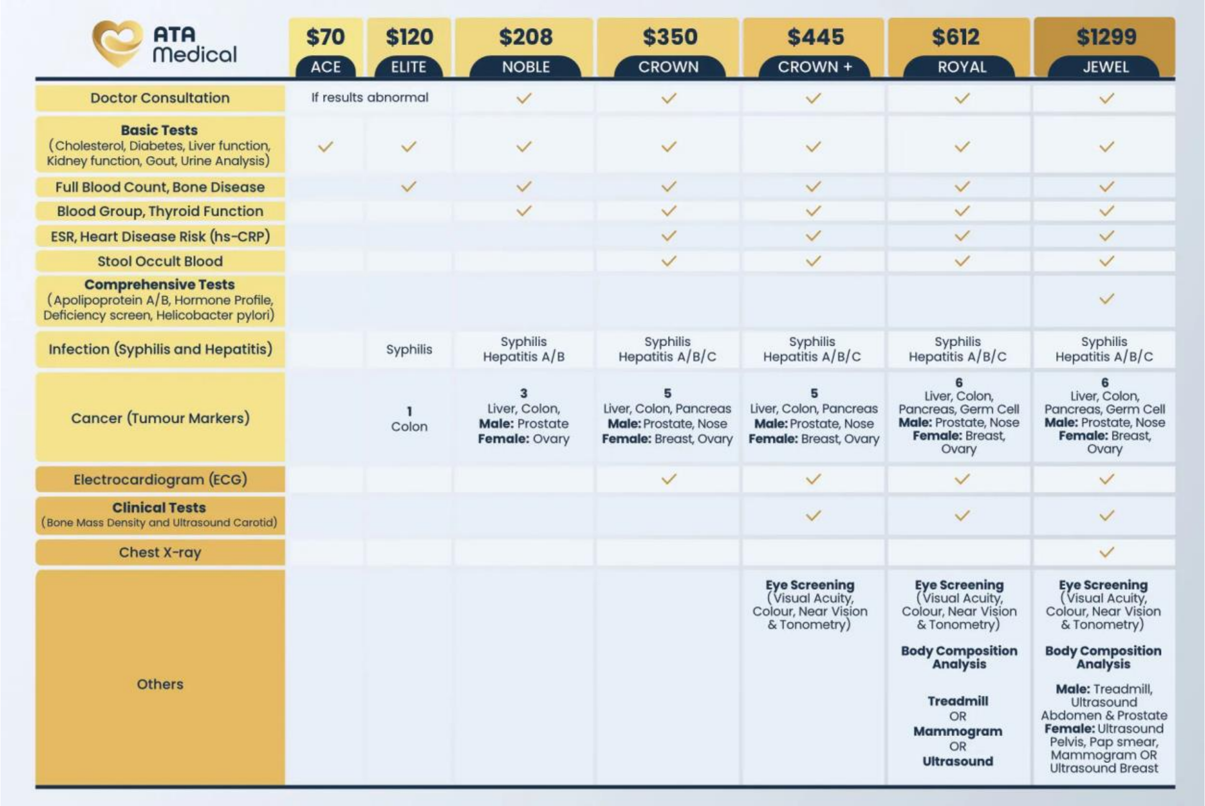
7.1 Appendix A: Business Understanding Research



Death rate from cardiovascular diseases, 2000 to 2021

Reported annual death rate from cardiovascular diseases per 100,000 people, based on the underlying cause listed on death certificates.

Figure 7.1.1: Costs of CVD Treatment



ATA Medical	$70 ACE	$120 ELITE	$208 NOBLE	$350 CROWN	$445 CROWN +	$612 ROYAL	$1299 JEWEL
Doctor Consultation	If results abnormal		✓	✓	✓	✓	✓
Basic Tests (Cholesterol, Diabetes, Liver function, Kidney function, Gout, Urine Analysis)	✓	✓	✓	✓	✓	✓	✓
Full Blood Count, Bone Disease		✓	✓	✓	✓	✓	✓
Blood Group, Thyroid Function			✓	✓	✓	✓	✓
ESR, Heart Disease Risk (hs-CRP)				✓	✓	✓	✓
Stool Occult Blood				✓	✓	✓	✓
Comprehensive Tests (Apolipoprotein A/B, Hormone Profile, Deficiency screen, Helicobacter pylori)							✓
Infection (Syphilis and Hepatitis)		Syphilis	Syphilis Hepatitis A/B	Syphilis Hepatitis A/B/C	Syphilis Hepatitis A/B/C	Syphilis Hepatitis A/B/C	Syphilis Hepatitis A/B/C
Cancer (Tumour Markers)		**1** Colon	**3** Liver, Colon, **Male:** Prostate **Female:** Ovary	**5** Liver, Colon, Pancreas **Male:** Prostate, Nose **Female:** Breast, Ovary	**5** Liver, Colon, Pancreas **Male:** Prostate, Nose **Female:** Breast, Ovary	**6** Liver, Colon, Pancreas, Germ Cell **Male:** Prostate, Nose **Female:** Breast, Ovary	**6** Liver, Colon, Pancreas, Germ Cell **Male:** Prostate, Nose **Female:** Breast, Ovary
Electrocardiogram (ECG)				✓	✓	✓	✓
Clinical Tests (Bone Mass Density and Ultrasound Carotid)					✓	✓	✓
Chest X-ray							✓
Others					**Eye Screening** (Visual Acuity, Colour, Near Vision & Tonometry)	**Eye Screening** (Visual Acuity, Colour, Near Vision & Tonometry) **Body Composition Analysis** **Treadmill** OR **Mammogram** OR **Ultrasound**	**Eye Screening** (Visual Acuity, Colour, Near Vision & Tonometry) **Body Composition Analysis** **Male:** Treadmill, Ultrasound Abdomen & Prostate **Female:** Ultrasound Pelvis, Pap smear, Mammogram OR Ultrasound Breast

Figure 7.1.2: Costs of CVD Treatment

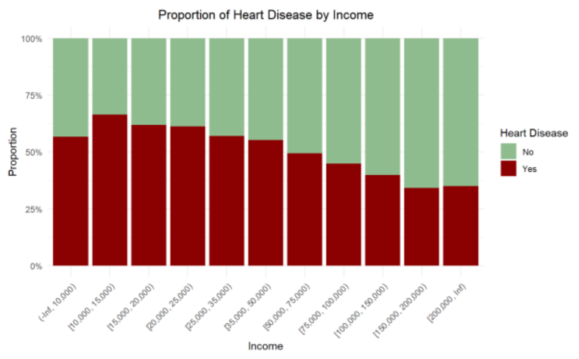
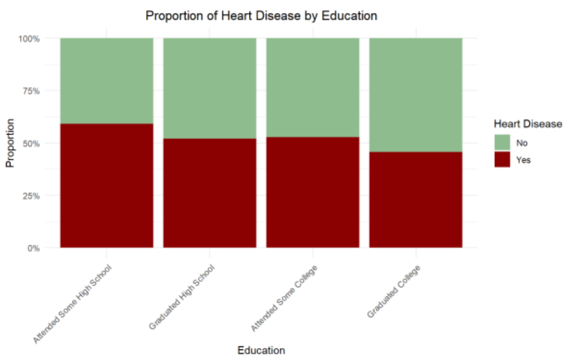
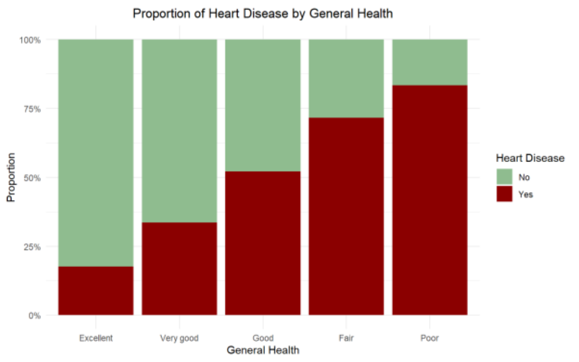
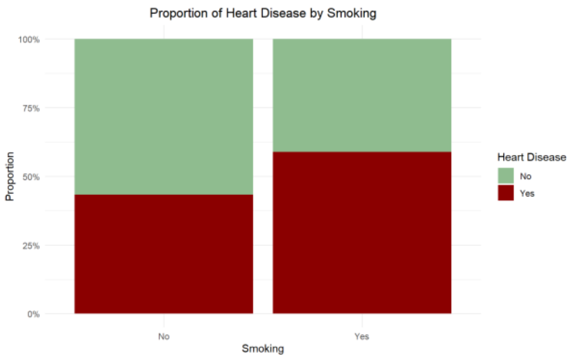
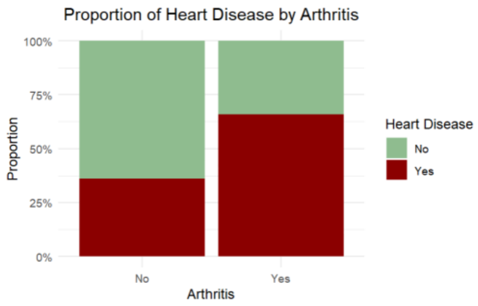
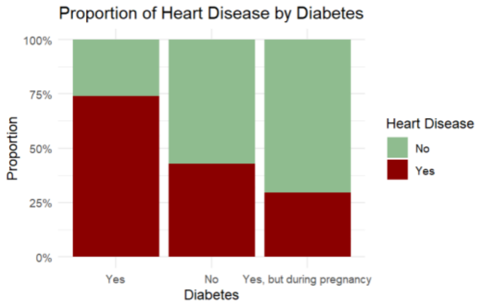
7.2 Appendix B: Data Dictionary

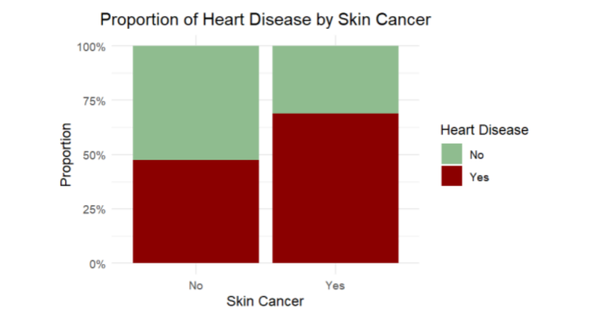
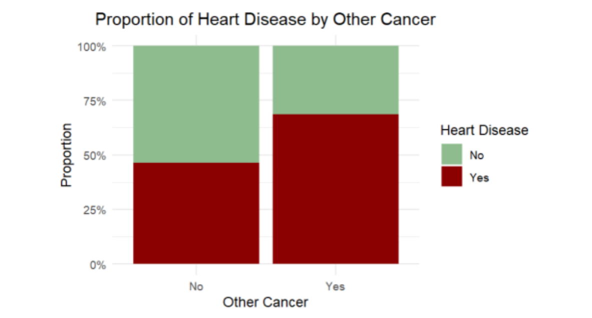
Variable Name	Data Type	Description
HEART_DISEASE	Categorical	Diagnosed with Coronary Heart Disease (CHD) [Yes/No]
SEX	Categorical	Sex of respondent
AGE_GROUP	Categorical	Age group of respondent
EDUCATION	Categorical	Level of education completed
INCOME	Categorical	Range of annual household income
UNABLE_TO_AFFORD_MED	Categorical	Inability to see a doctor in the past year due to cost [Yes/No]
WEIGHT_KG	Continuous	Weight of respondent
HEIGHT_M	Continuous	Height of respondent
BMI	Continuous	Calculated body mass index (BMI) of respondent
BMI_CATEGORY	Categorical	BMI category of respondent
GENERAL_HEALTH	Categorical	General health status from 1 to 5 (5 being excellent)
PHYSICAL_HEALTH	Continuous	Number of days in the past 30 days with poor physical health
MENTAL_HEALTH	Continuous	Number of days in the past 30 days with poor mental health
CHECKUP_DURATION_YR	Continuous	Time since the last routine doctor's checkup
EXERCISE	Categorical	Engagement in physical activity or exercise in the past 30 days outside of regular job [Yes/No]
SLEEP_DURATION_HR	Continuous	Average hours of sleep per 24-hour period
SMOKING	Categorical	Is currently a smoker [Yes/No]
CIGARETTES_PER_DAY	Continuous	Number of cigarettes smoked per day
DRINKING	Categorical	Consumed at least one drink of alcohol in the past 30 days
DRINKING_DAYS_PER_MONTH	Continuous	Number of days in the past 30 days where at least one drink of alcohol was consumed
HEAVY_DRINKING	Categorical	Heavy alcohol consumption (more than 14 drinks per week for men and 7 drinks per week for women) [Yes/No]
STROKE	Categorical	Ever diagnosed with a stroke [Yes/No]
ASTHMA	Categorical	Current diagnosis of asthma [Yes/No]
SKIN_CANCER	Categorical	Ever diagnosed with skin cancer [Yes/No]
OTHER_CANCER	Categorical	Ever diagnosed with other types of cancer [Yes/No]
LUNG_DISEASE	Categorical	Ever diagnosed with pulmonary-related disease(s) [Yes/No]
DEPRESSION	Categorical	Ever diagnosed with depressive disorders [Yes/No]
KIDNEY_DISEASE	Categorical	Ever diagnosed with kidney disease [Yes/No]
ARTHRITIS	Categorical	Ever diagnosed with arthritis [Yes/No]
OVERWEIGHT_OR_OBESE	Categorical	Has BMI greater than 25.00 (Overweight or obese) [Yes/No]

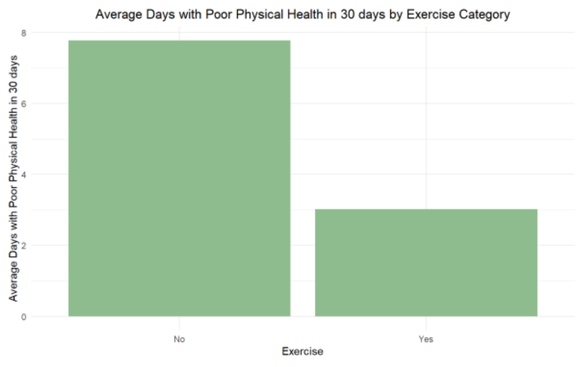
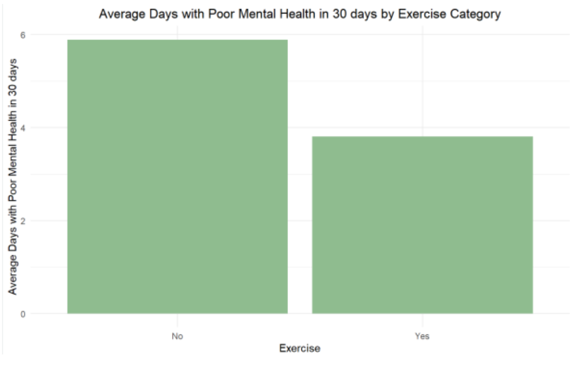
DIABETES	Categorical	Ever diagnosed with diabetes [Yes/No]

7.3 Appendix C: Exploratory Data Analysis

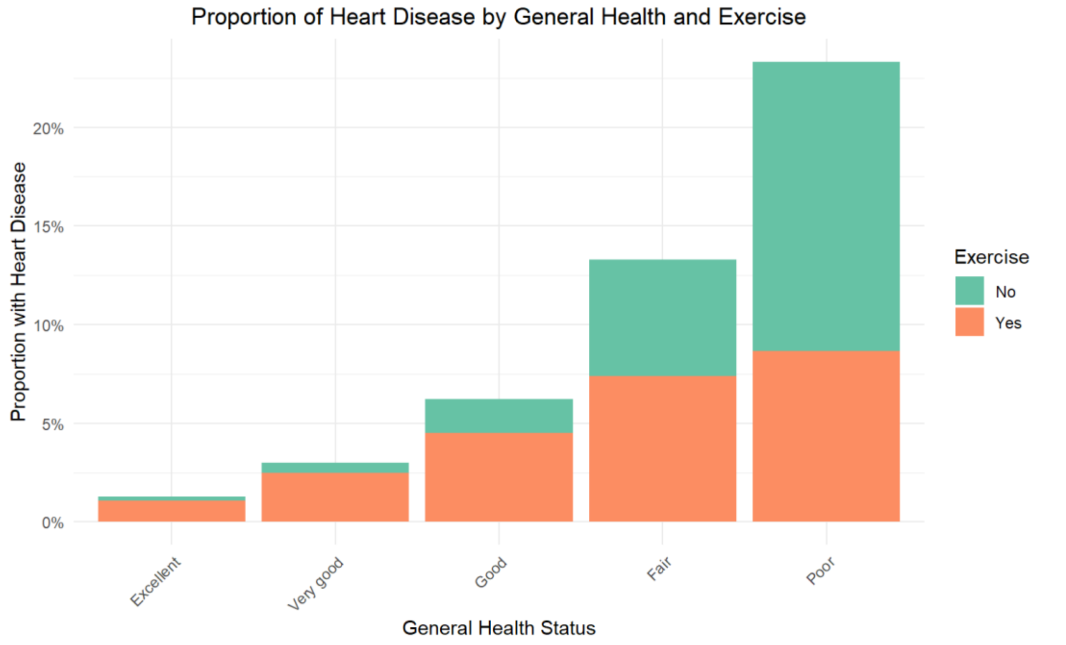
7.3.1 Bivariate EDA

<table>
<tr><th colspan="2">HEART DISEASE VS CATEGORICAL PREDICTOR VARIABLES</th></tr>
<tr><td><b>INCOME</b></td><td><b>EDUCATION</b></td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td><b>GENERAL HEALTH</b></td><td><b>SMOKING</b></td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td><b>ARTHRITIS</b></td><td><b>DIABETES</b></td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td><b>SKIN CANCER</b></td><td><b>OTHER CANCER</b></td></tr>
</table>

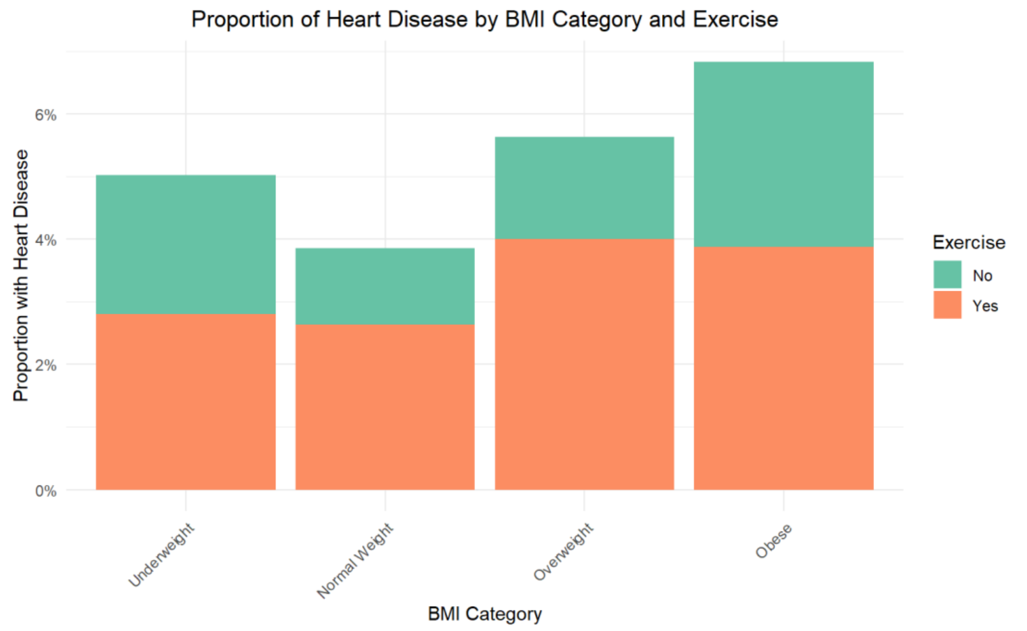
	

BIVARIATE ANALYSIS BETWEEN 2 PREDICTOR VARIABLES	
AVERAGE DAYS WITH POOR PHYSICAL HEALTH BY EXERCISE	AVERAGE DAYS WITH POOR MENTAL HEALTH BY EXERCISE
	

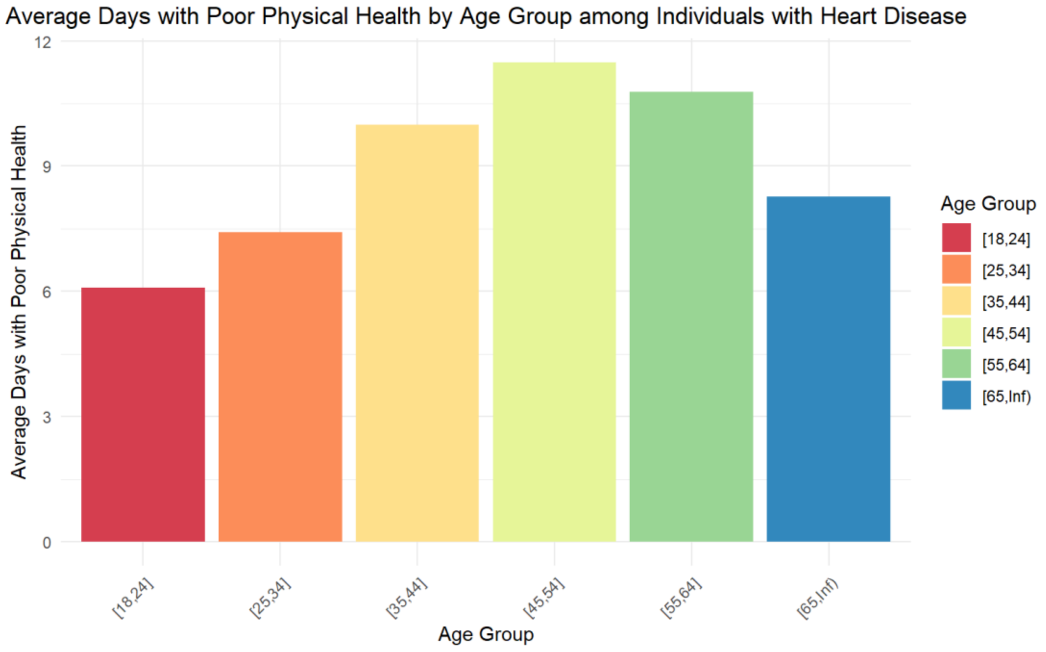
7.3.2 Multivariate EDA

HEART DISEASE BY GENERAL HEALTH AND EXERCISE


HEART DISEASE BY BMI CATEGORY AND EXERCISE

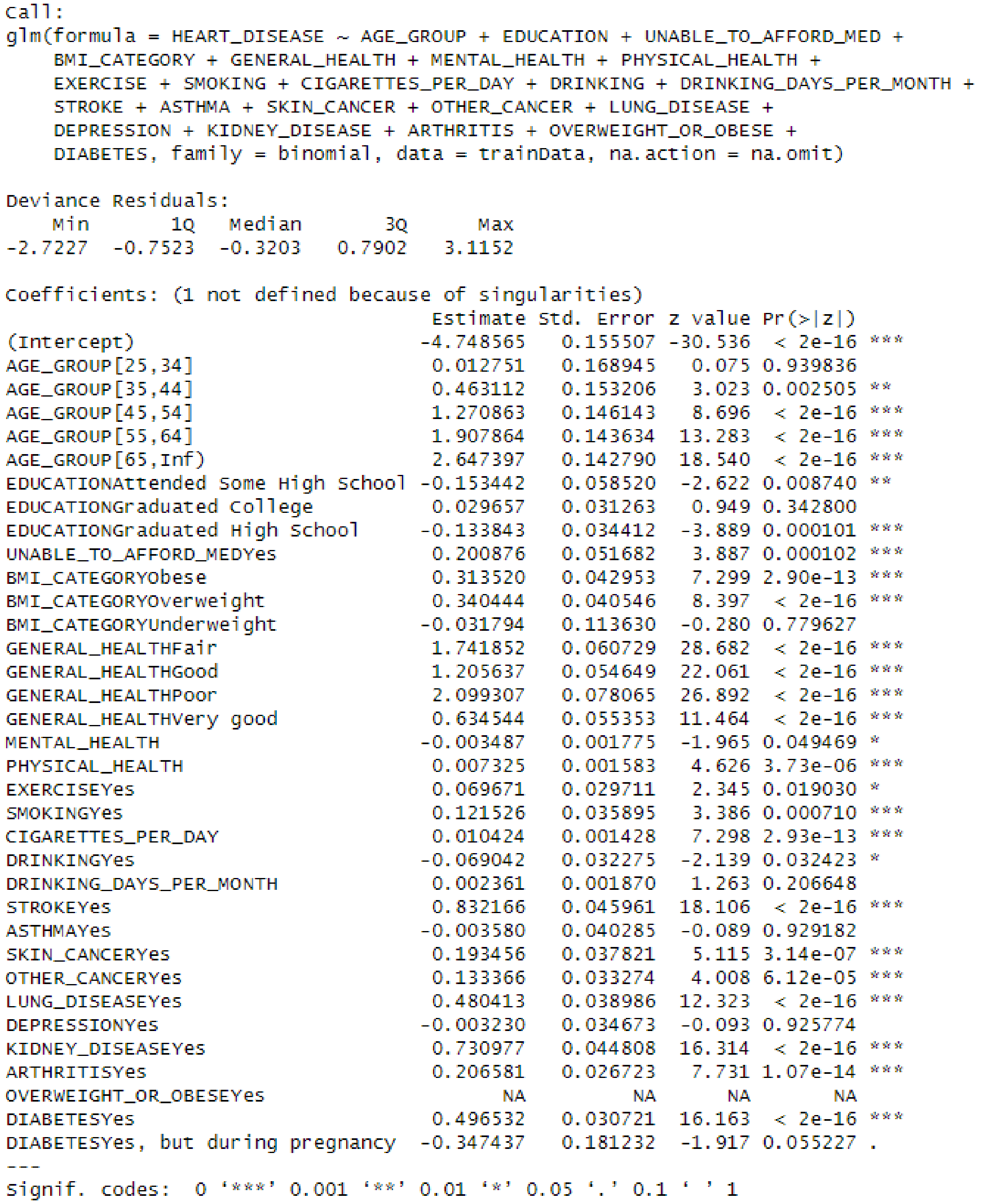


AVERAGE DAYS WITH POOR PHYSICAL HEALTH BY AGE GROUP AND HEART DISEASE



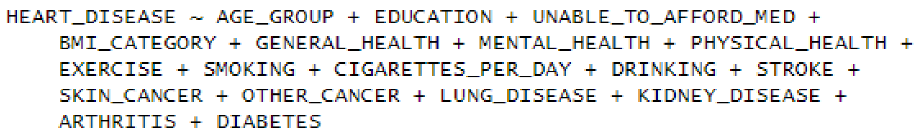
7.4 Appendix D: Modelling

7.4.1 Logistic Regression

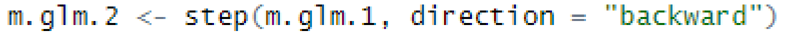


```
Call:
glm(formula = HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
    BMI_CATEGORY + GENERAL_HEALTH + MENTAL_HEALTH + PHYSICAL_HEALTH +
    EXERCISE + SMOKING + CIGARETTES_PER_DAY + DRINKING + DRINKING_DAYS_PER_MONTH +
    STROKE + ASTHMA + SKIN_CANCER + OTHER_CANCER + LUNG_DISEASE +
    DEPRESSION + KIDNEY_DISEASE + ARTHRITIS + OVERWEIGHT_OR_OBESE +
    DIABETES, family = binomial, data = trainData, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7227  -0.7523  -0.3203   0.7902   3.1152

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -4.748565   0.155507 -30.536  < 2e-16 ***
AGE_GROUP[25,34]                    0.012751   0.168945   0.075 0.939836
AGE_GROUP[35,44]                    0.463112   0.153206   3.023 0.002505 **
AGE_GROUP[45,54]                    1.270863   0.146143   8.696  < 2e-16 ***
AGE_GROUP[55,64]                    1.907864   0.143634  13.283  < 2e-16 ***
AGE_GROUP[65,Inf)                   2.647397   0.142790  18.540  < 2e-16 ***
EDUCATIONAttended Some High School -0.153442   0.058520  -2.622 0.008740 **
EDUCATIONGraduated College          0.029657   0.031263   0.949 0.342800
EDUCATIONGraduated High School     -0.133843   0.034412  -3.889 0.000101 ***
UNABLE_TO_AFFORD_MEDYes             0.200876   0.051682   3.887 0.000102 ***
BMI_CATEGORYObese                   0.313520   0.042953   7.299 2.90e-13 ***
BMI_CATEGORYOverweight              0.340444   0.040546   8.397  < 2e-16 ***
BMI_CATEGORYUnderweight            -0.031794   0.113630  -0.280 0.779627
GENERAL_HEALTHFair                  1.741852   0.060729  28.682  < 2e-16 ***
GENERAL_HEALTHGood                  1.205637   0.054649  22.061  < 2e-16 ***
GENERAL_HEALTHPoor                  2.099307   0.078065  26.892  < 2e-16 ***
GENERAL_HEALTHVery good             0.634544   0.055353  11.464  < 2e-16 ***
MENTAL_HEALTH                      -0.003487   0.001775  -1.965 0.049469 *
PHYSICAL_HEALTH                     0.007325   0.001583   4.626 3.73e-06 ***
EXERCISEYes                         0.069671   0.029711   2.345 0.019030 *
SMOKINGYes                          0.121526   0.035895   3.386 0.000710 ***
CIGARETTES_PER_DAY                  0.010424   0.001428   7.298 2.93e-13 ***
DRINKINGYes                        -0.069042   0.032275  -2.139 0.032423 *
DRINKING_DAYS_PER_MONTH             0.002361   0.001870   1.263 0.206648
STROKEYes                           0.832166   0.045961  18.106  < 2e-16 ***
ASTHMAYes                          -0.003580   0.040285  -0.089 0.929182
SKIN_CANCERYes                      0.193456   0.037821   5.115 3.14e-07 ***
OTHER_CANCERYes                     0.133366   0.033274   4.008 6.12e-05 ***
LUNG_DISEASEYes                     0.480413   0.038986  12.323  < 2e-16 ***
DEPRESSIONYes                      -0.003230   0.034673  -0.093 0.925774
KIDNEY_DISEASEYes                   0.730977   0.044808  16.314  < 2e-16 ***
ARTHRITISYes                        0.206581   0.026723   7.731 1.07e-14 ***
OVERWEIGHT_OR_OBESEYes                    NA         NA      NA       NA
DIABETESYes                         0.496532   0.030721  16.163  < 2e-16 ***
DIABETESYes, but during pregnancy  -0.347437   0.181232  -1.917 0.055227 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7.4.1.1: Logistic Regression based on EDA Insights

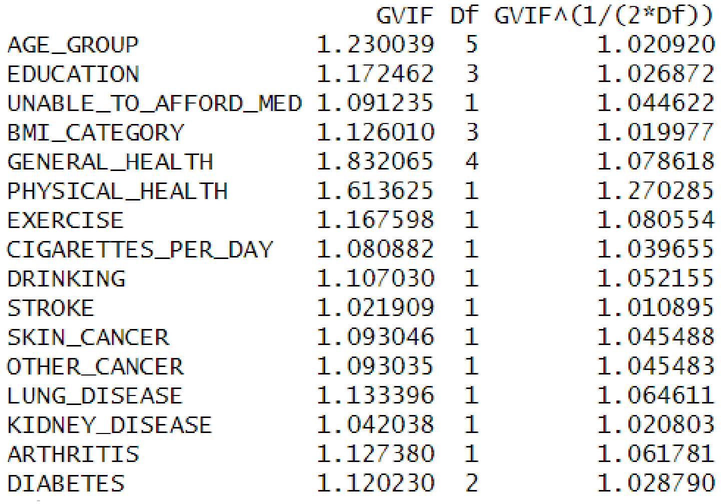


```
HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
    BMI_CATEGORY + GENERAL_HEALTH + MENTAL_HEALTH + PHYSICAL_HEALTH +
    EXERCISE + SMOKING + CIGARETTES_PER_DAY + DRINKING + STROKE +
    SKIN_CANCER + OTHER_CANCER + LUNG_DISEASE + KIDNEY_DISEASE +
    ARTHRITIS + DIABETES
```



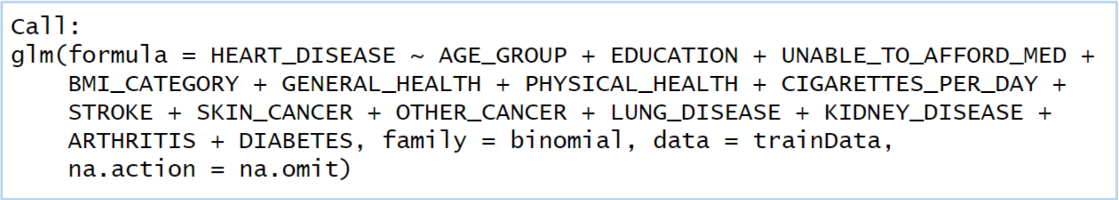
```
m.glm.2 <- step(m.glm.1, direction = "backward")
```

Figure 7.4.1.2: Backward Stepwise Selection



```
                         GVIF Df GVIF^(1/(2*Df))
AGE_GROUP            1.230039  5        1.020920
EDUCATION            1.172462  3        1.026872
UNABLE_TO_AFFORD_MED 1.091235  1        1.044622
BMI_CATEGORY         1.126010  3        1.019977
GENERAL_HEALTH       1.832065  4        1.078618
PHYSICAL_HEALTH      1.613625  1        1.270285
EXERCISE             1.167598  1        1.080554
CIGARETTES_PER_DAY   1.080882  1        1.039655
DRINKING             1.107030  1        1.052155
STROKE               1.021909  1        1.010895
SKIN_CANCER          1.093046  1        1.045488
OTHER_CANCER         1.093035  1        1.045483
LUNG_DISEASE         1.133396  1        1.064611
KIDNEY_DISEASE       1.042038  1        1.020803
ARTHRITIS            1.127380  1        1.061781
DIABETES             1.120230  2        1.028790
```

Figure 7.4.1.3: Checking VIF values for multicollinearity



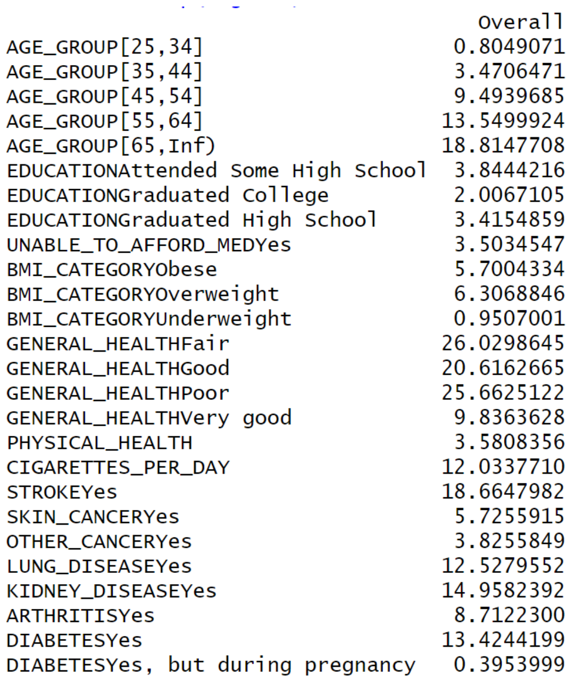
```
Call:
glm(formula = HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
    BMI_CATEGORY + GENERAL_HEALTH + PHYSICAL_HEALTH + CIGARETTES_PER_DAY +
    STROKE + SKIN_CANCER + OTHER_CANCER + LUNG_DISEASE + KIDNEY_DISEASE +
    ARTHRITIS + DIABETES, family = binomial, data = trainData,
    na.action = na.omit)
```

Figure 7.4.1.4: Formula for Logistic Regression Model

Variable	**Coefficient**	**Odds Ratio**	**OR.CI (2.5%)**	**OR.CI (97.5%)**
AGE_GROUP [25,34]	0.135012	1.145	0.829	1.602
AGE_GROUP [35,44]	0.535234	1.708	1.273	2.333
AGE_GROUP [45,54]	1.392668	4.026	3.049	5.425
AGE_GROUP [55,64]	1.953453	7.053	5.371	9.462
AGE_GROUP [65,Inf)	2.689822	14.729	11.246	19.717
EDUCATION [Attended Some High School]	-0.230690	0.794	0.706	0.893
EDUCATION [Graduated College]	0.063853	1.066	1.002	1.135
EDUCATION [Graduated High School]	-0.121367	0.886	0.826	0.950

UNABLE_TO_AFFORD_MED [Yes]	0.184060	1.202	1.084	1.332
BMI_CATEGORY [Obese]	0.251292	1.286	1.179	1.402
BMI_CATEGORY [Overweight]	0.263013	1.301	1.199	1.412
BMI_CATEGORY [Underweight]	0.109229	1.115	0.890	1.396
GENERAL_HEALTH [Fair]	1.598669	4.946	4.388	5.583
GENERAL_HEALTH [Good]	1.141090	3.130	2.811	3.492
GENERAL_HEALTH [Poor]	2.028180	7.600	6.513	8.879
GENERAL_HEALTH [Very good]	0.550779	1.735	1.556	1.937
PHYSICAL_HEALTH	0.005637	1.006	1.003	1.009
CIGARETTES_PER_DAY	0.018108	1.018	1.015	1.021
STROKE [Yes]	0.897232	2.453	2.233	2.696
SKIN_CANCER [Yes]	0.222356	1.249	1.157	1.348
OTHER_CANCER [Yes]	0.131806	1.141	1.066	1.221
LUNG_DISEASE [Yes]	0.493816	1.639	1.517	1.770
KIDNEY_DISEASE [Yes]	0.689367	1.992	1.821	2.181
ARTHRITIS [Yes]	0.237594	1.268	1.202	1.338
DIABETES [Yes]	0.424107	1.528	1.436	1.626
DIABETES [Yes, but during pregnancy]	-0.069846	0.933	0.654	1.308

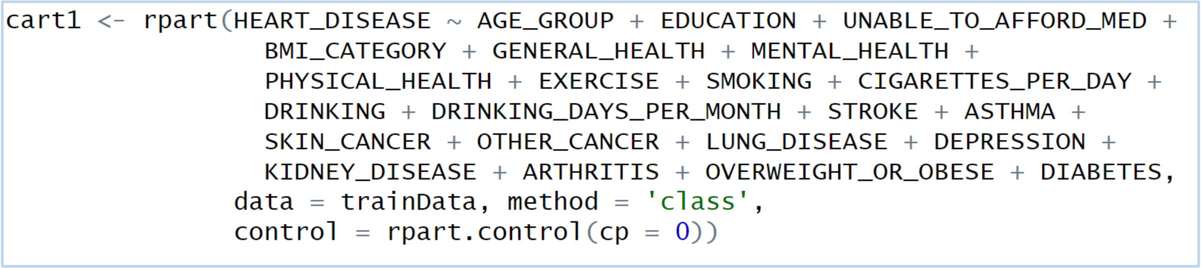
Figure 7.4.1.5: Summary of Logistic Regression Model



```
                                      Overall
AGE_GROUP[25,34]                    0.8049071
AGE_GROUP[35,44]                    3.4706471
AGE_GROUP[45,54]                    9.4939685
AGE_GROUP[55,64]                   13.5499924
AGE_GROUP[65,Inf)                  18.8147708
EDUCATIONAttended Some High School  3.8444216
EDUCATIONGraduated College          2.0067105
EDUCATIONGraduated High School      3.4154859
UNABLE_TO_AFFORD_MEDYes             3.5034547
BMI_CATEGORYObese                   5.7004334
BMI_CATEGORYOverweight              6.3068846
BMI_CATEGORYUnderweight             0.9507001
GENERAL_HEALTHFair                 26.0298645
GENERAL_HEALTHGood                 20.6162665
GENERAL_HEALTHPoor                 25.6625122
GENERAL_HEALTHVery good             9.8363628
PHYSICAL_HEALTH                     3.5808356
CIGARETTES_PER_DAY                 12.0337710
STROKEYes                          18.6647982
SKIN_CANCERYes                      5.7255915
OTHER_CANCERYes                     3.8255849
LUNG_DISEASEYes                    12.5279552
KIDNEY_DISEASEYes                  14.9582392
ARTHRITISYes                        8.7122300
DIABETESYes                        13.4244199
DIABETESYes, but during pregnancy   0.3953999
```

Figure 7.4.1.6: Variable Importance for Logistic Regression Model

7.4.2 CART – Classification Tree



```
cart1 <- rpart(HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
                    BMI_CATEGORY + GENERAL_HEALTH + MENTAL_HEALTH +
                    PHYSICAL_HEALTH + EXERCISE + SMOKING + CIGARETTES_PER_DAY +
                    DRINKING + DRINKING_DAYS_PER_MONTH + STROKE + ASTHMA +
                    SKIN_CANCER + OTHER_CANCER + LUNG_DISEASE + DEPRESSION +
                    KIDNEY_DISEASE + ARTHRITIS + OVERWEIGHT_OR_OBESE + DIABETES,
                  data = trainData, method = 'class',
                  control = rpart.control(cp = 0))
```

Figure 7.4.2.1: Formula for CART

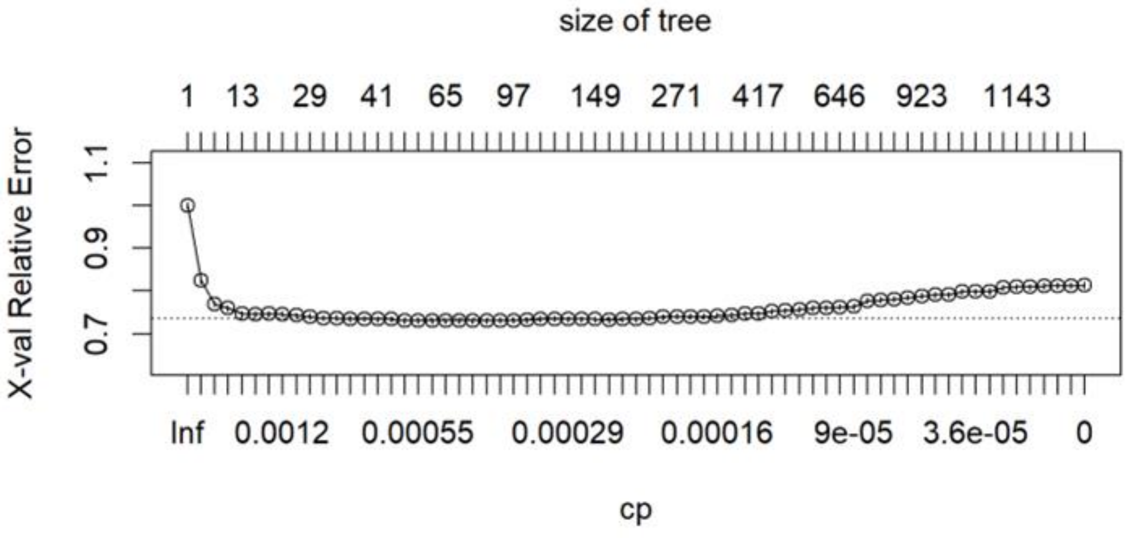


Figure 7.4.2.2: Cross-validation errors for different complexity parameters

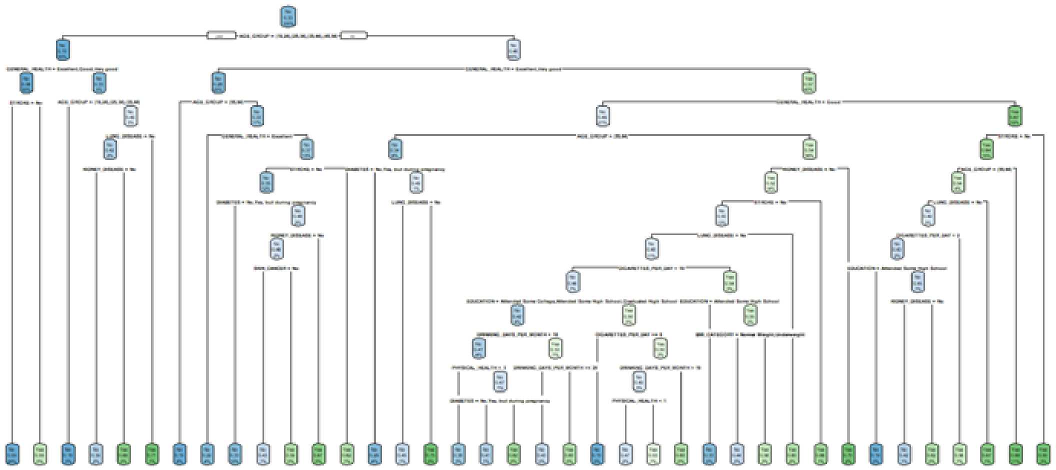


Figure 7.4.2.3 CART model (Optimal prune)

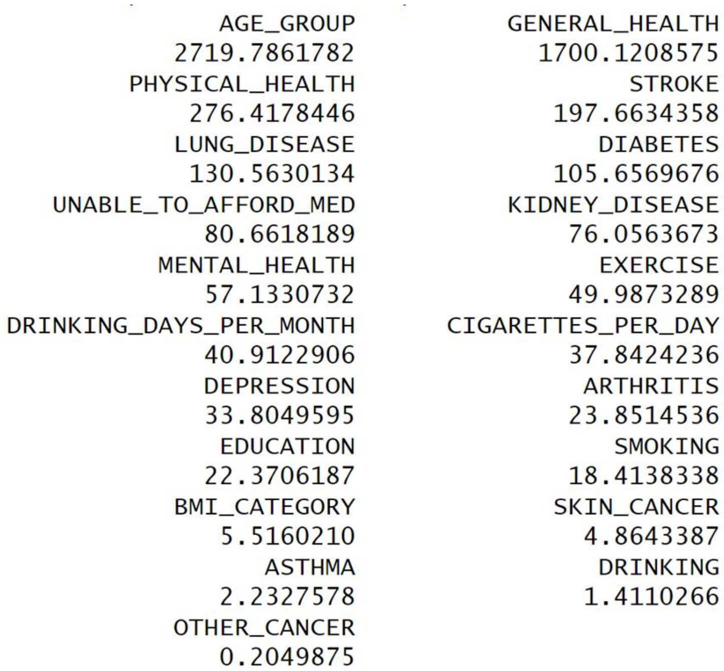
Full summary of the CART Model is shown due to the small font size above

```
n= 39204

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 39204 13068 No (0.66666667 0.33333333)
     2) AGE_GROUP=[18,24],[25,34],[35,44],[45,54] 13811  1401 No (0.89855912
0.10144088)
       4) GENERAL_HEALTH=Excellent,Good,Very good 11677   699 No (0.94013873
0.05986127)
         8) STROKE=No 11518   612 No (0.94686578 0.05313422) *
         9) STROKE=Yes 159    72 Yes (0.45283019 0.54716981) *
       5) GENERAL_HEALTH=Fair,Poor 2134   702 No (0.67104030 0.32895970)
        10) AGE_GROUP=[18,24],[25,34],[35,44] 1091   193 No (0.82309808 0.17690192) *
        11) AGE_GROUP=[45,54] 1043   509 No (0.51198466 0.48801534)
          22) LUNG_DISEASE=No 808   343 No (0.57549505 0.42450495)
            44) KIDNEY_DISEASE=No 712   278 No (0.60955056 0.39044944) *
            45) KIDNEY_DISEASE=Yes 96    31 Yes (0.32291667 0.67708333) *
          23) LUNG_DISEASE=Yes 235    69 Yes (0.29361702 0.70638298) *
     3) AGE_GROUP=[55,64],[65,Inf) 25393 11667 No (0.54054267 0.45945733)
       6) GENERAL_HEALTH=Excellent,Very good 9680  2663 No (0.72489669
0.27510331)
        12) AGE_GROUP=[55,64] 3051   447 No (0.85349066 0.14650934) *
        13) AGE_GROUP=[65,Inf) 6629  2216 No (0.66571127 0.33428873)
          26) GENERAL_HEALTH=Excellent 1541   336 No (0.78195977 0.21804023) *
          27) GENERAL_HEALTH=Very good 5088  1880 No (0.63050314 0.36949686)
            54) STROKE=No 4803  1703 No (0.64542994 0.35457006)
             108) DIABETES=No,Yes, but during pregnancy 4082  1355 No (0.66805488
0.33194512) *
             109) DIABETES=Yes 721   348 No (0.51733703 0.48266297)
```

```
              218) KIDNEY_DISEASE=No 639   293 No (0.54147105 0.45852895)
                436) SKIN_CANCER=No 530   229 No (0.56792453 0.43207547) *
                437) SKIN_CANCER=Yes 109    45 Yes (0.41284404 0.58715596) *
              219) KIDNEY_DISEASE=Yes 82    27 Yes (0.32926829 0.67073171) *
          55) STROKE=Yes 285   108 Yes (0.37894737 0.62105263) *
    7) GENERAL_HEALTH=Fair,Good,Poor 15713  6709 Yes (0.42697130 0.57302870)
      14) GENERAL_HEALTH=Good 8353  4052 No (0.51490482 0.48509518)
        28) AGE_GROUP=[55,64] 2181   737 No (0.66208161 0.33791839)
          56) DIABETES=No,Yes, but during pregnancy 1632   469 No (0.71262255
0.28737745) *
          57) DIABETES=Yes 549   268 No (0.51183971 0.48816029)
            114) LUNG_DISEASE=No 488   222 No (0.54508197 0.45491803) *
            115) LUNG_DISEASE=Yes 61    15 Yes (0.24590164 0.75409836) *
        29) AGE_GROUP=[65,Inf) 6172  2857 Yes (0.46289695 0.53710305)
          58) KIDNEY_DISEASE=No 5453  2644 Yes (0.48487071 0.51512929)
            116) STROKE=No 4917  2454 No (0.50091519 0.49908481)
              232) LUNG_DISEASE=No 4274  2066 No (0.51661207 0.48338793)
                464) CIGARETTES_PER_DAY< 9.5 2933  1345 No (0.54142516 0.45857484)
                  928) EDUCATION=Attended Some College,Attended Some High
School,Graduated High School 1577   666 No (0.57767914 0.42232086)
                    1856) DRINKING_DAYS_PER_MONTH< 9.5 1379   562 No (0.59245830
0.40754170)
                      3712) PHYSICAL_HEALTH< 2.5 1004   386 No (0.61553785 0.38446215)
*
                      3713) PHYSICAL_HEALTH>=2.5 375   176 No (0.53066667 0.46933333)
                        7426) DIABETES=No,Yes, but during pregnancy 274   113 No
(0.58759124 0.41240876) *
                        7427) DIABETES=Yes 101    38 Yes (0.37623762 0.62376238) *
                    1857) DRINKING_DAYS_PER_MONTH>=9.5 198    94 Yes (0.47474747
0.52525253)
                      3714) DRINKING_DAYS_PER_MONTH>=24.5 86    37 No (0.56976744
0.43023256) *
                      3715) DRINKING_DAYS_PER_MONTH< 24.5 112    45 Yes (0.40178571
0.59821429) *
                  929) EDUCATION=Graduated College 1356   677 Yes (0.49926254
0.50073746)
                    1858) CIGARETTES_PER_DAY>=7.5 10     1 No (0.90000000 0.10000000)
*
                    1859) CIGARETTES_PER_DAY< 7.5 1346   668 Yes (0.49628529
0.50371471)
                      3718) DRINKING_DAYS_PER_MONTH< 18.5 1191   585 No (0.50881612
0.49118388)
                        7436) PHYSICAL_HEALTH< 0.5 708   330 No (0.53389831 0.46610169)
*
                        7437) PHYSICAL_HEALTH>=0.5 483   228 Yes (0.47204969
0.52795031) *
                      3719) DRINKING_DAYS_PER_MONTH>=18.5 155    62 Yes (0.40000000
0.60000000) *
                465) CIGARETTES_PER_DAY>=9.5 1341   620 Yes (0.46234154
```

```
0.53765846)
                   930) EDUCATION=Attended Some High School 48    16 No (0.66666667
0.33333333) *
                   931) EDUCATION=Attended Some College,Graduated College,Graduated
High School 1293   588 Yes (0.45475638 0.54524362)
                    1862) BMI_CATEGORY=Normal Weight,Underweight 154    67 No
(0.56493506 0.43506494) *
                    1863) BMI_CATEGORY=Obese,Overweight 1139   501 Yes (0.43985953
0.56014047) *
               233) LUNG_DISEASE=Yes 643   255 Yes (0.39657854 0.60342146) *
             117) STROKE=Yes 536   181 Yes (0.33768657 0.66231343) *
            59) KIDNEY_DISEASE=Yes 719   213 Yes (0.29624478 0.70375522) *
        15) GENERAL_HEALTH=Fair,Poor 7360  2408 Yes (0.32717391 0.67282609)
          30) STROKE=No 6010  2137 Yes (0.35557404 0.64442596)
            60) AGE_GROUP=[55,64] 1711   780 Yes (0.45587376 0.54412624)
             120) LUNG_DISEASE=No 1187   580 No (0.51137321 0.48862679)
               240) CIGARETTES_PER_DAY< 1.5 616   262 No (0.57467532 0.42532468)
                 480) EDUCATION=Attended Some High School 64    12 No (0.81250000
0.18750000) *
                 481) EDUCATION=Attended Some College,Graduated College,Graduated High
School 552   250 No (0.54710145 0.45289855)
                   962) KIDNEY_DISEASE=No 457   191 No (0.58205689 0.41794311) *
                   963) KIDNEY_DISEASE=Yes 95    36 Yes (0.37894737 0.62105263) *
               241) CIGARETTES_PER_DAY>=1.5 571   253 Yes (0.44308231 0.55691769) *
             121) LUNG_DISEASE=Yes 524   173 Yes (0.33015267 0.66984733) *
            61) AGE_GROUP=[65,Inf) 4299  1357 Yes (0.31565480 0.68434520) *
          31) STROKE=Yes 1350   271 Yes (0.20074074 0.79925926) *
```

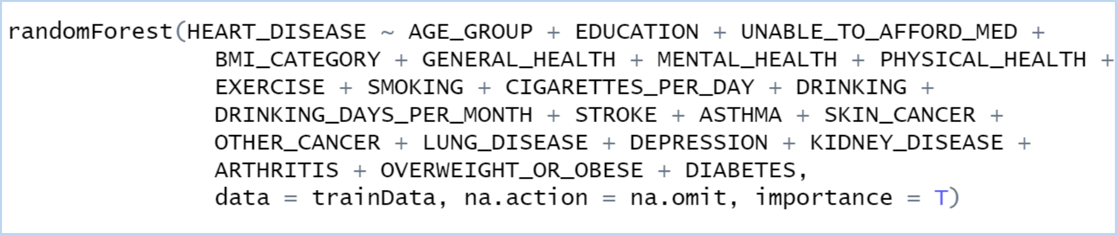
Figure 7.4.2.4: Full Summary of CART Model



```
             AGE_GROUP          GENERAL_HEALTH
          2719.7861782            1700.1208575
       PHYSICAL_HEALTH                  STROKE
           276.4178446             197.6634358
          LUNG_DISEASE                DIABETES
           130.5630134             105.6569676
   UNABLE_TO_AFFORD_MED          KIDNEY_DISEASE
            80.6618189              76.0563673
         MENTAL_HEALTH                EXERCISE
            57.1330732              49.9873289
 DRINKING_DAYS_PER_MONTH      CIGARETTES_PER_DAY
            40.9122906              37.8424236
            DEPRESSION               ARTHRITIS
            33.8049595              23.8514536
             EDUCATION                 SMOKING
            22.3706187              18.4138338
          BMI_CATEGORY             SKIN_CANCER
             5.5160210               4.8643387
                ASTHMA                DRINKING
             2.2327578               1.4110266
          OTHER_CANCER
             0.2049875
```

Figure 7.4.2.5 Variable Importance for CART

7.4.3 Random Forest



```
randomForest(HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
               BMI_CATEGORY + GENERAL_HEALTH + MENTAL_HEALTH + PHYSICAL_HEALTH +
               EXERCISE + SMOKING + CIGARETTES_PER_DAY + DRINKING +
               DRINKING_DAYS_PER_MONTH + STROKE + ASTHMA + SKIN_CANCER +
               OTHER_CANCER + LUNG_DISEASE + DEPRESSION + KIDNEY_DISEASE +
               ARTHRITIS + OVERWEIGHT_OR_OBESE + DIABETES,
               data = trainData, na.action = na.omit, importance = T)
```

Figure 7.4.3.1: Formula for Random Forest

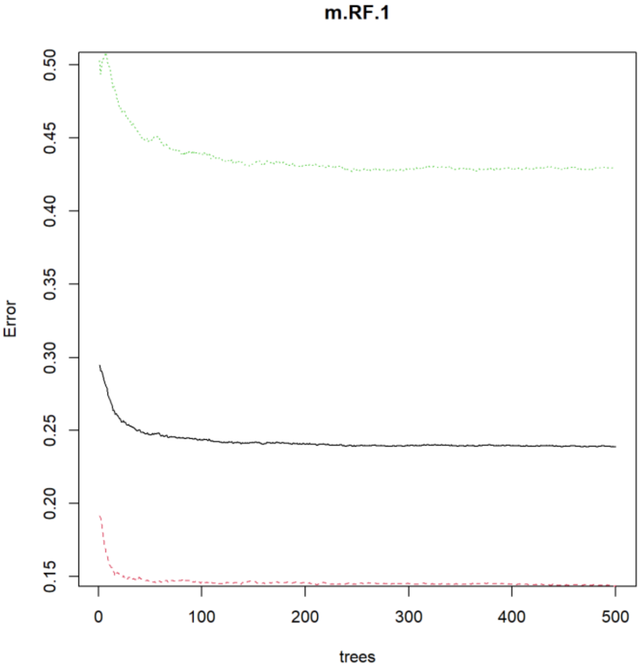


Figure 7.4.3.2: OOB Error against Number of Trees for Random Forest

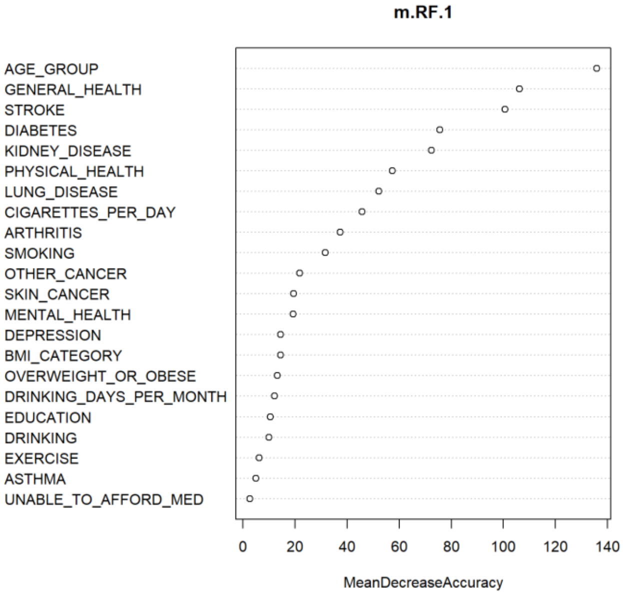
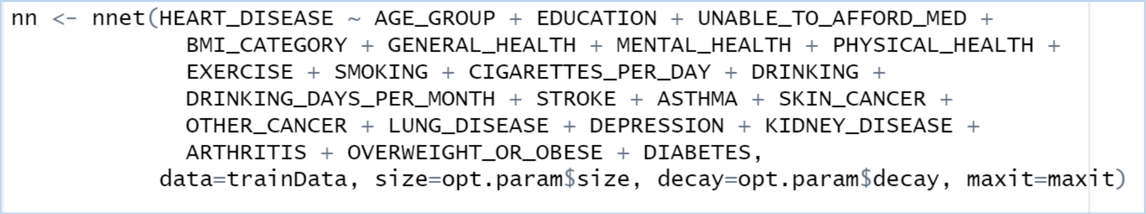


Figure 7.4.3.3: Variable Importance for Random Forest

7.4.4 Neural Network



```
nn <- nnet(HEART_DISEASE ~ AGE_GROUP + EDUCATION + UNABLE_TO_AFFORD_MED +
             BMI_CATEGORY + GENERAL_HEALTH + MENTAL_HEALTH + PHYSICAL_HEALTH +
             EXERCISE + SMOKING + CIGARETTES_PER_DAY + DRINKING +
             DRINKING_DAYS_PER_MONTH + STROKE + ASTHMA + SKIN_CANCER +
             OTHER_CANCER + LUNG_DISEASE + DEPRESSION + KIDNEY_DISEASE +
             ARTHRITIS + OVERWEIGHT_OR_OBESE + DIABETES,
           data=trainData, size=opt.param$size, decay=opt.param$decay, maxit=maxit)
```

Figure 7.4.4.1: Formula for Neural Network

7.5 Appendix E: NHCS’ Current Measures

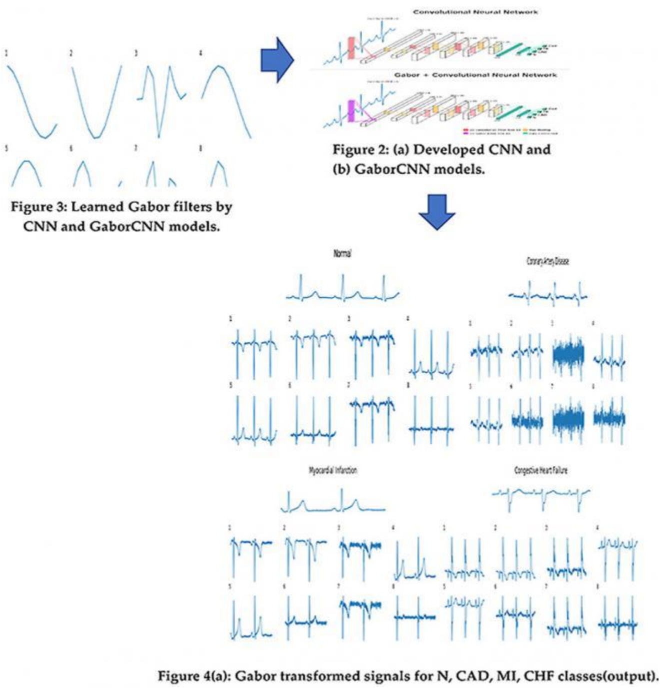


Figure 2: (a) Developed CNN and (b) GaborCNN models.

Figure 3: Learned Gabor filters by CNN and GaborCNN models.

Figure 4(a): Gabor transformed signals for N, CAD, MI, CHF classes(output).

Figure 7.5.1: Gabor-CNN Algorithm for ECG pattern recognition and CVD diagnosis

8. References

Avramova, N. (2018, December 5). *Too much sleep linked to a greater risk of disease and death, study finds*. Cable News Network. https://edition.cnn.com/2018/12/04/health/sleep-duration-linked-to-death-and-diseases-study-intl/index.html

Brouwer, E. D., Watkins, D., Olson, Z., Goett, J., Nugent, R., & Levin, C. (2015, November 26). *Provider costs for prevention and treatment of cardiovascular and related conditions in low- and middle-income countries: a systematic review*. BMC Public Health. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-2538-z

Centers for Disease Control and Prevention. (2020, May 6). *Heart Disease and Mental Health Disorders*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/mentalhealth.htm

Centers for Disease Control and Prevention. (2022, September 8). *Heart Disease and Stroke*. Centers for Disease Control and Prevention. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

Centers for Disease Control and Prevention. (2022b, September 8). Physical Inactivity. Centers for Disease Control and Prevention. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/physical-activity.htm

Centers for Disease Control and Prevention. (2023, November 15). *CDC - 2022 BRFSS Survey Data and Documentation*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2022.html

Centers for Disease Control and Prevention. (2024, February 5). *Smoking and Heart Disease, Stroke, and Peripheral Artery Disease*. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/campaign/tips/diseases/heart-disease-stroke.html

Ciumărnean, L., Milaciu, M. V., Negrean, V., Orășan, O. H., Vesa, S. C., Sălăgean, O., Iluţ, S., & Vlaicu, S. I. (2021, December 25). *Cardiovascular Risk Factors and Physical Activity for the Prevention of Cardiovascular Diseases in the Elderly*. Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/1660-4601/19/1/207

Cooper N. (2023, February 8). *World's Best Specialized Hospitals 2023*. Newsweek. https://www.newsweek.com/rankings/worlds-best-specialized-hospitals-2023

Corliss, J. (2022, August 1). *Too little sleep may be hard on your heart*. Harvard Health Publishing. http://www.health.harvard.edu/heart-health/too-little-sleep-may-be-hard-on-your-heart

Chase, B. (2023, June 12). *More evidence moderate drinking is good for your heart. Also: a reason.* The Harvard Gazette. https://news.harvard.edu/gazette/story/2023/06/is-drinking-in-moderation-good-for-your-heart/

Chen, H., Luo, X., Du, Y., He, C., Lu, Y., Shi, Z., & Zhou, J. (2023, August 31). *Association between chronic obstructive pulmonary disease and cardiovascular disease in adults aged 40 years and above: Data from NHANES 2013–2018 - BMC Pulmonary Medicine*. BioMed Central. https://bmcpulmmed.biomedcentral.com/articles/10.1186/s12890-023-02606-1

Dattani, S., Samborska, V., Ritchie, H., & Roser, M. (2023, December 28). *Cardiovascular Diseases*. Our World in Data. https://ourworldindata.org/cardiovascular-diseases

Gandler, G. Z. (2020, October 13). *Training models on Imbalanced Data*. Medium. https://towardsdatascience.com/training-models-on-imbalanced-data-561fa3f842b5

Interior Community Health Center. (2024, February). *February 2024: Signs and symptoms of heart disease*. Interior Community Health Center. https://www.interiorcommunityhealth.org/blog/february-2024-signs-and-symptoms-of-heart-disease#Symptoms_and_signs

Jadhav, H. (2023, August 2). *The Impact of Outliers in Machine Learning*. Medium. https://harshjadhav100.medium.com/the-impact-of-outliers-in-machine-learning-6cf4e4d33f21

Jahmunah, V., Ng, E. Y. K., Tan, R. S., & Acharya, U. R. (2021, May 7). *Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GABORCNN model with ECG signals*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0010482521002511?via%3Dihub

Jankowski, J., Floege, J., Fliser, D., Böhm, M., & Marx, N. (2021, March 15). *Cardiovascular Disease in Chronic Kidney Disease*. AHA Journals. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.120.050686

Klein, S., Gastaldelli, A., Yki-Järvinen, H., & Scherer, P. E. (2022, January). *Why does obesity cause diabetes?*. PubMed. https://pubmed.ncbi.nlm.nih.gov/34986330/

Lau, Q. Y. (2019, June 11). *Diagnosing the Heart of the Problem*. National Heart Centre Singapore. https://www.nhcs.com.sg/news/patient-care/diagnosing-the-heart-of-the-problem

Li, M., Zhou, B., & Hu, B. (2022, July 22). *Relationship between Income and Mental Health during the COVID-19 Pandemic in China*. International journal of environmental research and public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9330058/

Mark, P. B., Carrero, J. J., Matsushita, K., Sang, Y., Ballew, S. H., Grams, M. E., Coresh, J., Surapaneni, A., Brunskill, N. J., Chalmers, J., Chan, L., Chang, A. R., Chinnadurai, R., Chodick, G., Cirillo, M., de Zeeuw, D., Evans, M., Garg, A. X., Gutierrez, O. M., Heerspink, H. J. L., … Stengel, B. (2023). *Major cardiovascular events and subsequent risk of kidney* *failure with replacement therapy: a CKD Prognosis Consortium study*. European heart journal, 44(13), 1157–1166. https://doi.org/10.1093/eurheartj/ehac825

Miller, J. (2024, March 8). *Alcohol's Effects on the Heart*. AddictionHelp.com. https://www.addictionhelp.com/alcohol/effects/heart/

Nanyang Technological University. (2021, June 8). *New artificial intelligence tool invented by NTU, NP and NHCS scientists could speed up diagnosis of cardiovascular diseases*. Nanyang Technological University,. https://www.ntu.edu.sg/docs/default-source/corporate-ntu/hub-news/new-artificial-intelligence-tool-invented-by-ntu-np-and-nhcs-scientists-could-speed-up-diagnosis-of-cardiovascular-diseases78bcc0d4-268e-4e81-9885-44ddb462571e.pdf?sfvrsn=26fa5e92_3

NANYANG TECHNOLOGICAL UNIVERSITY. (2023, June 8). *New AI tool invented by NTU, NP and NHCS scientists could speed up diagnosis of cardiovascular diseases*. EurekAlert! https://www.eurekalert.org/news-releases/482134

Nicoll, D., Lu, C. M., & McPhee, S. J. (2017). *Benefits, costs and risks: Guide to diagnostic tests*. Anesthesia Central. https://anesth.unboundmedicine.com/anesthesia/view/GDT/619576/all/BENEFITS__COSTS_AND_RISKS

Pakhare, M., & Anjankar, A. (2024, January 4). *Critical Correlation Between Obesity and Cardiovascular Diseases and Recent Advancements in Obesity*. Cureus. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10838385/

Papaporfyriou, A., Bartziokas, K., Gompelmann, D., Idzko, M., Fouka, E., Zaneli, S., Bakakos, P., Loukides, S., & Papaioannou, A. I. (2023, May 31). *Cardiovascular diseases in COPD: From diagnosis and prevalence to therapy*. Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/2075-1729/13/6/1299

Parmar, M. P., Kaur, M., Bhavanam, S., Mulaka, G. S. R., Ishfaq, L., Vempati, R., C, M. F., Kandepi, H. V., Er, R., Sahu, S., & Davalgi, S. (2023, April 24). *A Systematic Review of the Effects of Smoking on the Cardiovascular System and General Health*. Cureus. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10208588/

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K., & Panguluri, S. K. (2019, April 27). *Cardiovascular Risks Associated with Gender and Aging*. U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616540/

Ryder, G. (2024, January 3). *Cardiac psychologists are pushing to protect heart patients' often-overlooked mental health*. STAT. https://www.statnews.com/2024/01/03/heart-disease-mental-health/

Paschalidis, Y. (2017, March 30). *How Machine Learning Is Helping Us Predict Heart Disease and Diabetes*. Harvard Business Review. https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes

Sehat, H. (2016, September 21). *Underweight people at elevated risk of heart diseases: Study*. The Jakarta Post. https://www.thejakartapost.com/life/2016/09/21/underweight-people-at-elevated-risk-of-heart-diseases-study.html

U.S. Department of Health and Human Services. (2022, March 24). *How Smoking Affects the Heart and Blood Vessels*. National Heart Lung and Blood Institute. https://www.nhlbi.nih.gov/health/heart/smoking

Volpe, M., & Gallo, G. (2023, March 13). *Obesity and cardiovascular disease: An executive document on pathophysiological and clinical links promoted by the Italian Society of Cardiovascular Prevention (SIPREC)*. Frontiers in cardiovascular medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10040794/

World Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1