NANYANG
TECHNOLOGICAL
UNIVERSITY

Lim Zi Xiang

Table of contents

Executive summary 1
1. Problem Statement 2
1.1 Introduction to problem statement 2
1.2 Reasons 2
2. Project Objectives 4
3. Data and Methodology 4
3.1 Data Sourcing 4
3.2 Data Cleaning 4
3.3 Data Exploration 7
4. Modelling 13
4.1 Overall approach 13
4.2 Pre-processing 13
4.3 Training 13
4.4 Models 13
5. Proposed Business Solution 20
5.1 Value proposition over existing models 21
6. Conclusion 21
References 22
Appendices 26

Executive summary

Singapore is facing an unprecedented health crisis, with a rapidly ageing population surging from 11.1% elderly to 18.4% over the past 10 years, further projected to increase to 25% in 2030 (Fang, 2022). Furthermore, increasingly more people suffer from lifestyle-related diseases from having a sedentary lifestyle (Lai, 2021), such as cardiovascular diseases, the second leading cause of death in Singapore amounting to nearly 1 in 3 deaths in 2018 (AIA, 2019). National Heart Centre Singapore (NHCS) already oversees hundreds of thousands of patients each year for cardiovascular diseases, and patient numbers are projected to further increase over the following years. The healthcare sector in Singapore shoulders an immense, growing burden that may eventually overwhelm our healthcare services.

To alleviate the stress on NHCS, we aim to make use of Artificial Intelligence (AI) and Machine Learning (ML) via predictive analysis for accurate and early detection of heart diseases. This allows NHCS to detect patients at a higher risk of heart disease, with the help of predictive analysis, and provide them with timely treatment. We also seek to reduce the resources needed for heart disease screening by providing a novel way of predicting heart disease based on non-medical data.

Currently, healthcare diagnostics tests are time consuming and expensive, often unaffordable to many patients. Furthermore, current AI models have high false negative rates where heart diseases go undetected.

Hence, this report aims to employ a model which effectively targets the above problem statements.

Firstly, we explored the different variables from the dataset via exploratory analysis. to examine their relation to heart disease and within the predictors. The emphasis is on how non-medical factors can influence heart disease, allowing us to extend these insights to patients subsequently help them reduce their likelihood of heart disease by correcting unhealthy habits.

Secondly, we applied five different models: Logistic Regression, Classification And Regression Tree (CART), Gradient Boosting Machine (GBM), Support Vector Machine (SVM) and Neural network. By comparing these models, we decided on the most accurate model in detecting the presence of heart disease.

Thirdly, the final model: Logistic regression model will be implemented into the app to allow patients to self-check their risk of heart disease through data available to them without needing advanced cardiac diagnostic tests or expert opinions from healthcare professionals. By accessing their risk of heart disease, individuals will be recommended on subsequent steps of action such as an appointment with the doctor.

Overall, this helps target our problem statement of time consuming and expensive tests using non-medical data to allow patients to gauge their risk prior and decide on their next course of action. With our model's high recall score, we also avoid giving high risk individuals a false sense of safety.

This will overcome the potential business problems as the app is easily accessible and cost effective. The comprehensive nature of the app allows individuals to make well informed decisions on whether it is necessary to consult a cardiologist based on the results of the diagnostic test taken through the app.

1. Problem Statement

1.1 Introduction to problem statement

Despite having advanced healthcare procedures in place to diagnose heart diseases, there are still limitations faced as stated below.

Time-Consuming Diagnostic Tests
Currently, NHCS uses a series of complex cardiac diagnostic tests in the diagnosing of heart diseases. These tests are time-consuming. For instance, the Ambulatory Electrocardiogram requires the patient to be monitored for 24-48 hours (Lau, 2019). However, nearly 50% of potentially salvageable heart muscle is lost within an hour of the coronary artery being blocked. Given, the time sensitivity of the test results, many patients could have succumbed to life-threatening episodes of heart attacks and other symptoms while waiting for the test results. Only 26% of 500 heart attack survivors managed to get treatment in time (SBS News, 2016).

High Cost of Tests
These diagnostic tests are also highly expensive (Table 1.1), potentially deterring high-risk patients from taking health screenings or seeking treatment for their conditions – the delay in early treatment of heart disease, would, besides worsening health outcomes for patients, exacerbate the strain on our healthcare sector as the treatment required would demand more attention and medical resources.

Consultation	Electrocardiogram	Echocardiogram	Treadmill Test
$180-$300	$60-$80	$450-$600	$350-$500

Table 1.1. Cardiac Diagnostic Test Costs in Singapore (Macdonald, 2022)

Manpower Shortage
Manpower, a critical resource for Singapore's healthcare system, is diverted away to conduct these time-consuming tests. This is especially pressing as there is an existing workforce shortage in Singapore's healthcare sector, to which the government has desperately responded by increasing reliance on foreign workers, subsidising salary increases, and rolling out staff support and wellness programmes (RSM, 2021). These approaches, while beneficial to the employees, are draining resources from both the government and the hospitals which can be better utilised.

Scant Medical Records
Cardiac patients are initially assessed based on their medical history, as well as that of their families. Therefore, the validity of the diagnosis relies heavily on the availability and accuracy of these records, which may not always be complete and available for medical practitioners.

1.2 Reasons

1.2.1 Incompleteness of current early detection models
Currently, there exist multiple models that predict heart disease based on medical data. However, beyond medical data, behavioural and psychological factors could also affect the risk of heart disease. There is tremendous, untapped potential in using these factors for more robust and powerful analytics than simply using medical data.

A study examined 4 conventional heart disease prediction approaches and found that the accuracy of the models ranged from a high of 85% to 95% (Table 2.1).

Table 1

Comparative analysis of different machine learning methods.

Methods	Accuracy	Precision	Recall	F1-measure
Naïve Bayes weighted approach	86.00	82.34	87.25	89.21
2 SVM's and XGBoost	94.03	86.56	94.78	92.79
SVM and DO	89.4	66.1	81.3	82.1
XGBoost	95.9	97.1	94.67	95.35

*Table 2.1. Analysis of the 4 heart disease detection models (*Nagavelli, et al., 2022)

Despite the deceptively high accuracy, these models had a concerning false negative rate of 40-45% for all 4 models. False negatives are especially dangerous as this means 40% of cases of heart disease go undetected, and consequently, untreated. Therefore, another priority of our model is to minimize the false negative rate in prediction of heart disease.

Naïve Bayes and weighted approach	2 SVM's and XGBoost	SVM and DO	XGBoost
44.5%	47.5%	43.4%	45.7%

*Table 2.2. A false negative rate of 4 heart disease detection models (*Nagavelli, et al., 2022)

1.2.2 Usefulness of non-medical factors in early detection models

The World Health Organisation states that the key behavioural factors of cardiovascular diseases are the use of tobacco, lack of physical activity and consumption of unhealthy food. (WHO, 2021). The usefulness of behavioural factors can also be seen by a quantised study that was conducted over 6 months at the diabetes units of health centres in Iran. The study showed that 9 modifiable risk factors such as smoking and consumption of fruits and vegetables were associated with more than 90% of the risk of a heart attack in this large global case-control study in 52 countries (Sabzmakan, et al., 2014). Furthermore, even physiological factors have been shown to be correlated with heart disease. Depression for instance, is associated with a 3-fold increase in the risk of heart disease. When Coronary Heart Disease (CHD) risk estimates were combined across studies, depression was associated with a significantly increased risk of CHD with a combined risk ratio (RR) of 1.64 (95% confidence interval [CI]=1.29-2.08, P<.001), with significant heterogeneity between studies. Hence, this highlights the importance of behavioural factors given their correlation to heart disease (Smith et al., 2011).

1.2.3 Current models' reliance on complex medical data

Most modern heart disease early detection models require complex medical data, which in turn requires extensive medical tests to be conducted prior. This may not be well-received, especially by the elderly. Additionally, people who have not done all these medical tests will not have the data required for conventional prediction models to predict, so they won't be able to receive an early diagnosis until they undergo the expensive and resource-draining cardiac diagnostic tests.

2. Project Objectives

We would like to build a model that predicts the probability of heart disease based on non-medical data which are easily accessible by everyone, such as daily vegetable intake, age, mental well-being and more, so that a heart disease diagnosis can be quickly generated anytime, anywhere, by anybody.

3. Data and Methodology

3.1 Data Sourcing

We aim to construct a model that predicts likelihood of CHD based on behaviour and demographics. To construct a complete and accurate model, we need to consider a wide range of behavioural & demographic factors that affect risk of CHD.

The Behavioural Risk Factor Surveillance System (BRFSS) conducted public health surveys through the telephone, collecting objective state data of U.S. citizens based on their behaviour which might result in chronic health problems in the long run. (CDC[1], 2022) The data comes from an official source, and contains an extensive collection of non-medical factors that could potentially contribute towards risk of CHD.

To better suit this dataset for Singapore use, we avoided taking U.S. specific variables like state of residence, whether the person has served in US armed forces, and housing types. Instead, we focused on more general factors that apply cross-borders.

3.2 Data Cleaning

BRFSS has provided a comprehensive dataset with 441,456 rows and 330 columns, but some columns are duplicated and unrelated, such as interview date & country of residence. Furthermore, the dataset is too large to perform data exploration and data modelling. Therefore, we put the dataset through a sequence of dimensionality reduction and cleaning procedures in our data processing pipeline, extracting only the useful information.

3.2.1 Recoding values

We start by correcting the encodings used by the survey. Based on the codebook, columns like MENTHLTH use the encodings: 88 for "None", 77 for "Don't Know/Not Sure", and 99 for "refused". Left untreated, our models will misinterpret these values. Therefore, these structural errors need to be fixed by recoding values as 0 for "None", NA for "Don't Know/Not Sure" and "Refused".

The sequence of questions also matters. For instance, if the respondent has previously responded to ALCDAY5 with 888[2], then AVEDRNK would not be asked, but will be recorded as NA. This explains why questions further down the survey have more missing values. In fact, AVEDRNK2 has 52.2% NA values. To fix this column, we recorded a 0 in AVEDRNK2 where ALCDAY5 == 888, reducing NA values to 5.82%. Same applies for SCNTWRK1.

3.2.2 Removing duplicates

The dataset had no duplicate rows.

3.2.3 Extracting relevant columns

We extracted a subset of the data frame to include non-medical factors. Behavioural factors like total number of alcoholic drinks consumed per week and demographic factors like weight were chosen. Unrelated columns like date month and state were filtered out.

3.2.4 Creating the response variable

We dropped CVDCRHD4 since it was found to be a subset of MICHD. MICHD was renamed to "Y".

3.2.5 Correcting numeric values

According to the codebook, WTKG3, BMI5, GRENDAY, FRUITDA1 and BEANDAY in the dataset have an implied 2 decimal points. Thus, we divide the values in these columns by 100.

3.2.6 Correcting data types

PA1MIN, PA1VIGM, SCNTWRK1, CHILDREN are set as integers.

SEX, HLTHPLN1, MEDCOST, ADDEPEV2, MARITAL, EMPLOY1, INCOME2, INTERNET are set as categorical variables.

SMOKER3, CHECKUP1, SEATBELT, EDUCA, AGEG5YR are set to ordinal variables.

3.2.7 Remove rows with missing response variable

In supervised learning, all the response variables must be available for testing and testing. Hence, we filtered out the rows where Y is missing since they are unusable.

3.2.8 Removed rows with missing values in columns where more than 20% data are missing

Some of the critical columns have high percentages of missing data. To ensure the usability of these columns while preserving the truthfulness of the data, we chose to drop all rows where data is missing in the columns with more than 20% missing values (SCNTWRK1, PA1VIGM, PA1MIN) as opposed to resorting to imputation which introduces error and biases into our data. With 441,456 records to work with, these errant rows could be removed without causing data insufficiency.

3.2.9 Removed rows with less than 80% of that rows' columns filled [only keep rows that have less than 20% NA value]

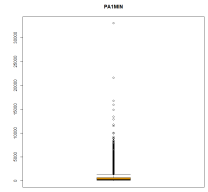
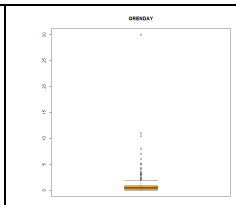
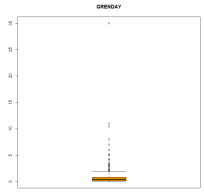
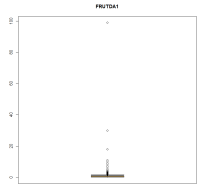
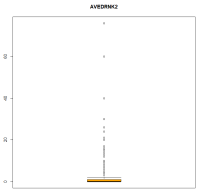
Given the large number of rows, we can further reduce the dimensionality and still have sufficient data for modelling. Hence, to ensure even greater accuracy in the models like regression, through the predominant usage of actual over imputed values, we removed rows with more than 20% missing values.

3.2.10 Balancing the dataset

Naturally, the dataset is highly imbalanced with a 1:9 proportion (heart disease: no heart disease). With imbalanced data, the model will develop a bias towards predicting the majority class, which is the case of no heart disease (Brownlee, 2020). This gives rise to a higher false negative rate which is a serious error as high-risk patients will be predicted to have no heart disease which could be fatal if ignored. Hence, we need to balance the data 1:1 through under sampling to reduce false negative rates.

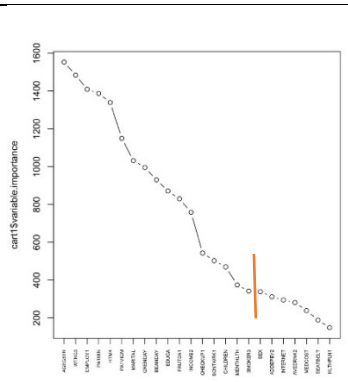
3.2.11 Removal of outliers

Through examination of statistical summary and boxplots, certain columns: PA1MIN, GRENDA, BEANDAY, FRUTDA and AVEDRNK2 have extreme or unreasonable values. Hence it is important to remove the outliers to make the data less skewed especially as outliers can drastically distort or change the fit estimates and predictions (Prabhakaran, 2017).

				
PA1MIN: Minutes of total Physical Activity per week. PA1MIN has extreme and unreasonable outliers. For example, the maximum value of PA1MIN is 33080 minutes (551 hours). There are many anomalous data points as such which exceed maximum hours in a week. These outliers must be removed.	GRENDAY: Dark green vegetable intake in times per day GRENDAY has extreme outliers too. Maximum green vegetables intake is 30 times a day. These extreme outliers will be removed.	BEANDAY: Bean intake in times per day BEANDAY has extreme outliers as well. The maximum value of BEANDAY for instance, which is very distant from the other points, so these outliers will be removed.	FRUTDA1: Fruit intake in times per day FRUITDA1 has unreasonable outliers too. For example, the maximum value of FRUITDA1 is 99. It is impossible for someone to consume fruits 99 times per day. This is unreasonable and thus outliers in FRUITDA1 were removed.	AVEDRNK2: Total number of drinks per day AVEDRNK2 has a max value of 72 drinks a day, way beyond the life-threatening limit. Thus, outliers were removed from AVEDRNK2 as well.

3.2.12 Using CART model to further reduce dimensionality through variable importance

The dataset set still has too many columns (27 columns) to perform modelling. Hence using a CART model, we further filtered the vital columns by variable importance.

	Through the plot, we decided to cut off between SMOKER and SEX to reduce dimensionality of the dataset by 30% for modelling. Hence utilising this as the cut-off point, the variables lesser than this threshold is removed from the dataframe.

3.2.13 Data imputation

After removal of rows, the remaining missing values are imputed using the K-Nearest Neighbour method (KNN).

KNN is our imputation method of choice as it allows for imputation of both numerical and categorical variables. KNN imputation can also be done in the presence of other missing values by utilizing alternative columns to compute distances, making it more robust than imputation via regression models.

KNN is a distance-based algorithm that typically uses Euclidean distance to compute closeness of different rows. However, we did not normalize the values because KNN from the VIM package uses Gower distance instead, where numeric variables are range normalized.

3.2.14 End of data cleaning

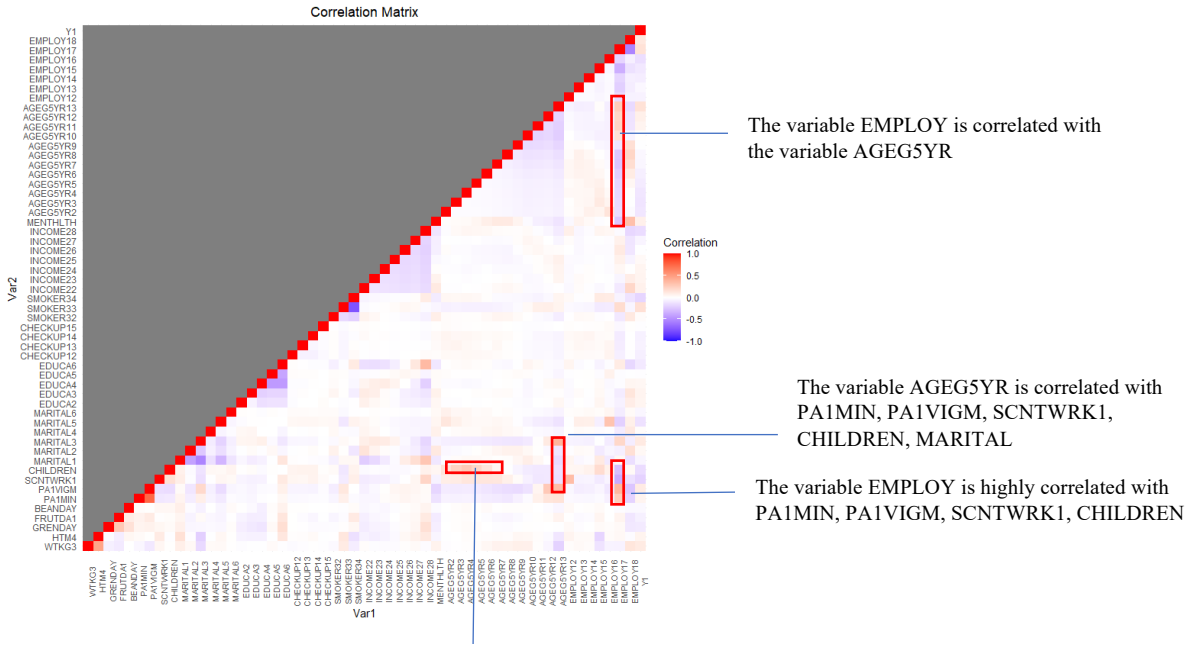
The final dataset had a shape of (21214, 18).

Finalized columns: WTKG3, HTM4, AGEG5YR, EMPLOY1, GRENDAY, FRUTDA1, MARITAL, BEANDAY, EDUCA, INCOME2, PA1MIN, PA1VIGM, CHECKUP1, MENTHLTH, SCNTWRK1, CHILDREN, SMOKER3, Y

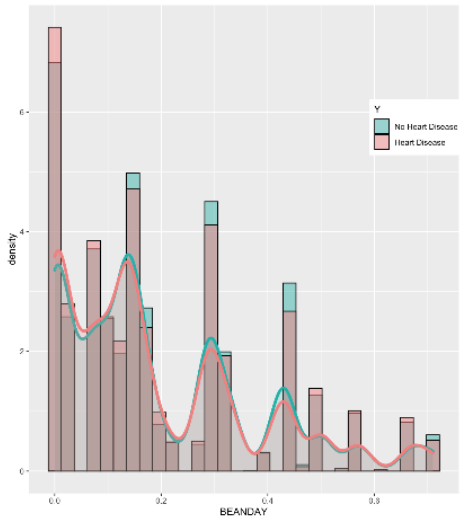
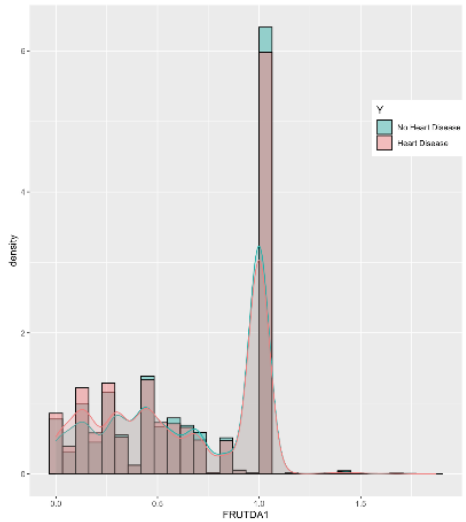
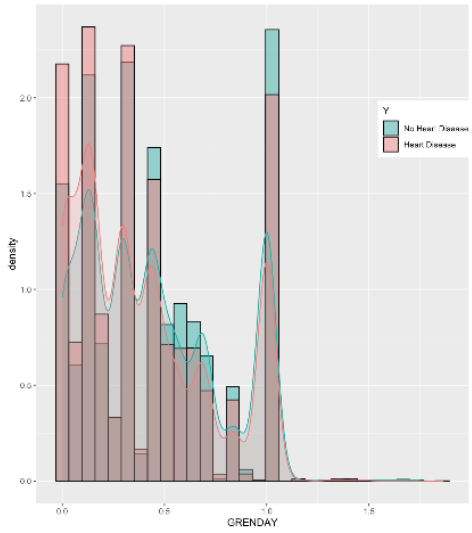
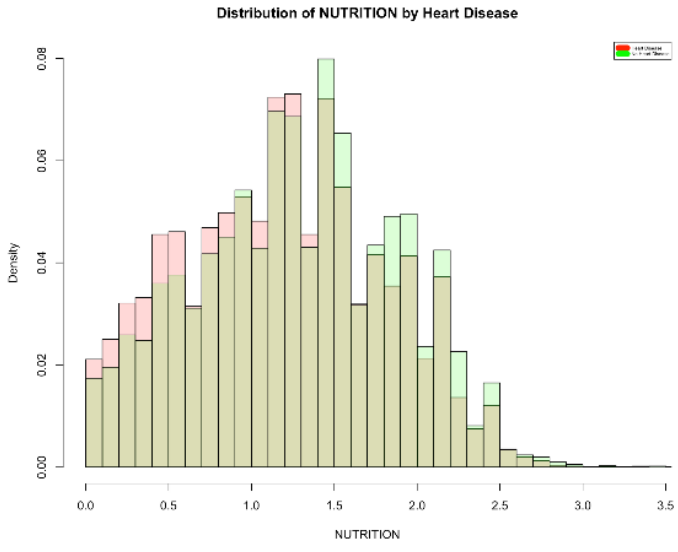
3.3 Data Exploration

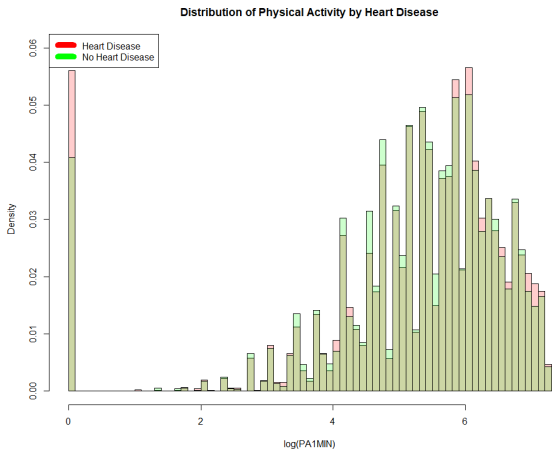
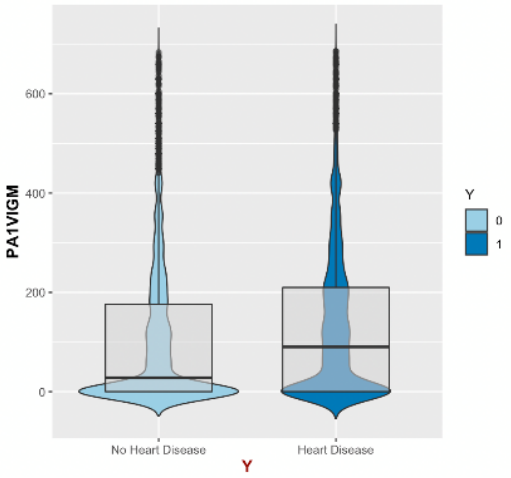
3.3.1 Bivariate analysis

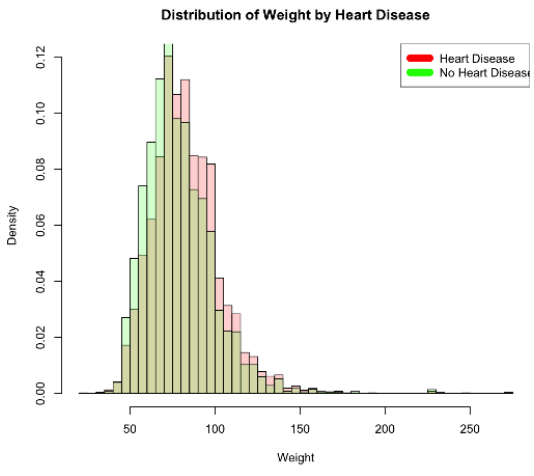
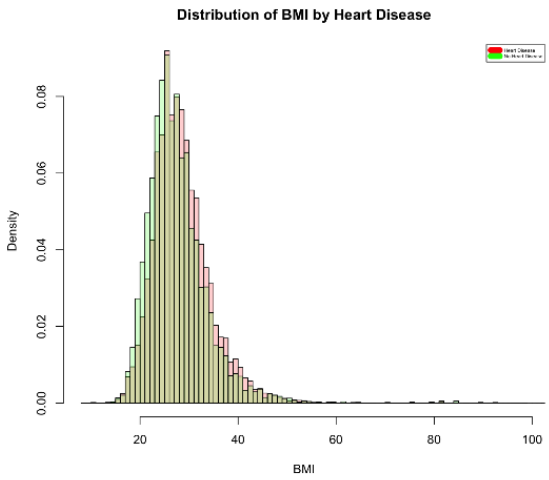
We used the correlation matrix to examine linear correlations between pairs of variables.



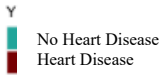
Continuous variables

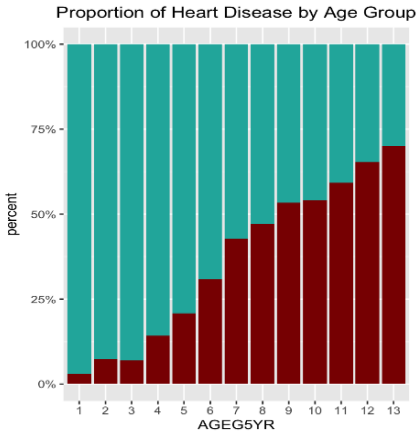
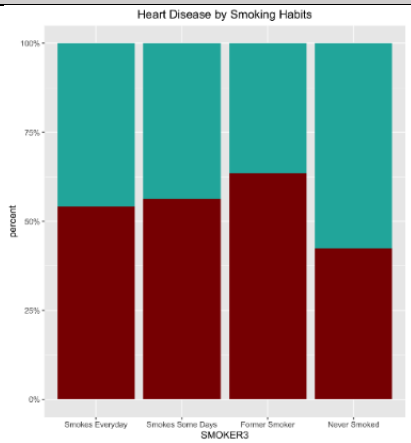
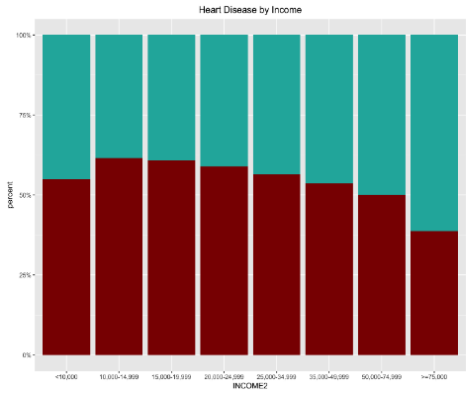
<table>
<tr><th colspan="3">NUTRITION VS HEART DISEASE</th></tr>
<tr><th>BEANS INTAKE</th><th>FRUITS INTAKE</th><th>VEG. INTAKE</th></tr>
<tr>
<td> 
</td>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>From the graph, the predominant trend is that the lower the bean intake, the greater the risk of heart disease. As beans intake increases beyond 0.3, there is a lower risk of heart disease.</td>
<td>From the graph, the predominant trend is that the lower the fruit intake, the greater the risk of heart disease. As fruit intake increases beyond 0.4, there is a lower risk of heart disease.</td>
<td>From the graph, the predominant trend is that the lower the vegetable intake, the greater the risk of heart disease. As vegetable intake increases beyond 0.45, there is a lower risk of heart disease.</td>
</tr>
<tr><th colspan="3">OVERALL NUTRION INTAKE</th></tr>
<tr>
<td colspan="2"> 
</td>
<td>The graph shows a linear combination of all three variables: GRENDAY, FRUTDA1, BEANDAY against the response variable, MICHD. As the nutrition intake increases beyond 1.4 times, there is a lower risk of heart disease. This is synchronous to research which shows that an increase in the consumption of vegetables or fruits by a serving reduces the risk of heart disease by 4%.(K J et.al, n.d.) Similarly, consumption of beans reduces blood cholesterol, a leading cause of heart disease. (American Heart Association, 2018).</td>
</tr>
</table>

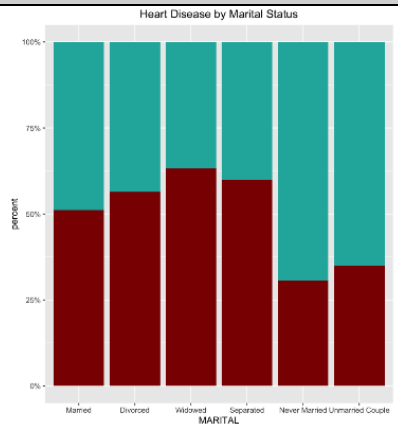
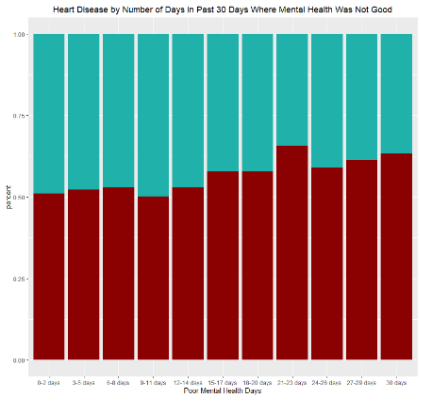
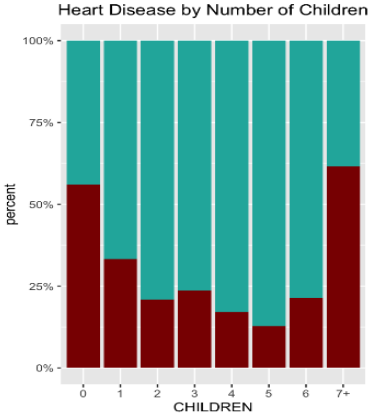
PHYSICAL FITNESS VS HEART DISEASE	
NORMAL PHYSICAL ACTIVITY	**INTENSE PHYSICAL ACTIVITY**
	
We observe that the longer the duration of physical activity, the lower the risk of heart disease. This is in line with the research which shows that adults who had 150 to 300 minutes of physical activity had a 63% lower risk of having heart disease. (American Heart Association, 2022)	By comparing the plots, it is observed that the more intense the physical activity is, the higher the risk of heart disease. This is synchronous with the research which shows that intensive exercise may increase the risk of getting cardiac arrest. (Healthessentials, 2020)

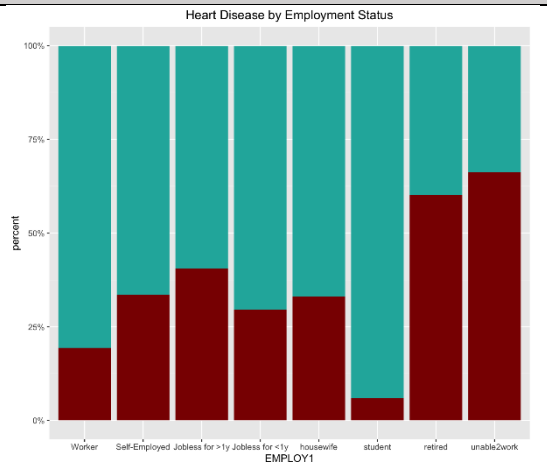
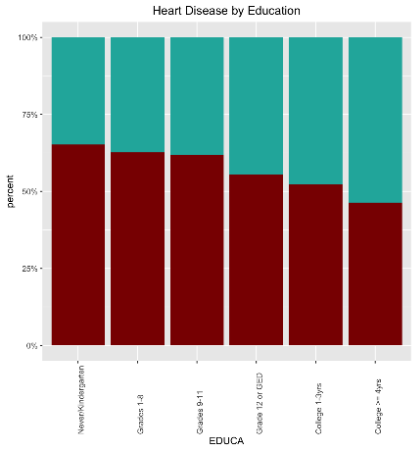
BODY MEASUREMENTS VS HEART DISEASE	
WEIGHT	**BMI**
	
The graph shows that the heavier an individual, the higher the risk of heart disease. This is synchronous with research which highlights that excess weight can lead to fatty material building up in your arteries which increases risk of heart disease (BHF, n.d.). This can be especially seen when weight increases beyond 70kg. This is accompanied by a higher risk of heart disease.	The graph shows that as BMI increases, the higher the risk of heart disease. This is synchronous with research which highlights the danger of crossing the healthy range of BMI as it increases the chance of contracting heart diseases (John Hopkins, n.d.). This can be especially seen when BMI increases beyond 30. This is accompanied by a higher risk of heart disease.

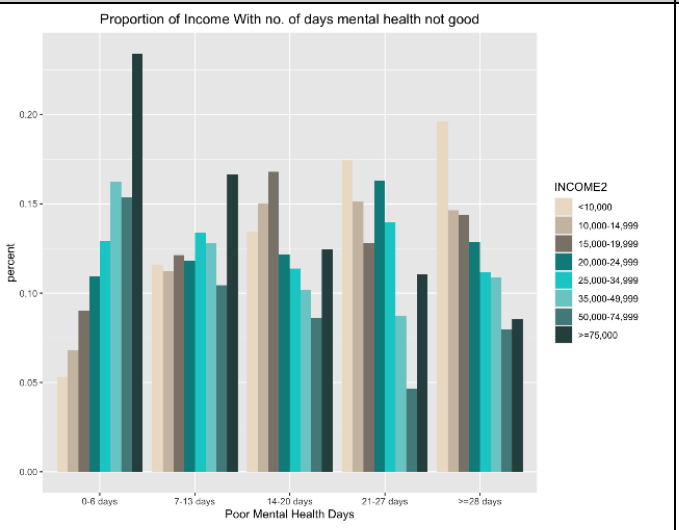
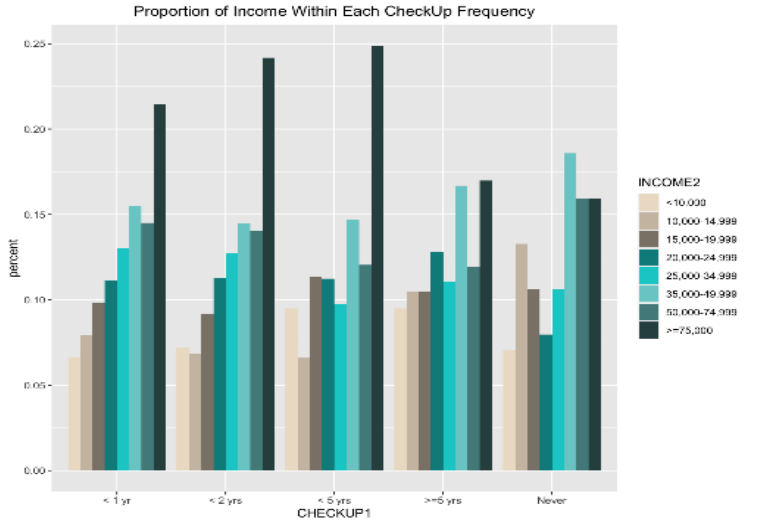
Categorical Variables

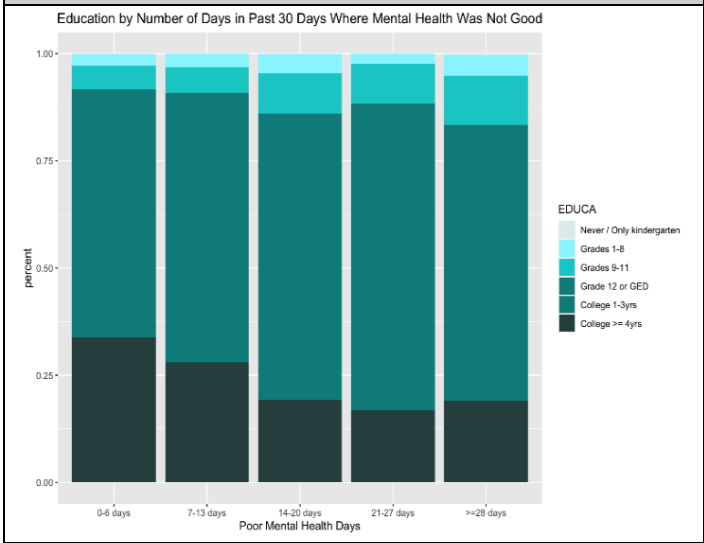
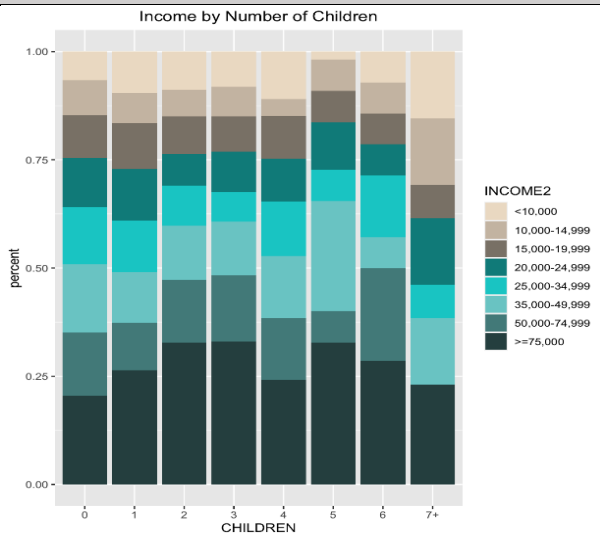


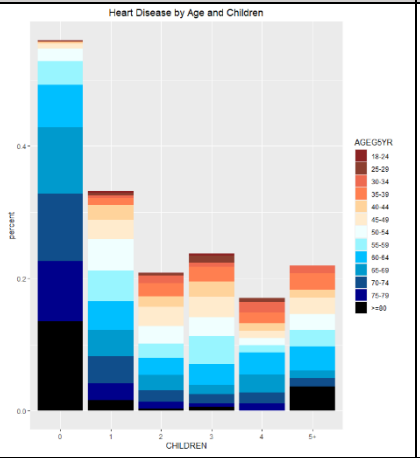
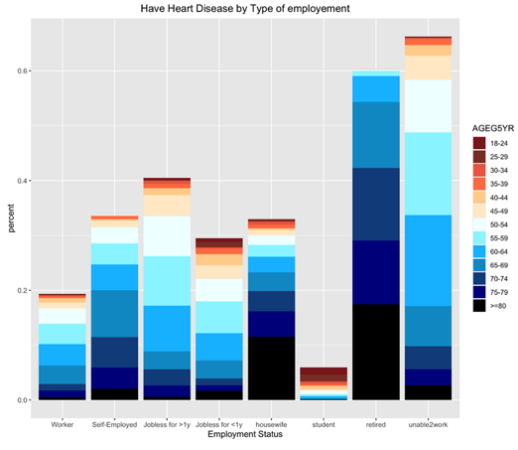
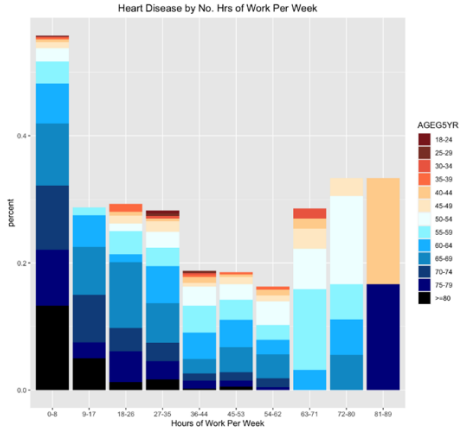
AGE GROUP	SMOKING HABITS	INCOME
		
Clear positive correlation between age and risk of heart disease. This is synchronous with research that emphasises that aging can cause changes in the heart and blood vessels that increases a person's risk of developing cardiovascular disease (NIH, n.d.).	Individuals who have or had smoked face a higher risk of heart disease. This is in line with the research which explains that 20% of smoking-related deaths is caused by heart disease. (Johns Hopkins Medicine, 2020)	As income increases, risk of heart disease decreases. This is in line with synchronous with research that emphasises that individuals with lower income have 50% higher risk of developing disease as they have poor access to health care treatments. (Rattue, 2011)

MARITAL STATUS	MENTAL HEALTH	NO. OF CHILDREN
		
Widowed, Divorced and Separated individuals face higher risk of heart disease. These 3 categories are associated with mental distress amongst individuals, synchronous with research that highlights that long periods of mental distress can trigger the physical effects of heart disease (SCL Health, 2022).	As the number of days of having poor mental health increases, risk of heart disease increases. This is in line with the research that explains how having a poor mental health for a prolonged period may lead to depression and stress which resulting in an increased cardiac reactivity. (CDC, 2020)	Individuals with no children and individuals with at least 7 children have a higher risk of heart disease. This is in line with the research which states that low sex hormone levels, that results in infertility, is a key factor that causes heart disease. (Rettner, 2011) Similarly, another study found that the individuals with more than 5 children face 26% higher chance of heart disease. (Williams, 2018)

EMPLOYMENT STATUS	EDUCATION
	
Unemployed individuals are at a higher risk of heart disease. This is most prominent in retired and unable-to-work individuals which could be largely attributed to age. However, the jobless individuals also have a greater likelihood of getting heart disease. This is synchronous with research as a study showed that unemployment and loss of jobs may increase the risk of heart attack (Mann, 2012).	The more educated an individual is, the lower the risk of heart disease. This is synchronous with the research that emphasises that with higher level of education, individuals are more likely to get better jobs which would enable them to get health insurance and thus seek a better treatment. Males with grade school level education were 17% more likely to have heart disease compared to males with graduate school education. (Merschel, 2020)

BIVARIATE ANALYSIS BETWEEN 2 PREDICTOR VARIABLES	
INCOME VS MENTAL HEALTH	**INCOME VS CHECKUP FREQUENCY**
	
Individuals with higher income tend to have better mental health. This is synchronous with the research which explains that individuals with lower income have higher suicide rate as one of the main concerns is to meet the basic needs. (Sareen et.al, 2011)	From the graph, it can be observed that individuals with higher income go for a routine check-up compared to individuals with a lower income. This is in line with the research that states how higher income individuals have greater affordability to access qualified hospitals and thus allowing them to go for routine medical check-ups. (Shin et.al, 2018)

EDUCATION VS MENTAL HEALTH	INCOME VS CHILDREN
	
From the graph, the more educated individuals have better mental health. This is synchronous with research which highlights that higher education levels are associated with greater stability and more life choices which in turn results in better mental health (William, 2022).	From 0 to 3 children, income generally grows, showing that finance is a main consideration in deciding how many children to have. With higher income, couples can afford to have more children. However, from 3 children onwards, having more children is associated with lower income. This is in line with the research which shows that lower income families have more children to support their livelihood. (World Vision, 2020).

MULTIVARIATE ANALYSIS		
HEART DISEASE BY AGE & CHILDREN	**HEART DISEASE BY AGE & EMPLOYMENT**	**HEART DISEASE BY AGE & HOURS OF WORK PER WEEK**
		
At first glance, it appears that the more children one has, the lower the chances of getting heart disease. However, the proportion of older people falls as children increases, so the lowered	Retires consist of older people, and hence have higher risk of heart disease. Conversely, students have the lowest risk due to lower age. Interestingly, people who are unable to work have the highest risk. Even though they are not the oldest, they	Individuals working 0-8 hours a week tend to be retirees, and hence face higher risk of heart disease due to their old age. We see a downward trend in incidence of heart disease from 9-17 to 54-62 hours a week as this corresponds to a generally younger age

incidence of heart disease could actually be a result of lower ages. Nonetheless, having children can reduce loneliness and even bring fulfilment, thereby improving mental health and lowering heart disease risk.	face a greater health risk due to pre-existing, crippling health conditions that may pose as risk factors for heart disease, such as severe mental illnesses and extreme obesity.	distribution. The minimum point is at 54-62 hrs of work / week, corresponding to a typical 9 hour, 5/6 day work week. Beyond this level, heart disease risk increases. Even if people in these categories are younger, the effects of overworking triumphs, drastically increasing risk of heart diseases (Sunrise Hospital and Medical Center, 2020).

4. Modelling

4.1 Overall approach

We utilized 5 different models to predict heart disease, and derived the best model based on context-specific metrics.

4.2 Pre-processing

Feature scaling was first conducted to normalize the numeric variables to between 0-1 since the variables have highly varying magnitudes. PA1MIN for instance, ranges from 0 to 1300+, whereas MENTHLTH ranges only from 0-30. Without feature scaling, distance-based models like SVM may be biased towards features with higher magnitudes (Bhandari, 2020).

Feature scaling also smoothens the gradient descent process, especially for Neural Network and SVM, helping the models converge quicker (Raj, n.d.).

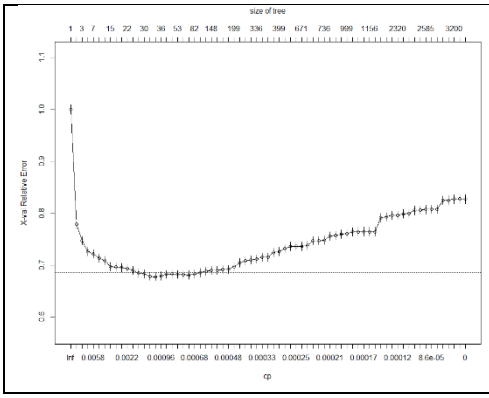
4.3 Training

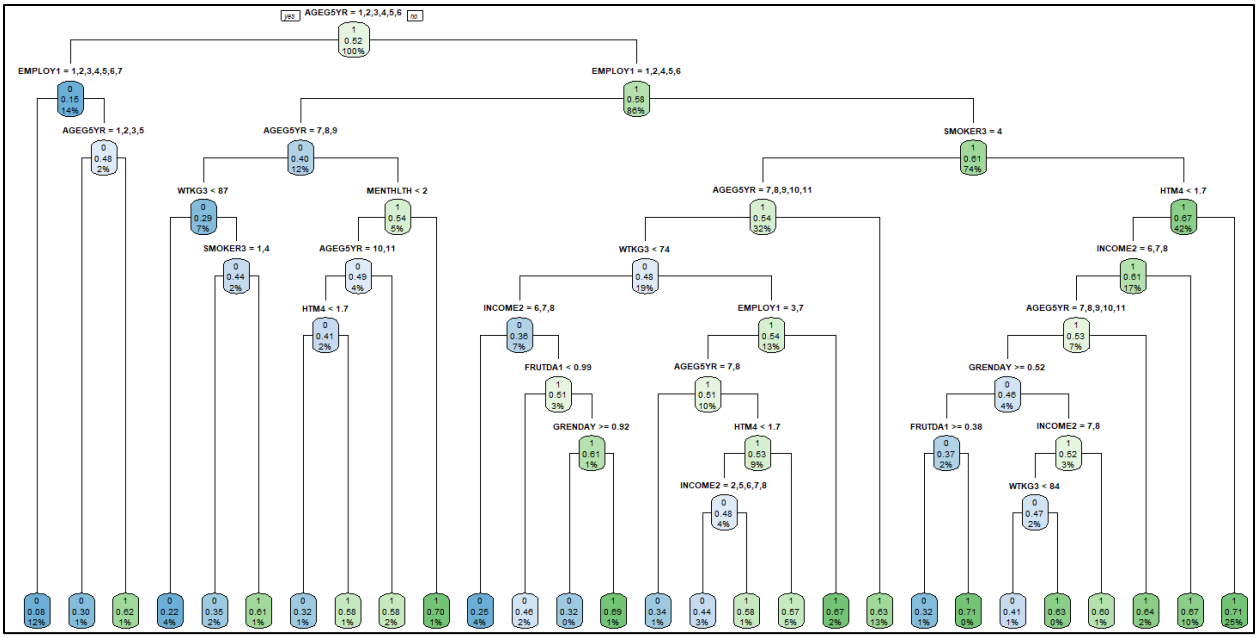
Since decision trees and ensemble methods like CART and GBM are not sensitive to variance in data (Thenraj, 2020), the non-scaled dataset is used for these models, whereas the scaled dataset is used for logistic regression, neural network and SVM. All 5 models were trained using a 70-30 train test split.

4.4 Models

4.4.1 CART

CART is a binary tree, where each split is based on the values of a predictor variable. The best splits are calculated based on the splitting criteria of minimizing Gini Impurity.

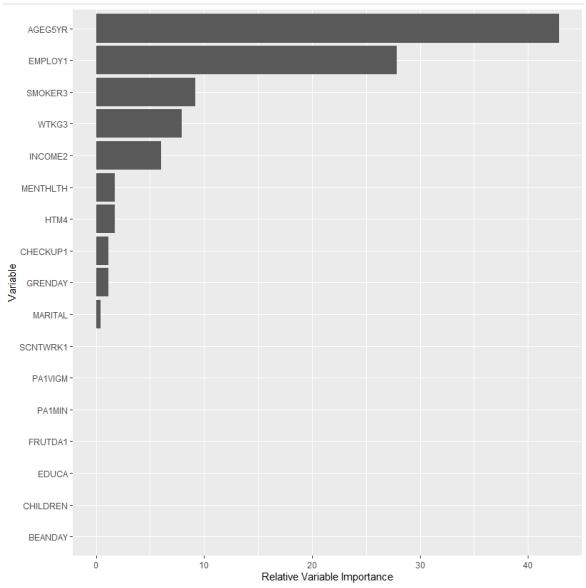
	After training on the train data set, the CART model was pruned to its optimal tree using the minimum Complexity Parameter (CP) + 1 standard error method. Optimal CP was found to be at 0.001760394.



4.4.2 GBM

Gradient boosting is an ensemble method where outputs from individual trees are combined. Weak learners are combined sequentially such that each new tree corrects the errors of the previous one.

Bernoulli loss function was used since this is a classification problem. 5 cross-validations were conducted for each of the 100 cross-validation iterations.



AGEG5YR and EMPLOY1 were identified as the most influential predictors, followed by SMOKER3, WTKG3, INCOME2, MENTHLTH, HTM4, CHECKUP1, GRENDAY and MARITAL.

SCNTWRK1, PA1VIGM, PA1MIN, FRUTDA1, EDUCA, CHILDREN and BEANDAY were deemed by the model to have 0 influence.

4.4.3 Logistic Regression

The logistic regression model takes a linear combination of the 17 independent predictors to model the probability of having heart disease.

Backward stepwise logistic regression is used to optimize the model, where variables are removed from the original logistic regression model on a step-by-step basis to prevent overfitting, reducing multicollinearity between the variables used.

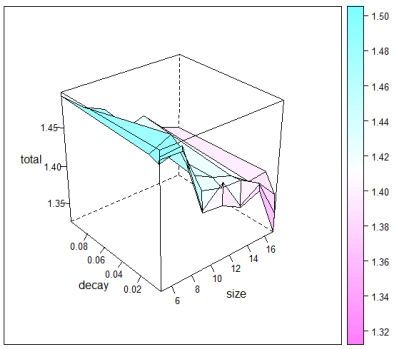
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.103378   0.363401 -11.292  < 2e-16 ***
WTKG3        3.013905   0.282567  10.666  < 2e-16 ***
HTM4         1.206896   0.230163   5.244 1.57e-07 ***
AGEG5YR2     0.735057   0.378435   1.942 0.052093 .
AGEG5YR3     0.607264   0.379061   1.602 0.109151
AGEG5YR4     1.041966   0.355110   2.934 0.003344 **
AGEG5YR5     1.323321   0.343964   3.847 0.000119 ***
AGEG5YR6     1.953981   0.329134   5.937 2.91e-09 ***
AGEG5YR7     2.384501   0.320822   7.432 1.07e-13 ***
AGEG5YR8     2.531512   0.318872   7.939 2.04e-15 ***
AGEG5YR9     2.894812   0.317700   9.112  < 2e-16 ***
AGEG5YR10    3.092737   0.318543   9.709  < 2e-16 ***
AGEG5YR11    3.374416   0.319867  10.549  < 2e-16 ***
AGEG5YR12    3.687066   0.322073  11.448  < 2e-16 ***
AGEG5YR13    4.063344   0.322218  12.611  < 2e-16 ***
EMPLOY12     0.546817   0.168483   3.246 0.001172 **
EMPLOY13     0.508429   0.136156   3.734 0.000188 ***
EMPLOY14     0.461076   0.154358   2.987 0.002817 **
EMPLOY15     0.318038   0.110805   2.870 0.004102 **
EMPLOY16     0.155447   0.268421   0.579 0.562512
EMPLOY17     0.539823   0.090150   5.988 2.12e-09 ***
EMPLOY18     1.228747   0.100355  12.244  < 2e-16 ***
GRENDAY     -0.515733   0.106800  -4.829 1.37e-06 ***
MARITAL2    -0.201367   0.060084  -3.351 0.000804 ***
MARITAL3    -0.217838   0.054272  -4.014 5.97e-05 ***
MARITAL4     0.078315   0.151495   0.517 0.605194
MARITAL5    -0.423937   0.077862  -5.445 5.19e-08 ***
MARITAL6    -0.086823   0.165257  -0.525 0.599317
INCOME22    -0.094818   0.100395  -0.944 0.344937
INCOME23    -0.084066   0.096584  -0.870 0.384089
INCOME24    -0.143234   0.096351  -1.487 0.137124
INCOME25    -0.354926   0.095434  -3.719 0.000200 ***
INCOME26    -0.416259   0.095078  -4.378 1.20e-05 ***
INCOME27    -0.575927   0.097771  -5.891 3.85e-09 ***
INCOME28    -0.786193   0.097239  -8.085 6.21e-16 ***
PA1VIGM     -0.247584   0.093024  -2.662 0.007779 **
CHECKUP12   -0.246928   0.071258  -3.465 0.000530 ***
CHECKUP13   -0.533181   0.105342  -5.061 4.16e-07 ***
CHECKUP14   -0.417487   0.110734  -3.770 0.000163 ***
CHECKUP15    0.258705   0.269812   0.959 0.337642
MENTHLTH     0.602149   0.079120   7.611 2.73e-14 ***
SMOKER32     0.007596   0.109358   0.069 0.944626
SMOKER33     0.002692   0.071436   0.038 0.969945
SMOKER34    -0.518613   0.070283  -7.379 1.60e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

After stepwise regression, only WTKG3, HTM4, AGEG5YR, EMPLOY1, GRENDAY, MARITAL, INCOME2, PA1VIGM, CHECKUP1, MENTHLTH and SMOKER3 remain.

PA1MIN, BEANDAY, FRUTDA1, EDUCA, CHILDREN, SCNTWRK1 were deemed statistically insignificant, and hence excluded as predictors.

4.4.4 Neural Network

In the nnet package, the neural network architecture is fixed at a single hidden layer. The model was trained over a maximum of 200 iterations or until convergence.



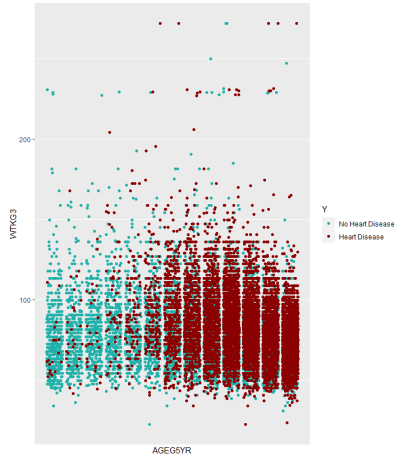
Hyperparameter tuning is done for number of nodes within the single hidden layer, as well as for weight decay.

Through grid search, we found that the optimal size of 7 nodes and weight decay of 0.01 gives the best recall and accuracy.

.

4.4.5 SVM

SVM draws hyperplanes in a N-dimensional space such that points on either side of the plane belongs to either one of the categories in a binary classification. The optimal hyperplane is one that maximizes its distance to points in either category



Using the example of AGEG5YR and WTKG3 in a 2-dimensional plane, we see that the original inputs are not linearly separable by class.

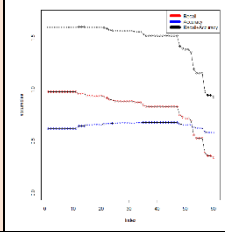
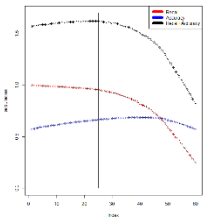
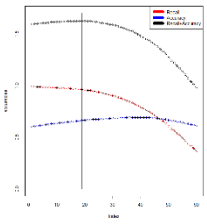
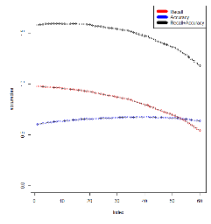
The kernel trick is therefore used with the linear transformation sigmoid function (Meyer. D, 2022):
tanh(γu0v + coef0)

Original input data is mapped into a higher dimensional space such that the classes become more linearly separable (Wilimitis, 2018).

4.5 Threshold Moving

After predictions were made, we sought to find the optimal threshold value under threshold moving to optimize the recall and accuracy of the models (Brownlee. J, 2020). Optimal threshold was found using a graphical approach, where optimal threshold is at maximum point of recall + accuracy as shown below.

SVM does not require a threshold for predictions.

	CART	GBM	Logistic Regression	Neural Network	SVM
Accuracy/ Recall trade-off					N/A
Optimal threshold	0.31	0.35	0.29	0.28	N/A

4.6 Models Evaluation

	CART	GBM	Logistic Regression	Neural Network	SVM
Accuracy	66.69285	66.59859	**66.41005**	66.77141	67.24273
Recall	93.19136	95.62088	**96.19076**	95.35093	82.69346
Precision	62.14000	61.68731	**61.46033**	61.86028	64.64244
F1 Score	74.56204	74.99412	**75.00000**	72.03836	72.56218

SVM has the highest classification accuracy, but in medical related analytics, recall is the best metric because we want to capture as many positives as possible to minimize chances of missing positive cases, especially when misclassifying a high-risk patient as having no heart disease can have fatal consequences due to late detection.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Recall evaluates the ratio: Number of correctly identified heart disease cases / (Number of correctly identified heart disease cases + Number of misclassified cases for positive heart disease cases). Therefore, the final model that we will be using is backward stepwise logistic regression which has the highest recall, paired with a considerably high accuracy.

4.7 Extracting Insights from models

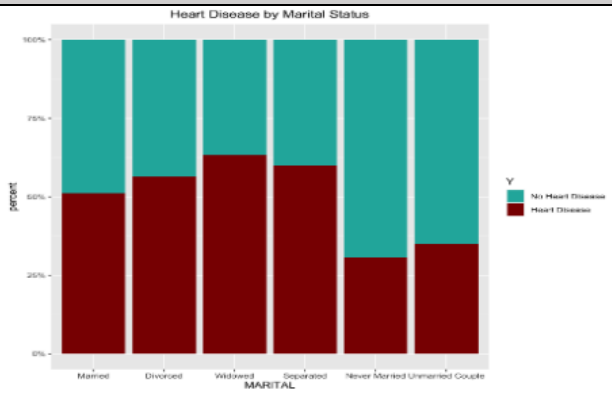
We will be drawing insights based on our best performing logistic regression model in determining the variables' significance in relation to Y. The model showed high significance scores for various variables such as age, smoking behaviour and income. However, we will not be diving into factors like age and weight, as they are well-known to be influential in detecting heart diseases. Instead, we will be analysing uncommon factors, specifically marital status and income.

MULTICOLLINEARITY	MARITAL	INCOME

`> vif(logreg)`
`GVIF Df GVIF^(1/(2*Df))`
`WTKG3 1.464980 1 1.210364`
`HTM4 1.511435 1 1.229404`
`AGEG5YR 3.670640 12 1.055677`
`EMPLOY1 3.113649 7 1.084510`
`GRENDAY 1.041151 1 1.020368`
`MARITAL 1.925540 5 1.067715`
`INCOME2 1.836777 7 1.044386`
`PA1VIGM 1.254109 1 1.119870`
`CHECKUP1 1.057691 4 1.007036`
`MENTHLTH 1.178218 1 1.085457`
`SMOKER3 1.195812 3 1.030253`
	Chi-square test of categorical association
	Variables: Y, MARITAL
	Hypotheses:
	null: variables are independent of one another
	alternative: some contingency exists between variables
	Observed contingency table:
	MARITAL
	Y 1 2 3 4 5 6
	0 5231 1320 1679 153 1519 224
	1 5469 1712 2888 229 670 120
	Expected contingency table under the null hypothesis:
	MARITAL
	Y 1 2 3 4 5 6
	0 5107 1447 2180 182 1045 164
	1 5593 1585 2387 200 1144 180
	Test results:
	X-squared statistic: 709.708
	degrees of freedom: 5
	p-value: <.001
	Other information:
	estimated effect size (Cramer's v): 0.183
		Chi-square test of categorical association
		Variables: Y, INCOME2
		Hypotheses:
		null: variables are independent of one another
		alternative: some contingency exists between variables
		Observed contingency table:
		INCOME2
		Y 1 2 3 4 5 6 7 8
		0 660 643 821 974 1181 1517 1516 2814
		1 805 1033 1270 1400 1531 1755 1518 1776
		Expected contingency table under the null hypothesis:
		INCOME2
		Y 1 2 3 4 5 6 7 8
		0 699 800 998 1133 1295 1562 1448 2191
		1 766 876 1093 1241 1417 1710 1586 2399
		Test results:
		X-squared statistic: 532.651
		degrees of freedom: 7
		p-value: <.001
		Other information:
		estimated effect size (Cramer's v): 0.158

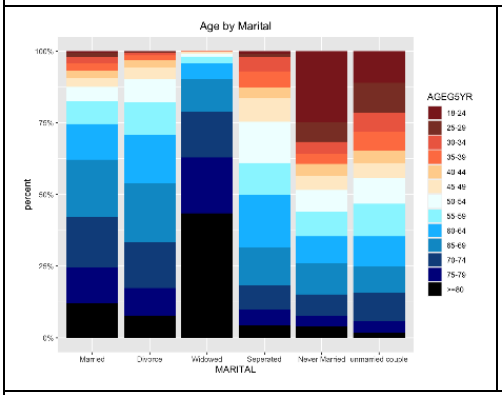
Multicollinearity does not exist among variables. Marital status and Income are independent predictors of heart disease.	Usage of chi-square test to prove that dependency exist between marital and heart disease as the p-value<0.001.	Usage of chi-square test to prove that dependency exist between income and heart disease as the p-value<0.001.

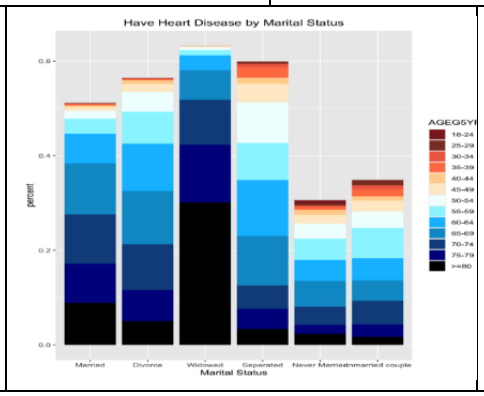
MARITAL

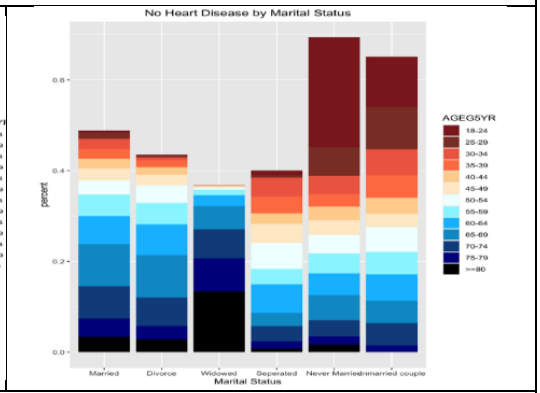


There is a higher percentage of those in “divorced", "married” and “Separated” who contracted heart disease as compared to other groups. Thus, we would like to deep dive to find the underlying factors behind Marital Status that results in one getting heart disease.

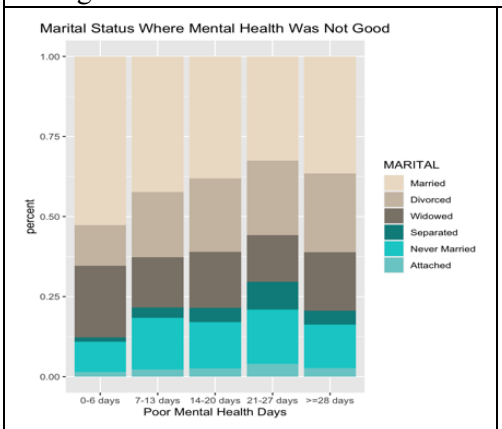
By using CART to model Marital against the other variables, we found that age is the top factor in determining marital status.

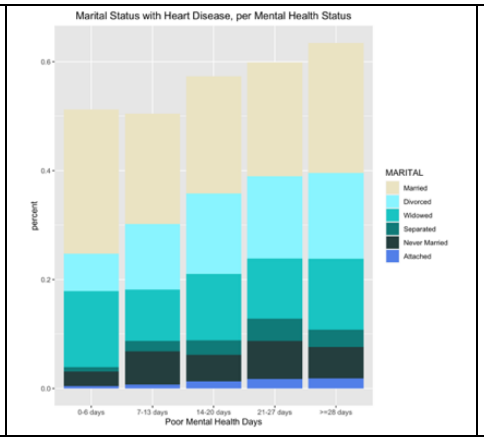


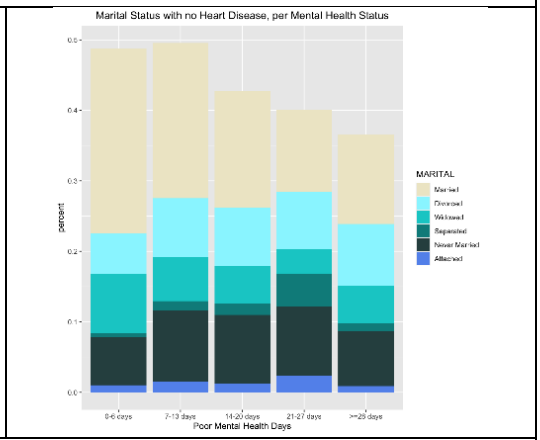




Respondents under 40 tend to fall under “non-married” and “unmarried couple”, while the seniors are equally distributed across the different groups. However, the widowers have the largest percentage of seniors, and this could explain why they also have the higher proportion of heart disease. These charts imply that the age factor in Marital Status could be the underlying factor in prediction of heart disease. Thus, we cannot conclude with a high degree of certainty that marital status by itself is a strong determinant of heart disease.



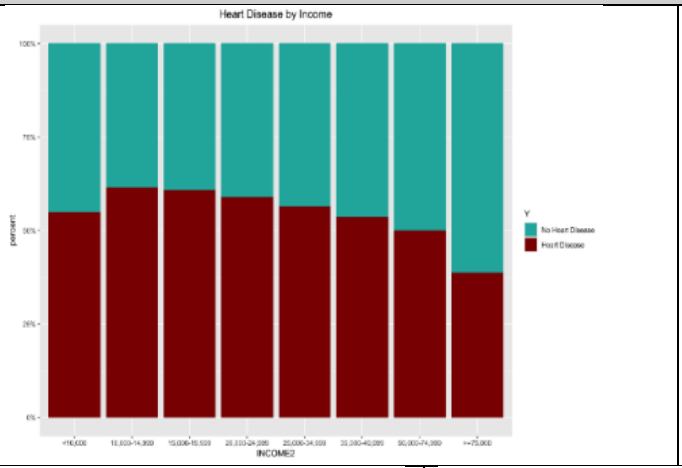




AHA (2022) mentioned the relatedness of marital status to heart diseases could be due to a syndrome named “broken heart”. It can be caused by emotionally stressful events resulting from divorced or breakups. Hence, we choose to investigate the relation between mental health, marital status, and heart disease even though the variable has a low variable significance.

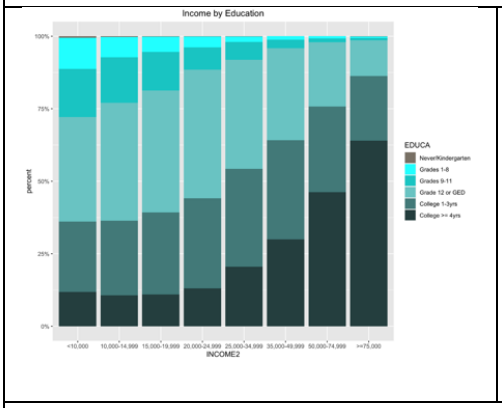
Indeed, we observe that as number of poor mental health days decreased, ratio of married to divorced and separated increases, showing a correlation between marriage and mental health where married couples generally have better mental health than divorcees and separated couples. Since we also observe that poor mental health is correlated with CHD risk, marriage could be a key contributing factor to CHD due to its influence on mental health.

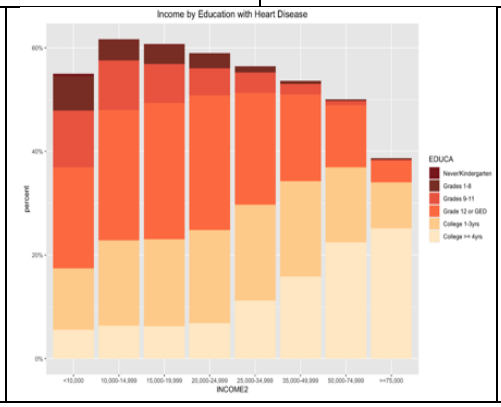
INCOME

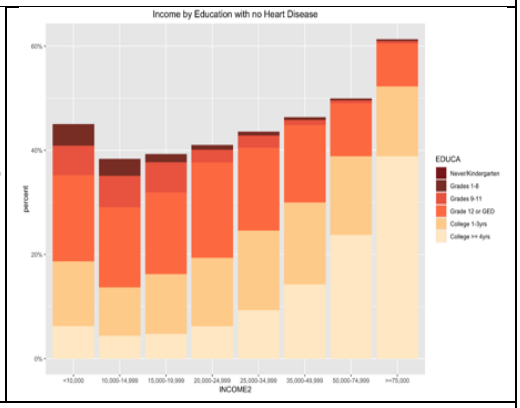


As income increases, the chances of one contracting heart diseases falls. Therefore, we would like to investigate the underlying factors behind income group that results in one getting heart disease.

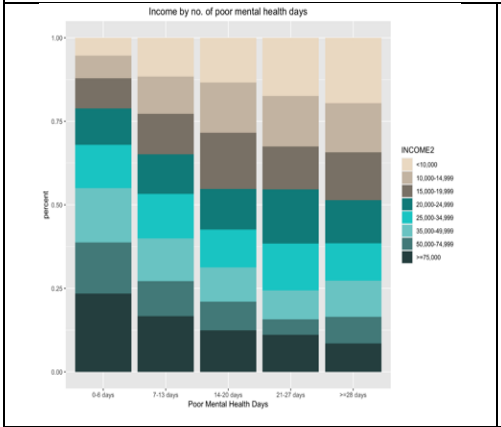
Once again, using CART to model income against other variables, education was found to be the most significant factor in determining income group

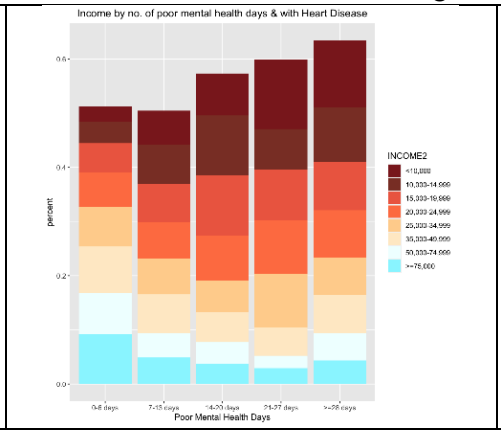


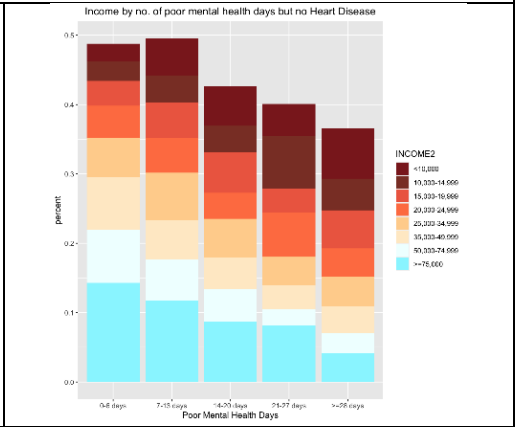




Income can be seen to be positively correlated to the level of education one attains. This trend is observed to be replicated among subset of groups of those with and without heart disease. Thus, this implies that if one has a higher level of education, they will have a higher income bracket which in turn, results in lower risk of getting heart disease.







Although deemed to be of low variable significance by the CART model, low income and work stress is researched to be the link between education and heart disease (ECS, 2019). Hence, there is a need for us to analyse and determine if mental health is the underlying factor, that is causing one to have heart disease.

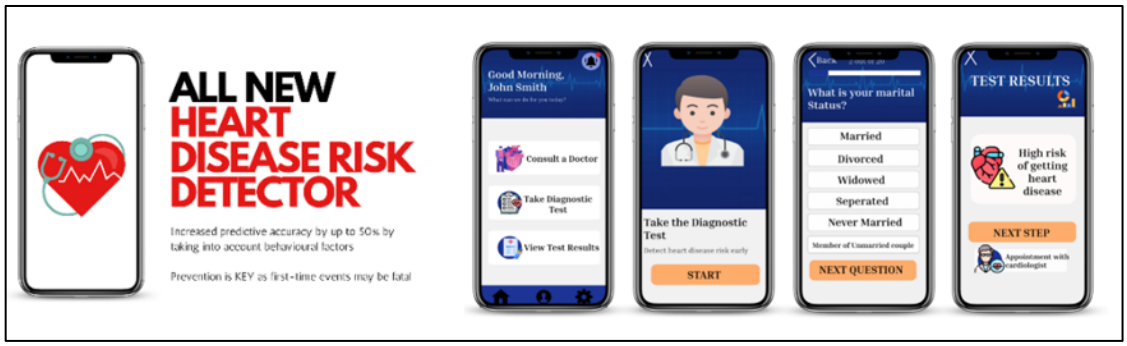
As income increases, the number of poor mental health days drop. The number of poor mental health days is also positively related to one getting heart disease as when the poor mental health days increase, the percentage of respondents getting heart disease is also higher. Hence, income is a good predictor of heart disease as it affects our mental health which in turn affects the risk of heart disease.

In summary, though marital status and income are assigned a high degree of variable importance by our primary logistic regression model, further analysis shows that their significance is strongly tied to mental health. In fact, the data shows that mental health is the key underlying consequence of marital status and income.

Mental health directly affects the risk of one getting heart disease. Many situations or factors contribute to our mental health, and one should not take lightly of it. Hence, these insights can be drawn on by doctors to explain to individuals the importance of placing emphasis on mental health and how it could have been a determining factor should they have been predicted to have a risk of heart disease.

5. Proposed Business Solution

Our proposed business solution is to implement a model that helps the NHCS make well-informed and timely decisions. It will act as a tool to detect patients, that may be at a high risk of heart disease, at an early stage based on non-medical factors via an app. This tool will not replace the existing heart disease tests, like X-rays, CT scans, MRI scans, echocardiograms and electrocardiograms; instead, it would supplement the existing heart disease tests in place. (Mayo Clinic, 2022)



Heart Disease Risk-Detector App

Step 1: Patient provides non-medical data on a survey

Patients can first provide non-medical data through a survey on the NHCS app. These non-medical factors will predominantly be those identified as important factors in the logistic regression model as stated above.

Step 2: Data is run through the regression model

Subsequently, these data will be run through a pre-trained model to predict whether an individual is likely to get heart disease based on the data given by them prior to taking any heart disease tests.

Step 3: Results of the survey

Based on the model, if the likelihood of the patient having heart disease is high, the survey will indicate that the patient is at a high risk of heart disease, along with the risk factors that contributed to this verdict. The patient can then consult a doctor for advice on the precautionary measures that must be taken to reduce the risk of heart disease. This doctor will obtain the details of the model prediction and data collected via the app. Moreover, the doctor may initiate a full body check up and take the necessary medical tests if the patient is in a very high risk of getting heart disease.

Step 4: Patients taking the necessary steps

Even if a patient is predicted to have a low likelihood of heart disease, they will be warned on the important behavioural factors identified by the model such as smoking that increases the risk of them developing heart disease in the future. Hence, the patients can take the necessary steps to help ensure that they mitigate any unhealthy habits that increase their risk in the future, such as reducing smoking.

5.1 Value proposition over existing models

Efficiency and Cost effectiveness

Current early detection models rely on medical data such as blood pressure and cholesterol levels requiring individuals to undergo medical health screening. (Mercy Health, n.d.) This is time consuming and incurs cost for the individuals who, require early detection to seek timely treatment in case they, have a heart disease. Since our model that we have proposed allows individuals to do a self-check, through participating in a survey, individuals can gauge whether they are at a high or low risk of developing heart disease with the help of the survey results and be advised accordingly. This would be time saving and cost effective as it does not necessitate individuals to consult medical practitioners. Only when the result of the survey indicates that the individual has a high risk of heart disease, will they be advised to consult a doctor and seek advice on the precautionary measures that should be taken.

Comprehensible and User-friendly

Furthermore, current early detection models require the results to be analysed by medical professionals due to the extensive use of medical jargons in the results. This may not be comprehensible to individuals who lack sufficient knowledge in the medical domain. This is in contrast with the survey which allows individuals to easily understand on why they have a higher risk of heart disease and their behaviour that they must be mindful of if they want to reduce the risk of developing a heart disease.

Alleviates stress on NHCS

Currently, cardiac diagnostic tests are all done by medical professionals. Individuals may frequently visit the doctors as they are unable to self-gauge their risk of developing heart disease. Such unnecessary consultations are both time consuming and stressful for the doctors. With our solution, individuals will be able to conduct a self-assessment on the likelihood of them developing a heart disease through the survey. Thus, individuals would only have to consult the doctor if they are at a greater risk of heart disease, alleviating the stress on medical practitioners and NHCS.

6. Conclusion

In conclusion, the behavioural factors of individuals were used to analyse the likelihood of them developing a heart disease using 5 different models: CART, GBM, logistic regression, neural network and SVM. The logistic regression model, with the highest recall value, chosen as the best model will aid individuals to gauge whether they are at higher risk of heart disease through the survey. Our proposed solution is accessible to all individuals as it is comprehensible, user friendly and cost effective. Ultimately, individuals can make timely and well-informed decisions, consulting the doctor when necessary, alleviating the stress on NHCS.

References

4 common health conditions in Singapore that you should protect yourself against. Protect Yourself: 4 Common Health Issues in Singapore | Life Matters. (n.d.). Retrieved October 29, 2022, from https://www.aia.com.sg/en/life-matters/health/4-common-health-issues-singapore.html

Advisory guidelines for healthcare sector - PDPC. (n.d.). Retrieved October 29, 2022, from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Sector-Specific-Advisory/advisoryguidelinesforthehealthcaresector28mar2017.pdf

Auto, H. (2022, September 28). *S'pore's population ageing rapidly: Nearly 1 in 5 citizens is 65 years and older*. The Straits Times. Retrieved October 29, 2022, from https://www.straitstimes.com/singapore/singapores-population-ageing-rapidly-184-of-citizens-are-65-years-and-older

Auto, H. (2021, November 18). *People in Singapore less healthy; COVID-19 may worsen situation: National Health Survey*. The Straits Times. Retrieved October 29, 2022, from https://www.straitstimes.com/singapore/health/people-in-singapore-less-healthy-and-covid-19-may-worsen-situation-national

Bhandari, A. (2020, April 3). *Feature scaling: Standardization vs normalization*. Analytics Vidhya. Retrieved October 29, 2022, from https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Brownlee, J. (2020, January 14). *Why is imbalanced classification difficult?* Machine Learning Mastery. Retrieved October 28, 2022, from https://machinelearningmastery.com/imbalanced-classification-is-hard/

Brownlee, J. (2020, February 10). *A gentle introduction to threshold-moving for Imbalanced Classification*. Machine Learning Mastery. Retrieved October 29, 2022, from https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

Boosting duration, intensity & frequency of physical activity may lower heart failure risk. American Heart Association. (n.d.). Retrieved October 29, 2022, from https://newsroom.heart.org/news/boosting-duration-intensity-frequency-of-physical-activity-may-lower-heart-failure-risk

Can emotional heartbreak impact heart health? SCL Health. (n.d.). Retrieved October 29, 2022, from https://www.sclhealth.org/blog/2022/02/can-emotional-heartbreak-impact-heart-health/

Centers for Disease Control and Prevention. (2022, August 29). *CDC - BRFSS*. Centers for Disease Control and Prevention. Retrieved October 29, 2022, from https://www.cdc.gov/brfss/index.html

Centers for Disease Control and Prevention. (2020, May 6). *Heart disease and mental health disorders*. Centers for Disease Control and Prevention. Retrieved October 29, 2022, from https://www.cdc.gov/heartdisease/mentalhealth.htm

Clare Oliver-Williams Junior Research Fellow. (2022, September 13). Having children is linked to increased risk of heart disease, new study suggests – but don't let that put you off. The Conversation. Retrieved October 29, 2022, from https://theconversation.com/having-children-is-linked-to-increased-risk-of-heart-disease-new-study-suggests-but-dont-let-that-put-you-off-105230

'Every second counts' in heart attacks. SBS News. (n.d.). Retrieved October 29, 2022, from https://www.sbs.com.au/news/article/every-second-counts-in-heart-attacks/wdzf1k24x

Heart health screenings. Mercy Health. (n.d.). Retrieved October 29, 2022, from https://www.mercy.com/health-care-services/heart-vascular/treatments/heart-health-screenings

How working too much affects your heart. Sunrise Hospital and Medical Center. (2020, February 17). Retrieved October 29, 2022, from https://sunrisehospital.com/blog/entry/how-working-too-much-affects-your-heart#:~:text=Another%20study%20published%20in%202015,hours%20on%20a%20weekly%20basis

Is broken heart syndrome real? www.heart.org. (2022, June 2). Retrieved October 29, 2022, from https://www.heart.org/en/health-topics/cardiomyopathy/what-is-cardiomyopathy-in-adults/is-broken-heart-syndrome-real

Jitender Sareen, M. D. (2011, April 4). Relationship between household income and mental disorders: Findings from a population-based longitudinal study. Archives of General Psychiatry. Retrieved October 29, 2022, from https://jamanetwork.com/journals/jamapsychiatry/fullarticle/211213

Joshipura KJ;Hu FB;Manson JE;Stampfer MJ;Rimm EB;Speizer FE;Colditz G;Ascherio A;Rosner B;Spiegelman D;Willett WC; (n.d.). *The effect of fruit and vegetable intake on risk for coronary heart disease*. Annals of internal medicine. Retrieved October 29, 2022, from https://pubmed.ncbi.nlm.nih.gov/11412050/

Mayo Foundation for Medical Education and Research. (2022, August 25). Heart disease. Mayo Clinic. Retrieved October 29, 2022, from https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124

MediLexicon International. (n.d.). Lower income individuals have 50% higher risk of heart disease. Medical News Today. Retrieved October 29, 2022, from https://www.medicalnewstoday.com/articles/233456

Merschel, M. (2020, August 12). *More school, less heart disease? researchers keep finding evidence*. www.heart.org. Retrieved October 29, 2022, from https://www.heart.org/en/news/2020/08/12/more-school-less-heart-disease-researchers-keep-finding-evidence

Meyer, D. (2022, October 24). *MISC functions of the Department of Statistics, Probability Theory Group (formerly: E1071), Tu Wien [R package E1071 version 1.7-12]*. The Comprehensive R Archive Network. Retrieved October 29, 2022, from https://cran.r-project.org/web/packages/e1071/

Obesity. BHF. (n.d.). Retrieved October 29, 2022, from https://www.bhf.org.uk/informationsupport/risk-factors/obesity#:~:text=How%20does%20obesity%20increase%20the%20risk%20of%20heart%20and%20circulatory%20diseases%3F&text=Excess%20weight%20can%20lead%20to,lead%20to%20a%20heart%20attack.

Prabhakaran, S. (n.d.). *If(typeof ez_ad_units != 'undefined'){ez_ad_units.push([[728,90],'r_statistics_co-box-3','ezslot_4',109,'0','0'])};__ez_fad_position('div-GPT-ad-r_statistics_co-box-3-0');outlier treatment*. Outlier Treatment With R | Multivariate Outliers. Retrieved October 29, 2022, from http://r-statistics.co/Outlier-Treatment-With-R.html

Raj, R. (n.d.). *Need of feature scaling in machine learning*. enjoyalgorithms. Retrieved October 29, 2022, from https://www.enjoyalgorithms.com/blog/need-of-feature-scaling-in-machine-learning/

Rettner, R. (2011, September 26). *Childlessness may increase men's heart disease risk*. Scientific American. Retrieved October 29, 2022, from https://www.scientificamerican.com/article/childlessness-may-increase/

Rupnarain, K. (2020, July 13). Why do the poor have large families? World Vision Canada. Retrieved October 29, 2022, from https://www.worldvision.ca/stories/why-do-the-poor-have-large-families

Shin, H.-Y., Kang, H.-T., Lee, J. W., & Lim, H.-J. (2018, March). The association between socioeconomic status and adherence to health check-up in Korean adults, based on the 2010-2012 Korean National Health and Nutrition Examination Survey. Korean journal of family medicine. Retrieved October 29, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876046/

Singh, S. (2021, July 27). *An emphasis on the minimization of false negatives/false positives in binary classification*. Medium. Retrieved October 29, 2022, from https://medium.com/@Sanskriti.Singh/an-emphasis-on-the-minimization-of-false-negatives-false-positives-in-binary-classification-9c22f3f9f73#:~:text=To%20minimize%20the%20number%20of,optimally%20reaches%20a%20global%20minimum.

Smoking and cardiovascular disease. Smoking and Cardiovascular Disease | Johns Hopkins Medicine. (2020, July 20). Retrieved October 29, 2022, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease

Team, H. and V. (2022, March 11). *Heart risks associated with extreme exercise*. Cleveland Clinic. Retrieved October 29, 2022, from https://health.clevelandclinic.org/can-too-much-extreme-exercise-damage-your-heart/

The benefits of beans and legumes. www.heart.org. (2022, July 22). Retrieved October 29, 2022, from https://www.heart.org/en/healthy-living/healthy-eating/eat-smart/nutrition-basics/the-benefits-of-beans-and-legumes#:~:text=Beans%20are%20high%20in%20minerals,keep%20you%20feeling%20full%20longer.

Thenraj, P. (2020, July 22). *Do decision trees need feature scaling?* Retrieved October 29, 2022, from https://towardsdatascience.com/do-decision-trees-need-feature-scaling-97809eaa60c6

U.S. Department of Health and Human Services. (n.d.). *Heart health and aging*. National Institute on Aging. Retrieved October 29, 2022, from https://www.nia.nih.gov/health/heart-health-and-aging#:~:text=Adults%20age%2065%20and%20older,risk%20of%20developing%20cardiovascular%20disease

Vanessa Lim @VanessaLimCNA, Lim, V., & Bookmark Bookmark Share WhatsApp Telegram Face. (n.d.). *Longer waiting times at hospitals with some patients told to wait up to 50 hours for a bed*. CNA. Retrieved October 29, 2022, from https://www.channelnewsasia.com/singapore/hospitals-beds-waiting-time-50-hours-admission-ng-teng-fong-sengkang-3014596

WebMD. (2012, November 19). Unemployment takes toll on the heart. WebMD. Retrieved October 29, 2022, from https://www.webmd.com/heart/news/20121119/unemployment-toll-heart

World Health Organization. (n.d.). *Cardiovascular diseases (cvds)*. World Health Organization. Retrieved October 29, 2022, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

World Health Organization. (n.d.). *Tobacco responsible for 20% of deaths from coronary heart disease*. World Health Organization. Retrieved October 29, 2022, from https://www.who.int/news/item/22-09-2020-tobacco-responsible-for-20-of-deaths-from-coronary-heart-disease

Williams, D. N. (2021, November 17). How does education affect mental health? News. Retrieved October 29, 2022, from https://www.news-medical.net/health/How-does-Education-Affect-Mental-Health.aspx#:~:text=and%20mental%20health-,Higher%20levels%20of%20education%20have%20been%20associated%20with%20better%20mental,earn%20more%20throughout%20their%20lifetimes.

Wilimitis, D. (2018, December 12). *The kernel trick in support vector classification*. Towards Data Science. Retrieved October 29, 2022, from https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Appendices

Behavioral Risk Factor Surveillance System (BRFSS) 2015 Codebook Report:

Variable Name	Description
MENTHLTH	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? 1 – 30: Number of days 88: None 77: Don't know/Not sure 99: Refused
HLTHPLN1	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? 1: Yes 2: No 7: Don't know/ Not sure 9: Refused
MEDCOST	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 1: Yes 2: No 7: Don't know/ Not sure 9: Refused BLANK: Not asked or Missing
CHECKUP1	About how long has it been since you last visited a doctor for a routine checkup? *[A routine checkup is a general* *physical exam, not an exam for a specific injury, illness, or condition.]* 1: Within past year (anytime less than 12 months ago) 2: Within past 2 years (1 year but less than 2 years ago) 3: Within past 5 years (2 years but less than 5 years ago) 4: 5 or more years ago 7: Don't know/Not sure 8: Never 9: Refused BLANK: Not asked or Missing

CVDCRHD4	(Ever told) you had angina or coronary heart disease? 1: Yes 2: No 7: Don't know/ Not sure 9: Refused BLANK: Not asked or Missing
ADDEPEV2	(Ever told) you that you have a depressive disorder, including depression, major depression, dysthymia, or minor depression? 1: Yes 2: No 7: Don't know/ Not sure 9: Refused
MARITAL	Are you: (marital status) 1: Married 2: Divorced 3: Widowed 4: Separated 5: Never married 6: A member of an unmarried couple 9: Refused
EDUCA	What is the highest grade or year of school you completed? 1: Never attended school or only kindergarten 2: Grades 1 through 8 (Elementary) 3: Grades 9 through 11 (Some high school) 4: Grade 12 or GED (High school graduate) 5: College 1 year to 3 years (Some college or technical school) 6: College 4 years or more (College graduate) 9: Refused
EMPLOY1	Are you currently…? 1: Employed for wages 2: Self-employed 3: Out of work for 1 year or more 4: Out of work for less than 1 year 5: A homemaker 6: A student 7: Retired 8: Unable to work 9: Refused

CHILDREN	How many children less than 18 years of age live in your household? 1 – 87: Number of children 88: None 99: Refused BLANK: Not asked or missing
INCOME2	Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") 1: Less than $10,000 Notes: If "no," code 02 2: Less than $15,000 ($10,000 to less than $15,000) Notes: If "no," code 03; if "yes," ask 01 3: Less than $20,000 ($15,000 to less than $20,000) Notes: If "no," code 04; if "yes," ask 02 4: Less than $25,000 ($20,000 to less than $25,000) Notes: If "no," ask 05; if "yes," ask 03 5: Less than $35,000 ($25,000 to less than $35,000) Notes: If "no," ask 06 6: Less than $50,000 ($35,000 to less than $50,000) Notes: If "no," ask 07 7: Less than $75,000 ($50,000 to less than $75,000) Notes: If "no," code 08 8: $75,000 or more 77: Don’t know/ Not sure 99: Refused BLANK: Not asked or Missing
INTERNET	Have you used the internet in the past 30 days? 1: Yes 2: No 7: Don’t know/ Not sure 9: Refused BLANK: Not asked or Missing
SMOKER3	Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker 1: Current smoker - now smokes every day 2: Current smoker - now smokes some days 3: Former smoker 4: Never smoked 9: Don’t know/Refused/Missing
HTM4	Reported height in metres

	91 – 244: Height in meters [2 implied decimal places] BLANK: Not asked or Missing
WTKG3	Reported weight in kilograms 2300 – 29500: Weight in kilograms [2 implied decimal places] 99999: Don't know/Refused/Missing
AGEG5YR	Fourteen-level age category 1: Age 18 to 24 2: Age 25 to 29 3: Age 30 to 34 4: Age 35 to 39 5: Age 40 to 44 6: Age 45 to 49 7: Age 50 to 54 8: Age 55 to 59 9: Age 60 to 64 10: Age 65 to 69 11: Age 70 to 74 12: Age 75 to 79 13: Age 80 or order 14: Don't know/Refused/Missing
SEATBELT	How often do you use seat belts when you drive or ride in a car? Would you say— 1: Always 2: Nearly always 3: Sometimes 4: Seldom 5: Never 7: Don't know/ Not sure 8: Never drive or ride in a car 9: Refused BLANK: Not asked or Missing
SCNTWRK1	About how many hours do you work per week on all of your jobs and businesses combined? 1 – 96: Hours (1-96 or more) 97: Don't know/ Not sure 98: Zero (none) 99: Refused BLANK: Not asked or Missing

AVEDRNK2	One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. During the past 30 days, on the days when you drank, about how many drinks did you drink on the average? (A 40 ounce beer would count as 3 drinks, or a cocktail drink with 2 shots would count as 2 drinks.) 1 – 76: Number of drinks 77: Don't know/ Not sure 99: Refused BLANK: Not asked or Missing
ALCDAY5	During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor? 101 – 199: Days per week Notes: 1 _ _ = Days per week 201 – 299: Days in past 30 days Notes: 2_ _ = Days in past 30 777: Don't know/Not sure 888: No drinks in past 30 days 999: Refused BLANK: Not asked or Missing
FRUTDA1	Fruit intake in times per day 0 – 9999: Times per day (two implied decimal places) BLANK: Don't know/ Not Sure or Refused/Missing
BEANDAY	Bean intake in times per day 0 – 9999: Times per day (two implied decimal places) BLANK: Don't know/ Not Sure or Refused/Missing
GRENDAY	Dark green vegetable intake in times per day 0 – 9999: Times per day (two implied decimal places) BLANK: Don't know/ Not Sure or Refused/Missing
MICHD	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) 1: Reported having MI or CHD 2: Did not report having MI or CHD BLANK: Not asked or Missing
PA1MIN	Minutes of total Physical Activity per week 0 – 99999: Minutes of Activity per week

	BLANK: Not asked or Missing
PA1VIGM	Minutes of total Vigorous Physical Activity per week 0 – 99999: Minutes of Activity per week BLANK: Not asked or Missing
SEX	Indicate sex of respondent. 1: Male 2: Female