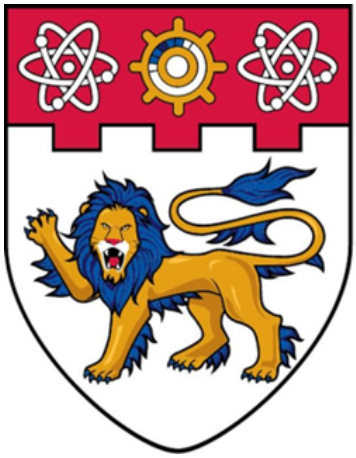

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Prediction of Heart Disease using Patient Data (Categorical)
Prediction of Stroke based on Heart Disease Data (Regression)

Team 7

Dylan Cheong De Wei
Ho Jia Jun Brandon
Julian Chung Zhong Wei
How Rui Yang
Wang Rui Xian

Executive Summary

Background

In this project, the cardiovascular disease and stroke datasets are used to generate a predictive model to detect cardiovascular disease and stroke in patients. Every feature in the datasets is analysed to determine its usefulness in the model generation.

Important Findings

Amongst the 14 features in the cardiovascular disease (CVD) dataset, the following 5 key features are deemed as the most significant - cholesterol levels, glucose levels, age, systolic and diastolic blood pressure. Using these most important data features, some interesting findings are: CVD tends to affect older people (specifically from age 55 and above), patients with CVD tend to have higher than average blood pressures, and cholesterol level is the biggest factor for predicting CVD (highest correlation). Amongst the 11 features in the stroke dataset, the following 5 key features are deemed as the most significant - hypertension, heart disease, marital status, smoking status and glucose level. Using these most important data features, some interesting findings are: people who are married tend to have higher chances of getting a stroke as compared to people who are not, patients with stroke tend to have hypertension and heart disease, and heart disease is the biggest factor for predicting stroke.

Business Impact

The prediction model developed can be used in two settings - Telehealth Application and Physical Visits. In these two settings, the model can be used to facilitate quicker and more accurate preliminary diagnoses. By doing so, the number of visits to NHCS can be reduced, allowing doctors to cater to cases that pose a real threat. This would not only increase the efficiency of consultations but also allow NHCS to accept more patients; both of which have an effect in helping the hospital generate more revenue. Additionally, the prediction model can be marketed as a commercial good in order to generate extra revenue from hospitals overseas.

Project Requirements

To develop our project further, the following are required:

1. **Data Collection in Singapore**
   To make the prediction model more catered towards the local market, data collection needs to be carried out to gain information on the local population.

2. **Development of Telehealth App for patients and Web App for doctors**
   This will be the interface for communication between our users and the prediction model.

Table of Content

1. Business Problems

Health data is usually collected during triage whenever a patient schedules a health checkup. Such data often contains a wealth of untapped information which can be used meaningfully in machine learning to predict diseases present in patients.

1.1 Prediction of Heart Disease using Patient Data (Categorical)

Cardiovascular diseases (CVD) remain as a top cause of death globally, taking an estimated 17.9 million lives annually [1]. Large investments into early detection systems have been popular in recent years due to the large increase in survival rate of the patient [1]. Early detections have also been shown to prevent attacks from CVD, reducing the need for dangerous emergency procedures which place the patient at risk [1]. However, it's not uncommon for CVD patients to pass on despite displaying experienced symptoms not taken into consideration by medical professionals [4]. Therefore, the potential for medical institutes to implement systems that can handle vast quantities of patient information and accurately profile high-risk patients exists.

Although medical institutions and practitioners generate a wealth of patient data, its effectiveness in terms of hidden but useful predictive data knows no bounds [3]. This report thus details an exploratory example of converting basic patient data into a cleaned dataset for useful categorical classification of CVD likelihood [5]. Common features that are of interest which may relate to the possibility of heart attacks include smoking, physical inactivity, high cholesterol or even over-usages of alcohol [5].

1.2 Prediction of Stroke based on Heart Disease Data (Regression)

It has been reported that more than 877,500 Americans suffer from heart disease, stroke, or other cardiovascular diseases every year [13]. Often, such diseases bear a correlation as the underlying cause is due to congested arteries. This means that detecting one such disease can potentially help in predicting another [13]. This report will thus attempt to use heart disease data to predict the possibility of a related disease - Stroke.

1.3 Business Value

User Flow

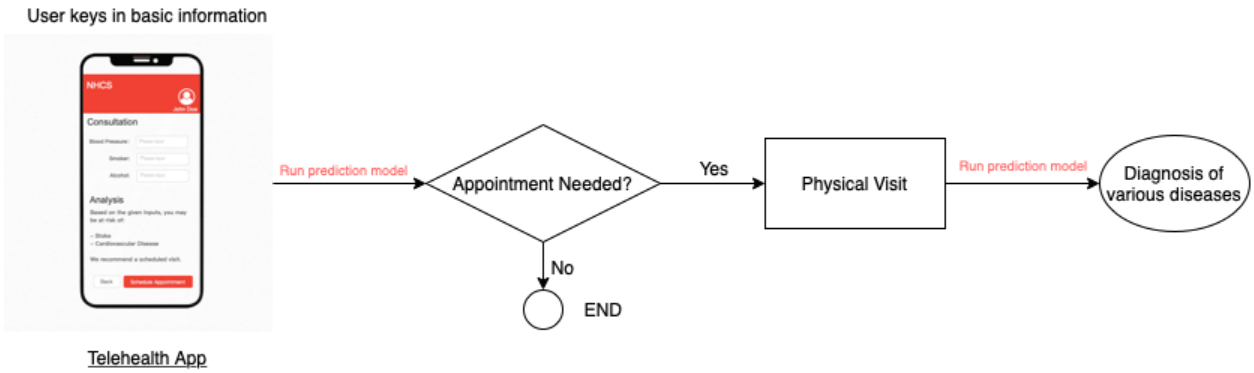

Figure 1: Proposed User Flow

1.3.1 New Revenue Line - Telehealth

With a sophisticated system to detect the presence of CVD and stroke, based on a few given inputs, a telehealth app can be set up to assess the risk of such diseases within patients. By doing so, a pre-check is done before patients visit the hospital. This would reduce the number of visits to the physical hospital and allow the hospital to only process "serious" cases. As a result, the hospital will not only free up capacity to see more patients but also get more patients with genuine concerns.

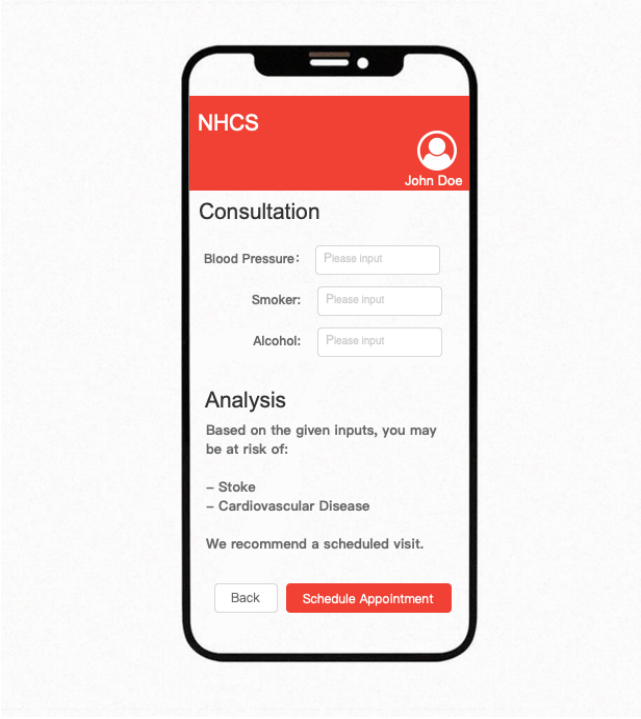

Figure 2: Telehealth User Interface

1.3.1 Increase Efficiency in Diagnosis

Using triage data to predict heart disease can serve to be a diagnosis aid for doctors. Through this, doctors can not only make quicker decisions but also more accurate ones. This will significantly reduce consultation times and open up treatment opportunities to more patients.

Additionally, we can also use heart disease data to predict the likelihood of related diseases like a stroke. This would allow the doctor to prescribe preventive maintenance which is beneficial towards preventing future visits from the patient. This not only improves patients' well-being but also aids them in reducing their healthcare expenditure. Using a stroke predictor would also help doctors get an overall understanding of the patient's cardiovascular health which enables them to make a more holistic diagnosis.

2. Motivations

Our team strongly believes that machine learning can both improve the prognosis accuracy and speed of CVD for patients around the world. This powerful tool can also do so consistently, making up for the liabilities in human error or fatigue in medical professionals across long sustained periods. Any diagnosis provided via machine learning can also act as guidance or serve as a reference for entrant medical specialists or even experienced doctors. The pronounced positive impact of this technology in the medical field is why our team strongly advocates for advances in its research and for testing in practical use.

Furthermore, a communal element also drives the team in this project. On average, heart disease is responsible for 32% of all deaths in Singapore [14]. This number is largely contributed by the elderly population within the country, which are also at higher risks of manifesting and succumbing to CVD. As the Singapore community is an ageing population, the team hopes that the project may contribute in some form to our greying society.

Lastly, our team was also motivated by the desire to innovate upon existing technologies in its application in new areas. Stroke was the chosen focus as the extension of the project as it holds the title of being the fourth leading cause of death in Singapore. Statistics also show that 4% of the elderly above the age of 50 are more prone to the illness [14], with it being a likely cause of long-term disability. This project aims to repurpose predictive machine learning models to refine any predictive system that can flag out CVD patients that are at high-risk of having a stroke. The team hopes the implementation of such systems can avoid unnecessary costs and inconvenience to medical institutes.

3. Heart Attack Prediction

3.1 Choice of Dataset

For the prediction of a CVD, we have 2 different datasets: a 70k dataset from Kaggle and a 920 dataset from UCI. However, there exist little similarities to merge the 2 datasets. Specifically, between the 14 features in the Kaggle dataset and the 15 features in the UCI dataset, only 3 features are consistent - age, gender, and cholesterol. Among which, the UCI dataset focuses on having continuous data while the Kaggle dataset focuses on having categorical data. For example, cholesterol is categorical data in Kaggle indicating how high it is while it is a continuous data in UCI indicating the exact cholesterol level. This makes it hard to effectively merge the datasets as there are little similarities and little basis of how "ranking" was done.

Thus, we have made the decision to focus on the larger dataset (Kaggle dataset) as this allows us to better split the dataset into a train-validation-test and also avoid possible biases or lack of corner cases a small dataset might have.

3.2 Data Exploration

3.2.1 Number of People with CVD Across Different Ages

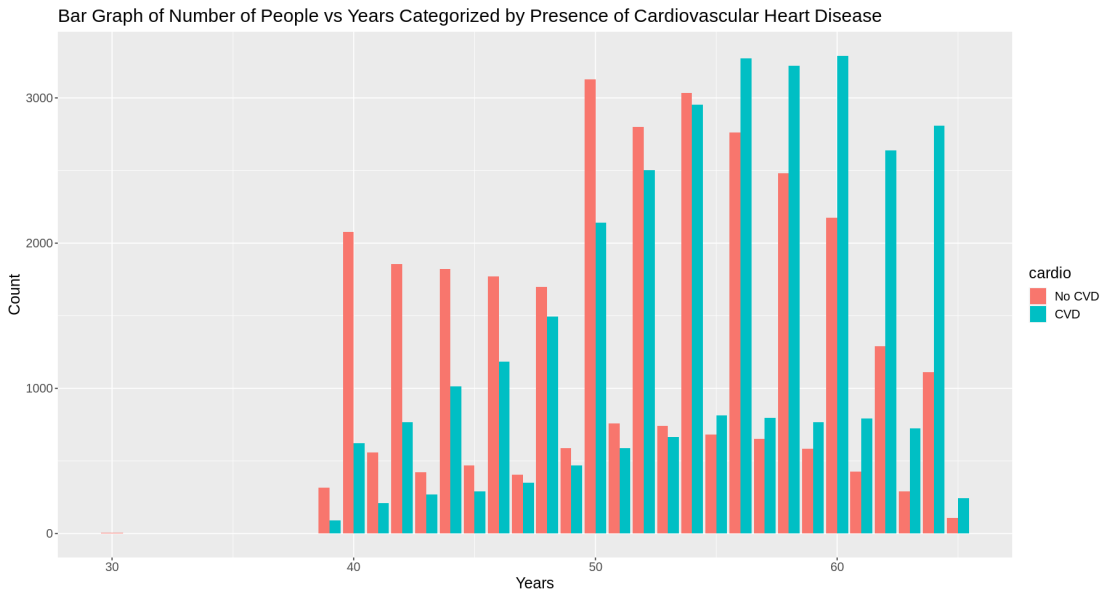

Figure 3: Number of People with CVD Across Different Ages

Based on Figure 3, it is observed that the majority of CVD patients come from an older age group between 55-70. From age 39 to 54, the number of CVD patients were fewer than those without CVD. However this phenomena reverses from age 55 onwards, and there were more

CVD patients than non-CVD patients. To refine our understanding, we performed additional analysis focusing on the trends between age and the number of CVD patients.

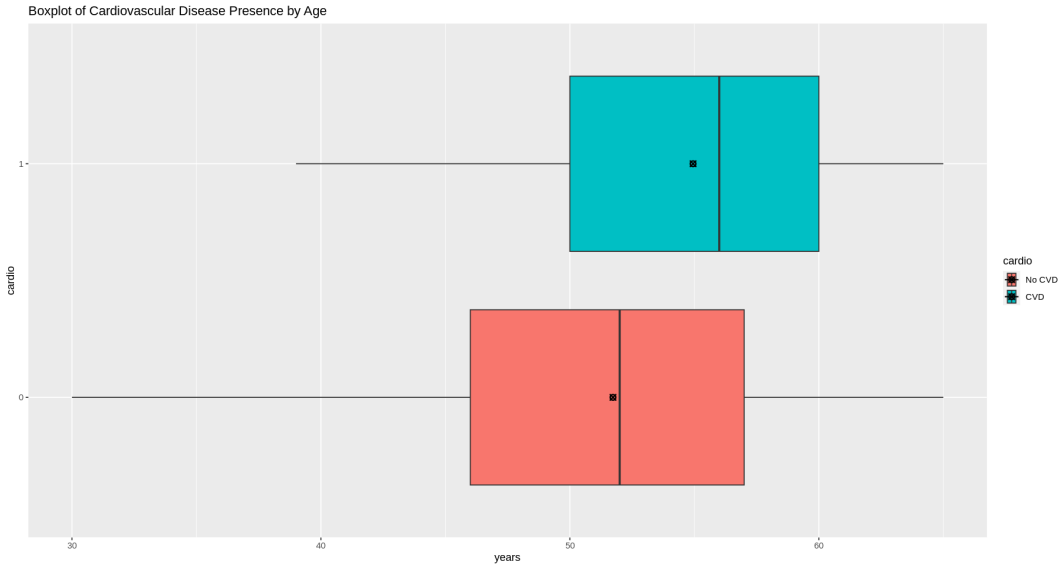

Figure 4: Boxplot of Number of People with CVD Across Different Ages

Upon further analysis, it was found that the median age of people with CVD was 56. This is higher than the median age of people without CVD at 52. Similarly, the average age of the population without CVD is approximately 51.73 while that of the population with CVD is 54.95. This suggests a strong trend between growing older and the likelihood of contracting CVD.

3.2.2 Number of People with CVD Across Different Blood Pressures

When measuring blood pressure, it is important to distinguish between systolic and diastolic blood pressures as different biomarkers as they may reveal different facets of a patient's health.

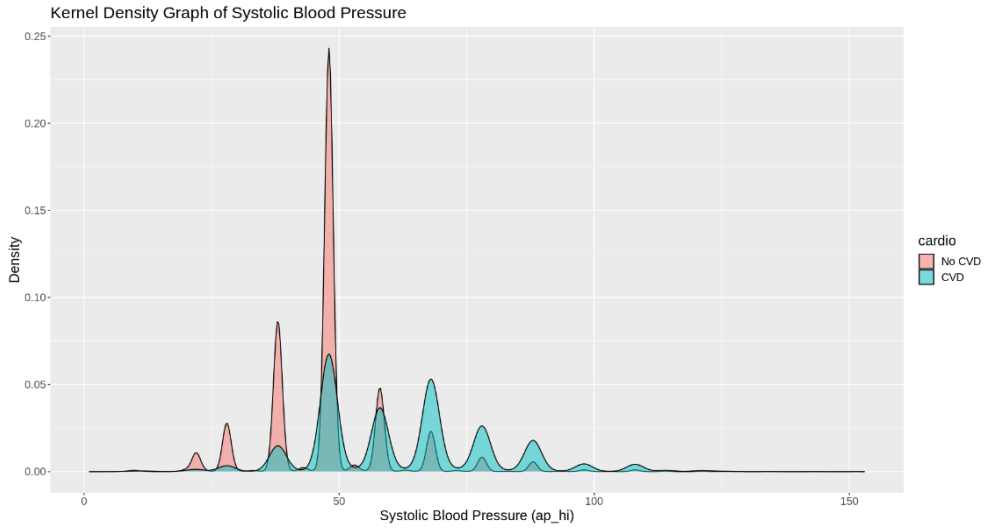

Figure 5: Kernel Density Graph of Systolic Blood Pressure

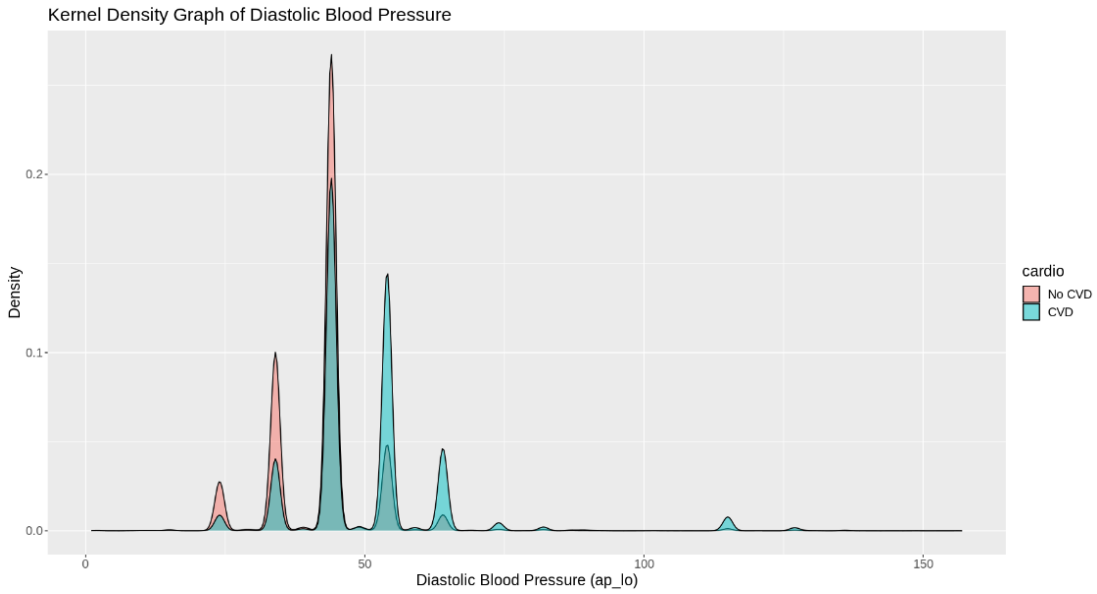

Figure 6: Kernel Density Graph of Diastolic Blood Pressure

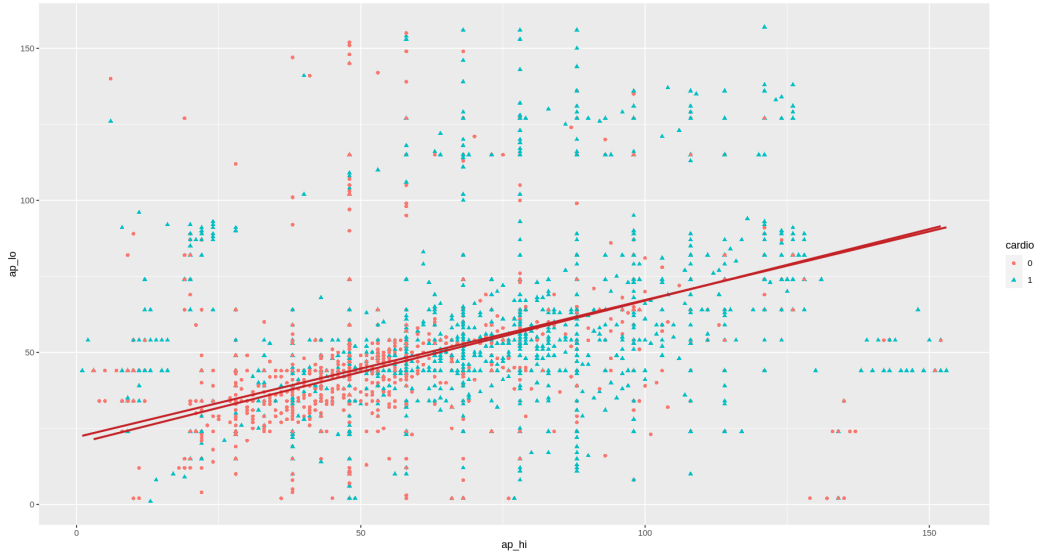

Figure 7: Scatter Plot Between Diastolic Blood Pressure (ap_lo) and Systolic Blood Pressure (ap_hi)

Figures 5 and 6 show that people without CVD have lower systolic and diastolic blood pressures compared to people with CVD. Upon further analysis, it was also found that 80.2% of the patients with CVD have higher than average blood pressures. Figure 7 also shows that there is a 62.8% correlation between systolic and diastolic blood pressures.

3.2.3 Other Categorical Factors Affecting CVD

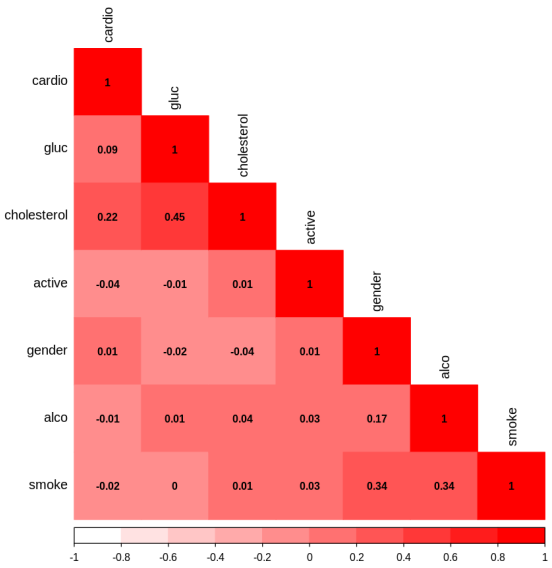

Figure 8: Correlation Matrix Between Categorical Factors and CVD

The phi coefficient is used to measure the association between the two variables. The phi coefficient is used when both variables are binary and ranges from -1 to 1. A value of 0 indicates no association between the variables, while a value of -1 or 1 indicates a perfect negative or positive association, respectively. Figure 8 shows that cholesterol has the highest phi correlation to CVD out of the other categorical factors.

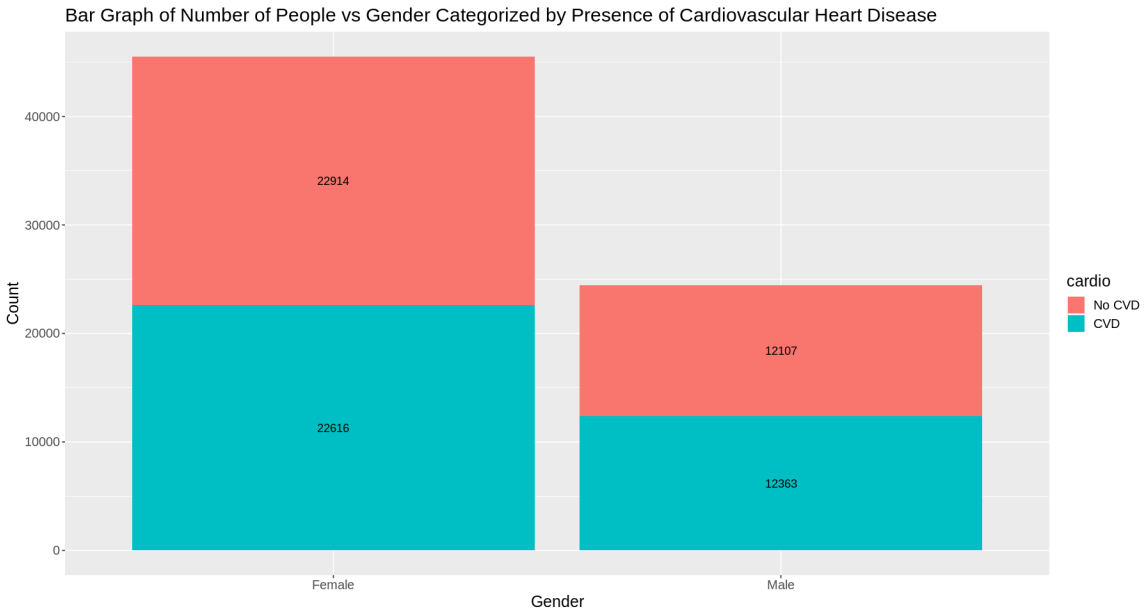

Figure 9: CVD Analysis by Gender

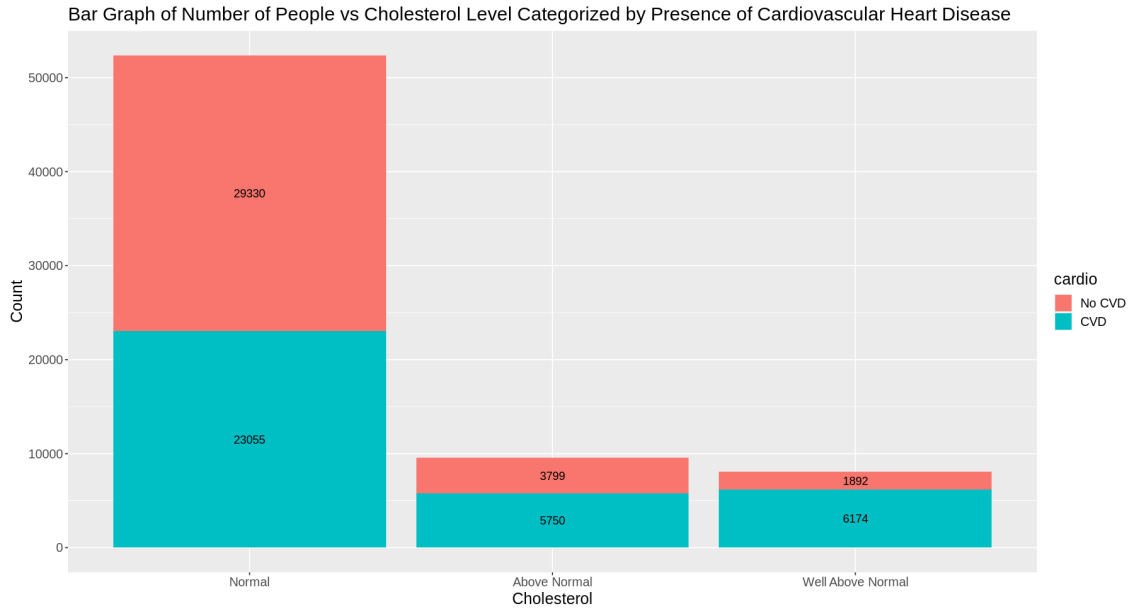

Figure 10: CVD Analysis by Cholesterol

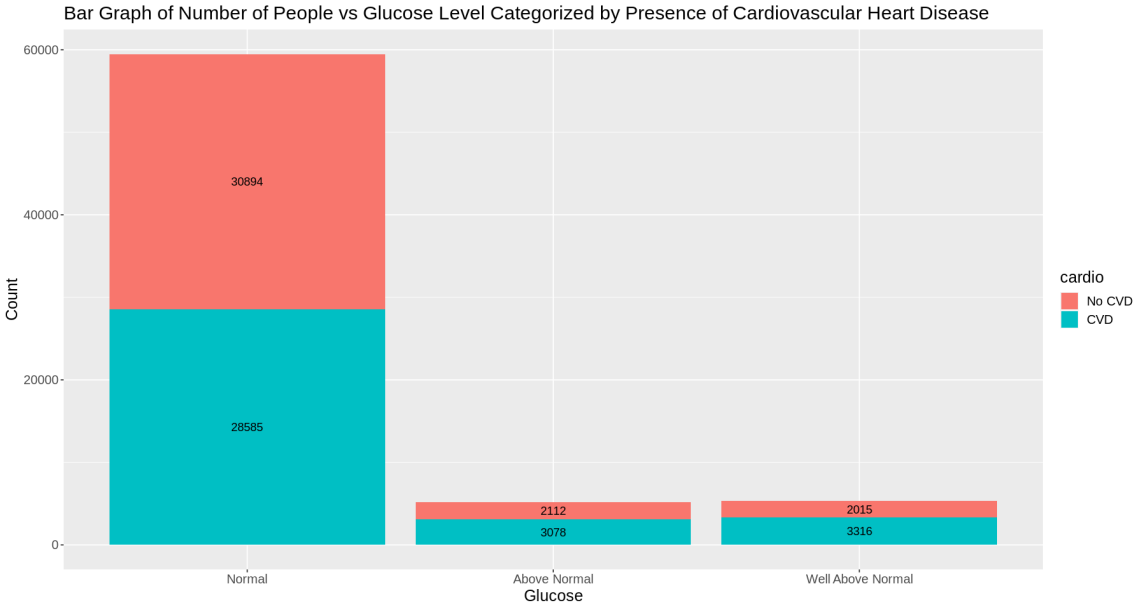

Figure 11: CVD Analysis by Glucose

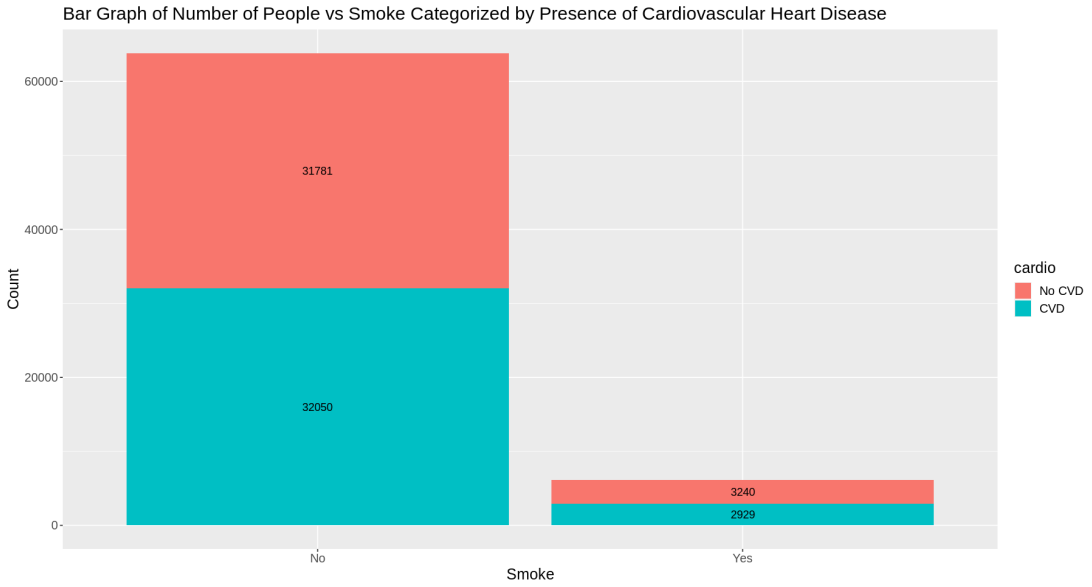

Figure 12: CVD Analysis by Smoke

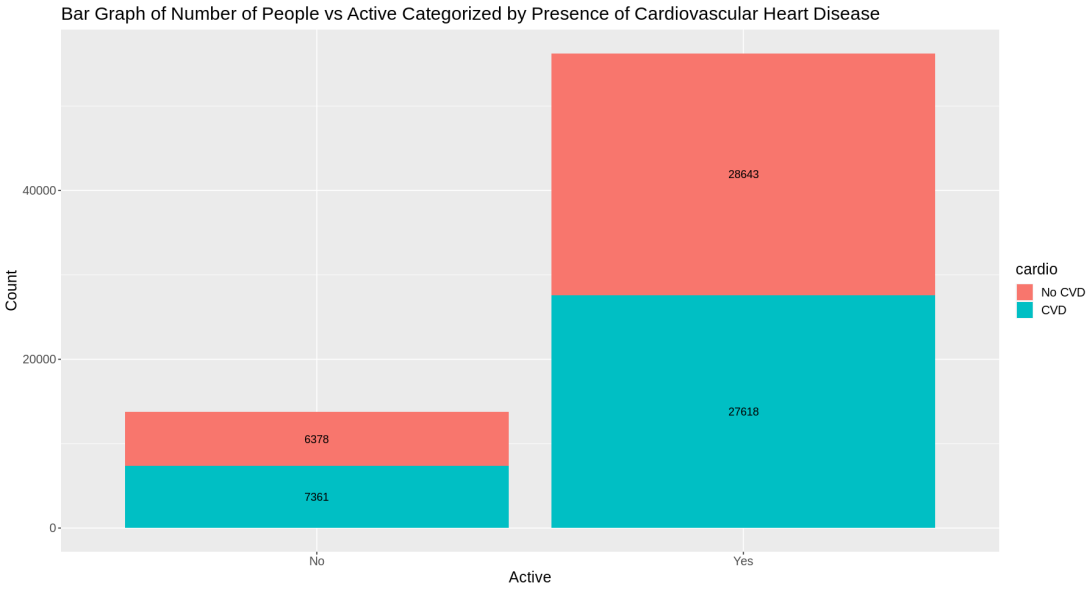

Figure 13: CVD Analysis by Active

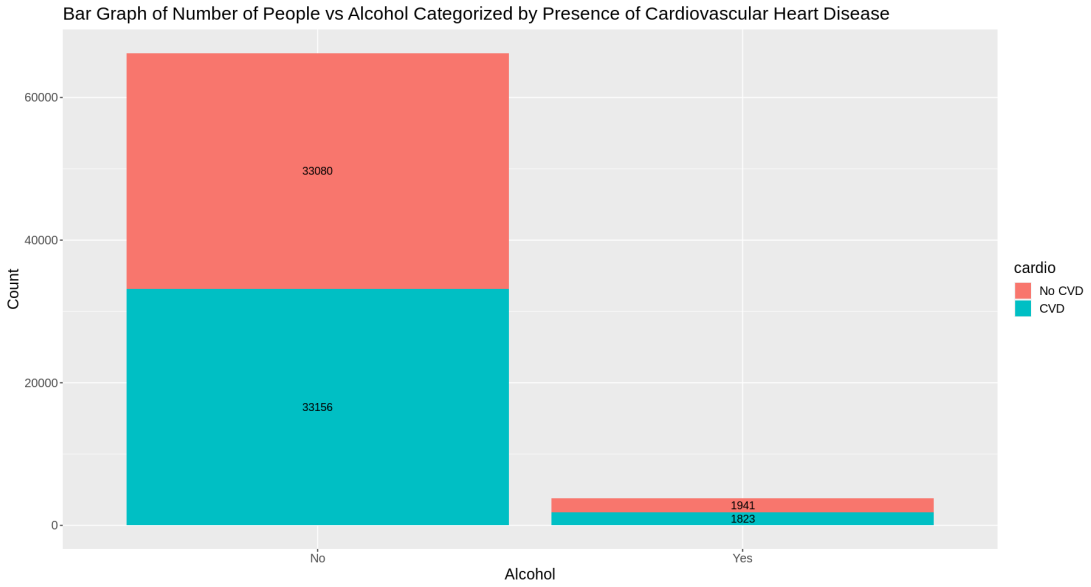

Figure 14: CVD Analysis by Alcohol

Figures 9 to 14 present the data of a few categorical factors. These factors are gender, cholesterol, glucose level, and whether the patient smokes, exercises or drinks alcohol. The following findings are made:

- The dataset has 45,530 females and 24,470 males. About 50% of each gender has CVD.
- Cholesterol level is a huge determinant of whether the patients have cardiovascular disease. It can be seen that there are 27.2% more healthy patients than CVD patients with normal cholesterol levels, while there are 51.4% more CVD patients than healthy ones with above normal cholesterol levels and 226.3% with well above normal cholesterol levels.
- Glucose level is also another huge determinant of whether patients have cardiovascular disease. It can be seen that there are 8.08% more healthy patients than CVD patients with normal glucose levels, while there are 45.7% more CVD patients than healthy ones with above normal glucose levels and 64.6% with well above normal glucose levels.
- There are more inactive patients with CVD compared to inactive patients without CVD (15.4%), while active patients without CVD are 3.71% more than active patients with CVD. This means that while exercising and maintaining an active lifestyle is good, it does not necessarily mean that you are less likely to have CVD.
- Smoking and alcohol consumption does not seem to have an adverse effect on whether the patients have CVD or not.

3.3 Predictive Model

Our team utilised two methods to generate predictive models for heart disease from patient data. For each method, we mainly focused on 3 different ways to generate the model - using the entire dataset, using a predetermined curated dataset, and using a generative lightest dataset.

3.3.1 Generalised Linear Model (GLM)

The GLM model was chosen as a model because of its computational efficiency, robustness and interpretability. The model enabled us to quickly understand each effect of predictor variables on the final output and was also able to handle our chosen dataset quickly.

Our efforts with GLM can largely be summarised into three trials:

Trial 1- A predictive model on the likelihood of heart attack with all patients data (72.8%)
Trial 2- A predictive model on the likelihood of heart attack with features chosen from EDA (72.62%)
Trial 3- A brute force method to determine the best predictive model on the likelihood of heart attack with the minimum combination of patient data (72.67%)

The results of all three trials allowed us to produce a predictive model with **72.8%** accuracy on the testset.

3.3.2 Logistic Regression

The logistic regression is also an apt model as it is commonly used to estimate the probability of a binary outcome (presence of a CVD in our case). Logistic regression is a type of GLM that is based on the logit function instead of the binomial function (used in the above model). Even though they are similar, the behaviour of the estimation methods can differ in corner cases and for singular types of data.

Our efforts with Logistic Regression can largely be summarised into two trials:

Trial 1 - A predictive model on the likelihood of heart attack with all patients data (72.79%)
Trial 2- A brute force method to determine the best predictive model on the likelihood of heart attack with the minimum combination of patient data (72.66%)

The results of the Logistic Regression models provide very similar results to the GLM models, largely because they both represent the same underlying model and fit by maximum likelihood estimation. The Logistic Regression models give us a predictive model with **72.79%** accuracy on the testset.

3.3.3 Classification And Regression Trees (CART)

CART is a commonly used tree-based model that is not only performant but also easily understandable. By breaking down the decision process into a tree, the insights are clear and self-explanatory across varying levels of technical background. In our trials, we will aim to establish a system of explaining the optimal decisions made by CART for practical diagnosis of patients.

Our efforts with CART can be largely be summarised into three iterative trials:

Trial 1 - A predictive model on the likelihood of heart attack with all patients data (71.5%)
Trial 2 - A predictive model on the likelihood of heart attack focusing on secondary health factors (62.7%)

With Trial 1, we included all relevant patient health data for prediction of CVD with CART, and we obtained a testset prediction accuracy of **71.5%**. In descending order of importance, systolic blood pressure has the highest importance (52%), followed by diastolic blood pressure (31%), and cholesterol (6%), and the only variable used in the construction of the decision tree is systolic blood pressure as seen in Figure B.1. This shows the importance of systolic blood pressure in the prediction of heart attack, but disregards other biomarkers which carry less weight. Thus, Trial 2 will explore the decision tree generated by CART after excluding these strong indicators.
For Trial 2, we exclude systolic and diastolic blood pressure from the training data, and we obtained a testset prediction accuracy of **62.7%**. In descending order of importance, cholesterol has the highest importance (53%), followed by age (37%), glucose levels (8%), weight (5%) and history of smoking (1%). The decision tree generated by Trial 2 as seen in Figure B.2 provides clearer in-depth insights on a possible methodology for medical professionals to ascertain whether a patient is at high-risk of CVD.

		Predicted	
		False	True
Real Values	False	5641	2628
	True	1368	4368

		Predicted	
		False	True
Real Values	False	4904	3116
	True	2100	3880

Table 1: Confusion matrix for Trial 1 (left) and Trial 2 (right)

To summarise, CART has ranked the importance of biomarkers in determination of CVD. In decreasing order, we have systolic blood pressure, diastolic blood pressure, cholesterol, age, glucose levels, weight and history of smoking.

As blood pressure can be easily measured at home, this presents an opportunity for integration into a possible telehealth platform for patients to easily perform preliminary checks before scheduling a follow-up appointment. Subsequently during the appointments, the other biomarkers can be accurately measured to supplement or provide a broad direction to investigate the possibility that the patient has CVD. The wide range of factors aggregated over time will also be useful for the medical professional to monitor the health of patients during regular health checks and flag out patients who are possibly at high risk of CVD to suggest medical intervention before it is too late.

On the flip side, the factors that are not statistically important to CART such as gender, activity levels also help medical professionals to zoom in on the important factors, reducing the consultation time required per patient and increasing the number of available consultation slots for other patients.

4. Stroke Prediction

4.1 Choice of Dataset

It was crucial to find a dataset that contained both stroke and CVD data. This was done so that the relationship between getting CVD and stroke can be found. It was found that the stroke prediction data from Kaggle set fulfilled this need. The dataset also consists of other features that overlap the dataset used in the CVD dataset, allowing similar data exploration techniques to be applied.

4.2 Data Exploration

4.2.1 Pre-existing Conditions Affecting Stroke

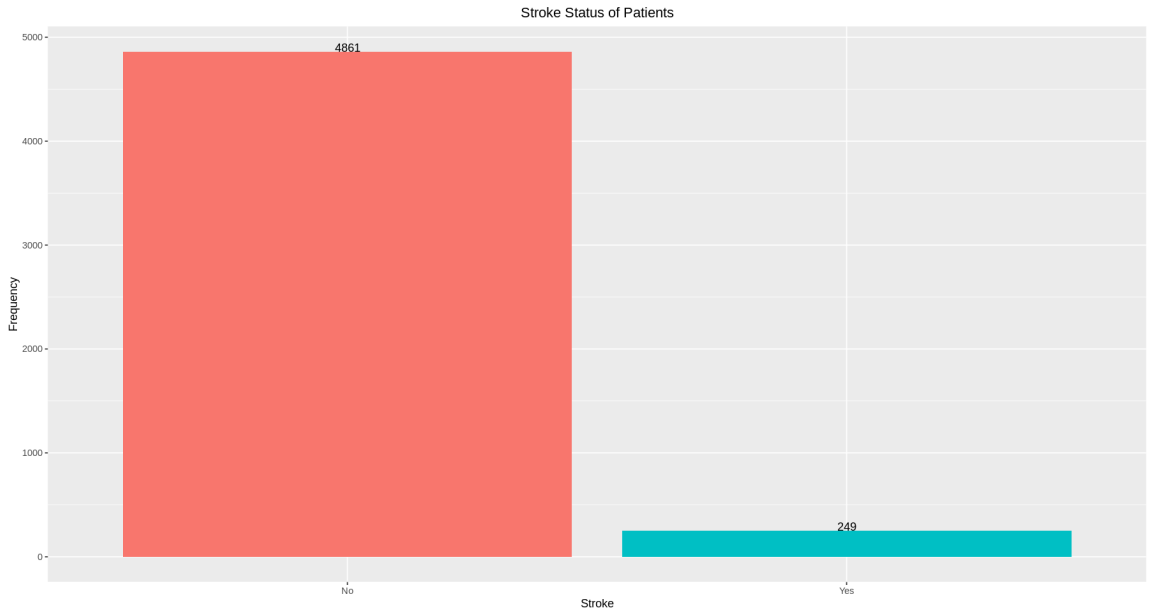

Figure 15: Number of People with Stroke in Dataset

Figure 15 shows that a majority of people surveyed did not have a stroke. Out of 5000 patients, 249 people suffered from a stroke.

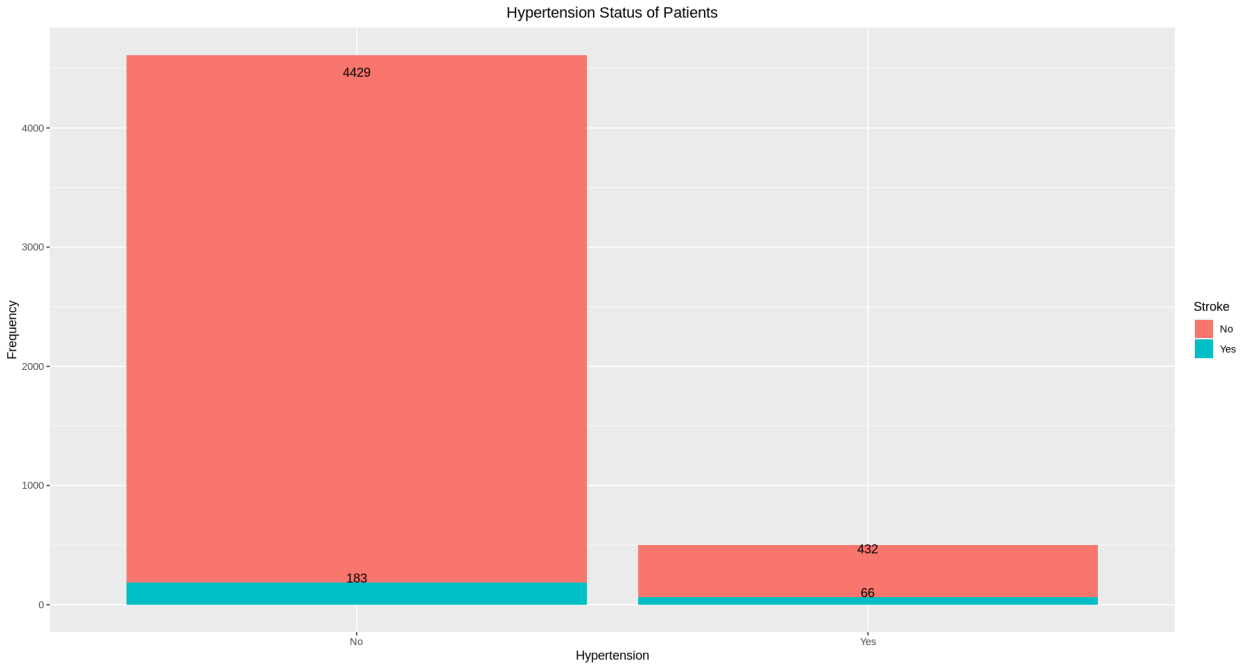

Figure 16: Number of People with Stroke and Hypertension in Dataset

From Figure 16, the number of patients without hypertension is vastly greater than the number of patients with hypertension. However, the gap is slightly less when we compare stroke data. 4.13% of patients who do not have hypertension have stroke, while 15.28% of hypertension patients have stroke. This is an indicator that there may be a correlation between having hypertension and getting a stroke.

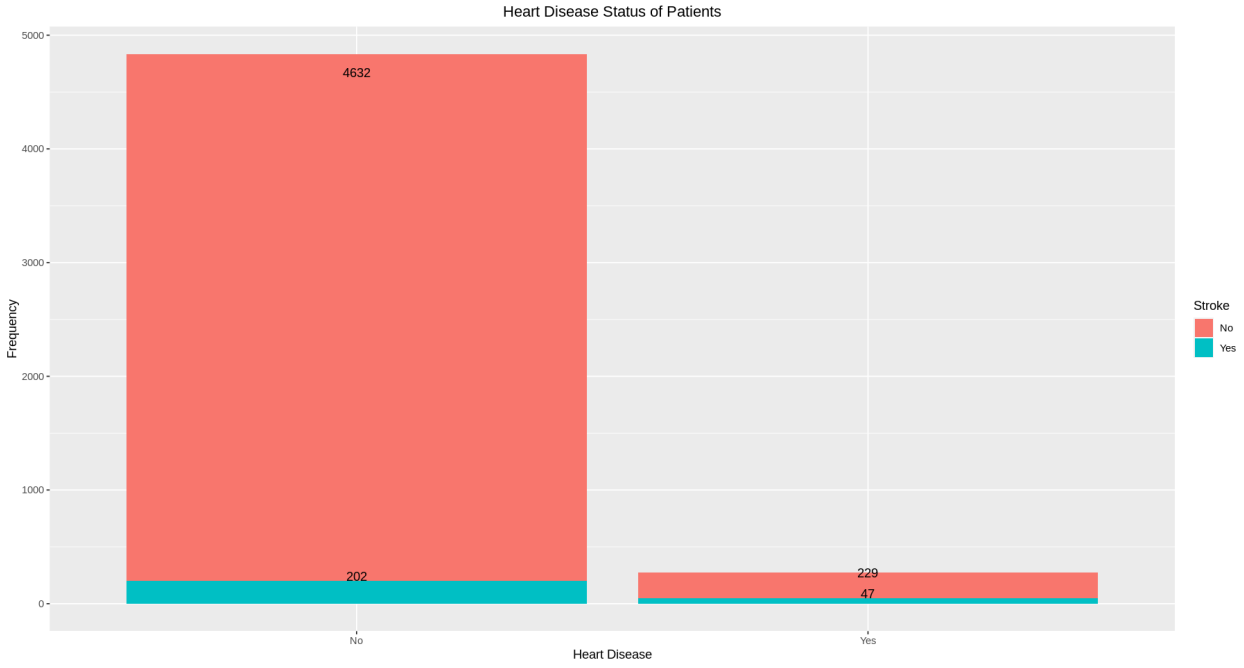

Figure 17: Number of People with Stroke and CVD in Dataset

Figure 17 shows that 4.36% of patients who do not have CVD have stroke, while 20.52% of CVD patients have stroke. This is an indicator that there may be a correlation between having CVD and getting a stroke.

4.2.2 Patient Profiles Affecting Stroke

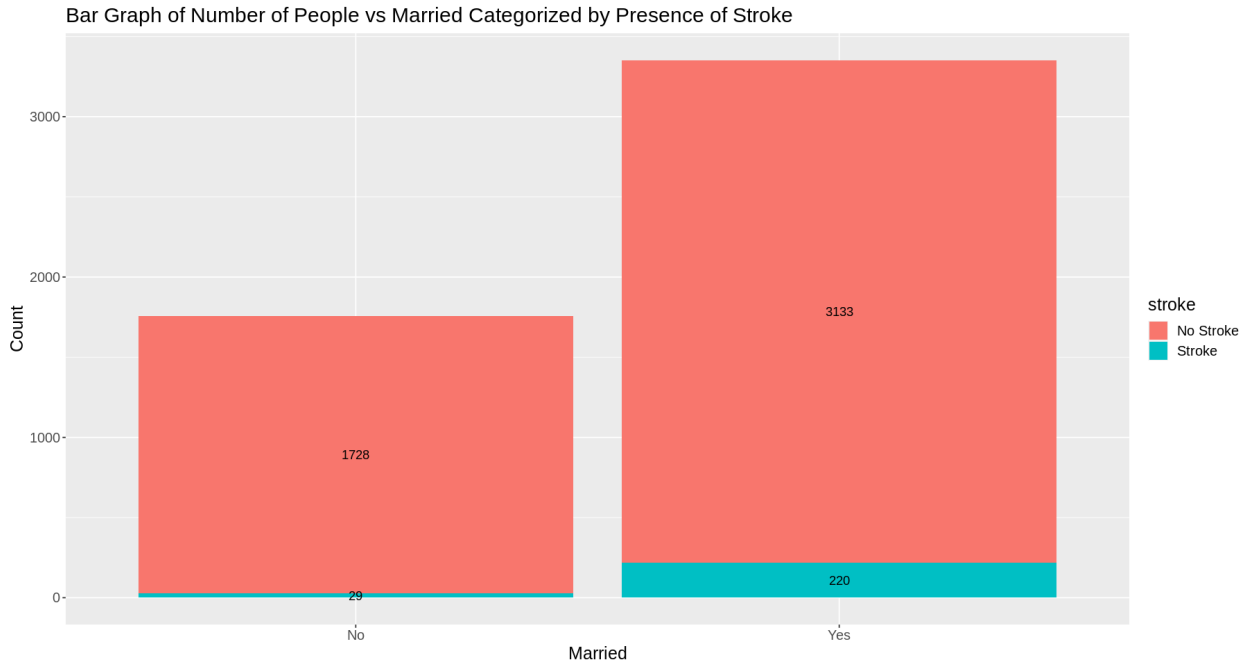

Figure 18: Number of Married People with Stroke in Dataset

Figure 18 shows that out of the married patients, 7.02% have stroke. For those who are unmarried, only 1.678% have stroke. This is an interesting observation which is proven true by statistics.

4.2.3 Patient Profiles Affecting Stroke

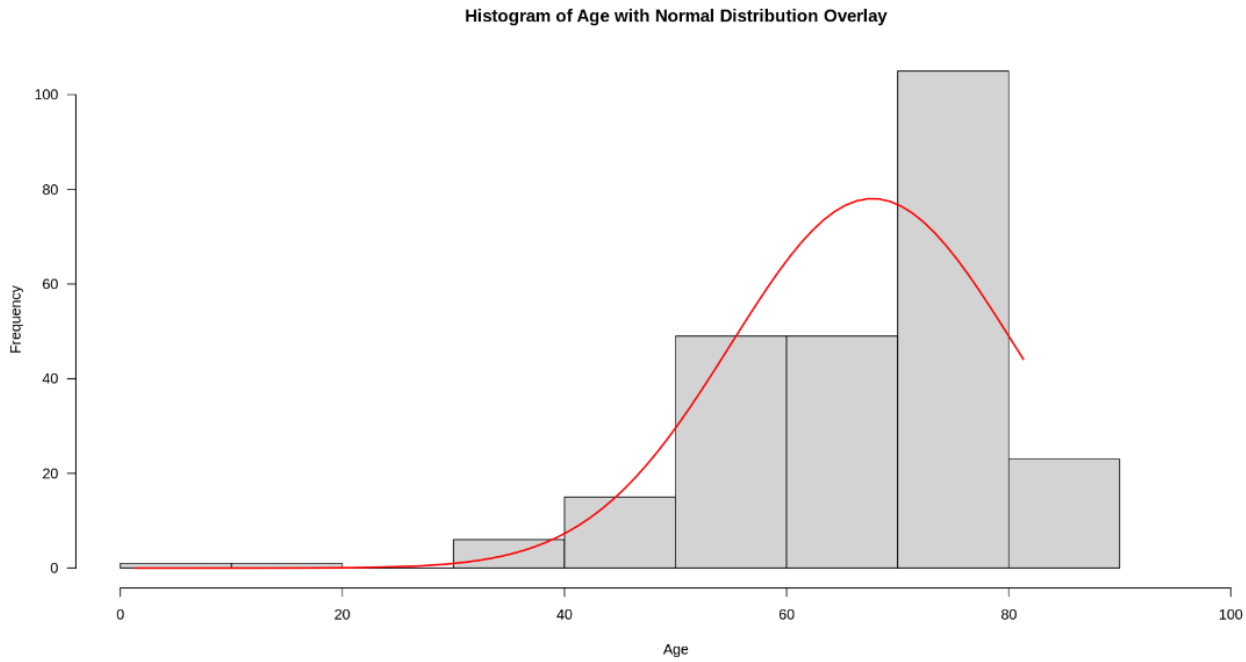

Figure 19: Histogram of Stroke Patients by Age

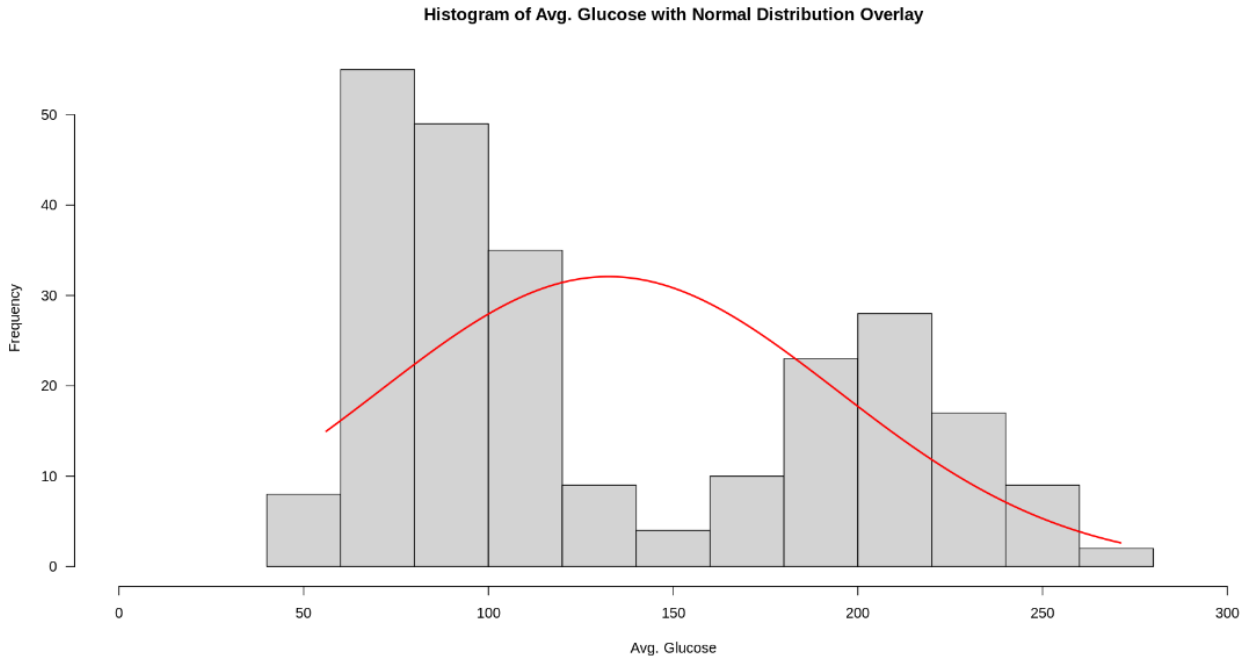

Figure 20: Histogram of Stroke Patients by Glucose Level

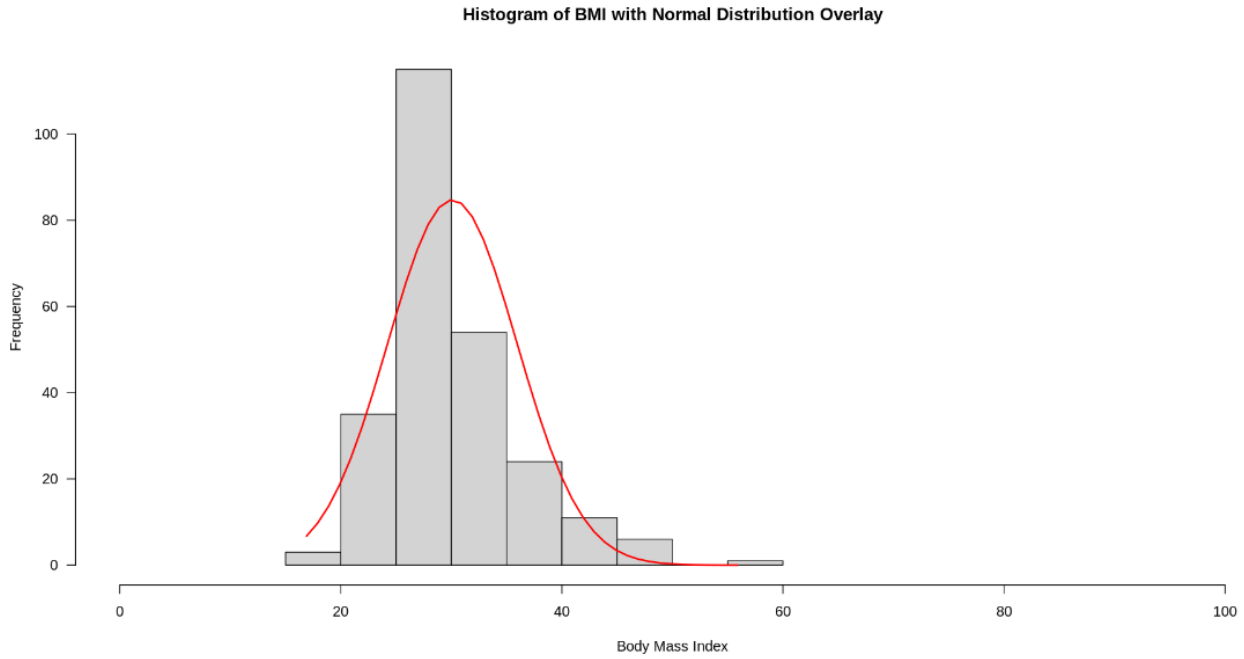

Figure 21: Histogram of Stroke Patients by Body Mass Index

Figure 19 shows that the median age of stroke patients is about 45. This shows that people of older ages are more susceptible to getting a stroke. Figure 20 shows that the median glucose level for stroke patients occurs around 91.89. Figure 21 shows that the median BMI for stroke patients lies around 28.1. This is almost within the "Obese" BMI range.

4.3 Predictive Model

Our team utilised two methods to generate predictive models for stroke from patient data. For each method, we mainly focused on 3 different ways to generate the model - using the entire dataset, using a predetermined curated dataset, and using a generative lightest dataset.

4.3.1 Generalised Linear Model (GLM)

As mentioned previously, the GLM model was chosen as a model because of its computational efficiency, robustness and interpretability. The model enabled us to quickly understand each effect of predictor variables on the final output and was also able to handle our chosen dataset quickly.

Our efforts with GLM can largely be summarised into three trials:

Trial 1- A predictive model on the likelihood of stroke with all patients data (70.25%)
Trial 2- A predictive model on the likelihood of stroke with features chosen from EDA (79.35%)

Trial 3- A brute force method to determine the best predictive model on the likelihood of stroke with the minimum combination of patient data (70.25%)

The results of all three trials allowed us to produce a predictive model up to **79.35%** accuracy on the testset.

4.3.2 Logistic Regression

The logistic regression is also an apt model as it is commonly used to estimate the probability of a binary outcome (presence of a CVD in our case). Logistic regression is a type of GLM that is based on the logit function instead of the binomial function (used in the above model). Even though they are similar, the behaviour of the estimation methods can differ in corner cases and for singular types of data.

Our efforts with Logistic Regression can largely be summarised into two trials:

Trial 1 - A predictive model on the likelihood of stroke with all patients data (95.01%)
Trial 2- A predictive model on the likelihood of stroke with features chosen from EDA (95.11%)
Trial 3- A brute force method to determine the best predictive model on the likelihood of stroke with the minimum combination of patient data (95.01%)

The Logistic Regression models give us a predictive model with **95.11%** accuracy on the testset.

However, the confusion matrix shows that GLM is potentially a better prediction model. For the GLM model, the true positive rate is 54% while the false positive rate is 19.34%. This allows However, for the logistic regression model, the true positive rate and false positive rate are both 0%. It is posited that it might be due to the bias of the data (more than 95% of participants do not have a stroke) causing the model to have a steep bias towards predicting everything as false. The objectively decent true positive rate allows hospitals to still pick up on at-risk patients and its lower false positive rate allows hospitals to avoid wasting overt resources on false predictions.

		Predicted	
		False	True
Real Values	False	784	188
	True	23	27

		Predicted	
		False	True
Real Values	False	972	0
	True	50	0

Table 2: Confusion Matrix for GLM model (left) vs Logistic Regression model (right)

5. Future Exploration

This section of the report details the potential future steps that could be taken to supplement our current research if provided the necessary times, resources and data.

5.1 Heart Attack Relapse

Amongst patients who have experienced a heart attack before, 20% of them are likely to have a relapse within the next five years [15]. Prevention measures currently in place are to consistently monitor a patient's key risk factors, such as blood pressure. However, in the case of secondary heart attacks, time has been proven to be extremely vital in the preservation of heart tissue, as well as survival rate.

Our team believes that a repurpose of our researched technology into the discovery of higher-risk individuals (those who are likely to have a heart attack relapse) would be beneficial to medical institutes and patients. This comes with the assumption that patients' data are recorded post-heart attack and that such specific data exists with a sufficiently large enough quantity.

The provided dataset only contains 14 features out of which 6 have been extracted as important features. The UCI datasets also contain other features such as fasting blood sugar, heart rate, etc. This means that it is possible to have other highly correlated features that were not considered. We believe that a detection system for heart attack relapse can add value to both patients and medical institutes. Patients greatly benefit from the increased assurance regarding the state of their health. This would also enable their loved ones to better prepare in the event of emergencies. For example, this might be in the form of additional preparation and monitoring equipment for individuals at higher-risk of relapse. Medical institutes on the other hand, are able to profit via enhanced treatment programs tailored to the risk of the CVD patient. This might include diet regimes or assisted monitoring systems that are targeted towards patients who require them.

5.2 Diet Recommendations

Another future expansion of this project is to include patients' food consumption patterns and other related information such as calorie intake, dietary restrictions, proportion of different nutrients in meals. This information can then be substantiated with our current datasets containing columns like age, height, weight, BMI, presence of stroke and presence of cardiovascular heart disease. With these combined datasets, it is then possible for us to use machine learning to predict the proportion of different nutrient types for the patients to nudge the patients towards healthier lifestyles with more balanced eating habits. Such a tool will help NHCS better assist their patients by providing a more comprehensive and well-rounded treatment plan for the patients.

6. References

1. Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... & American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2020). Heart disease and stroke statistics—2020 update: a report from the American Heart Association. Circulation, 141(9), e139-e596.
2. Kim, J., Thayabaranathan, T., Donnan, G. A., Howard, G., Howard, V. J., Rothwell, P. M., ... & Thrift, A. G. (2020). Global stroke statistics 2019. International Journal of Stroke, 15(8), 819-838.
3. Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.
4. Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, No. 1, pp. 25-29).
5. Amin, S. U., Agarwal, K., & Beg, R. (2013, April). Genetic neural network based data mining in prediction of heart disease using risk factors. In 2013 IEEE conference on information & communication technologies (pp. 1227-1231). IEEE.
6. Su, T. T., Amiri, M., Mohd Hairi, F., Thangiah, N., Bulgiba, A., & Majid, H. A. (2015). Prediction of cardiovascular disease risk among low-income urban dwellers in metropolitan Kuala Lumpur, Malaysia. BioMed research international, 2015.
7. Cardiovascular Disease dataset. (2019, January 20). Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
8. UCI Machine Learning Repository: Heart Disease Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Heart+Disease
9. Diabetes Dataset. (2020, August 5). Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
10. Stroke Prediction Dataset. (2021, January 26). Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
11. Govender, R.D., Al-Shamsi, S., Soteriades, E.S. et al. Incidence and risk factors for recurrent cardiovascular disease in middle-eastern adults: a retrospective study. BMC Cardiovasc Disord 19, 253 (2019). https://doi.org/10.1186/s12872-019-1231-z
12. Quer, G., Arnaout, R., Henne, M., & Arnaout, R. (2021). Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. Journal of the American College of Cardiology, 77(3), 300–313. https://doi.org/10.1016/j.jacc.2020.11.030
13. Heart Disease and Stroke | CDC. (n.d.). https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

14. Heart Disease Statistics - Singapore Heart Foundation. (n.d.). Singapore Heart Foundation. https://www.myheart.org.sg/health/heart-disease-statistics/#:~:text=In%20Singapore%2C%2021%20people%20die,to%20heart%20diseases%20or%20stroke.
15. Proactive steps can reduce chances of a second heart attack. (2021, August 16). www.heart.org.https://www.heart.org/en/news/2019/04/04/proactive-steps-can-reduce-chances-of-second-heart-attack#:~:text=Yet%2C%20about%20one%20in%20five,blood%20supply%20to%20that%20area.

Appendix

Appendix A - Project Schedule/ Gantt Chart
Appendix B - Trees generated by CART

Appendix A - Project Schedule/Gantt Chart

Tasks	JAN				FEB				MAR				APR			
Dataset Research																
Project Proposal and Data Cleaning																
Data Exploration																
Predictive Model Generation																
Report Writing and Slides																

Appendix B - Trees generated by CART

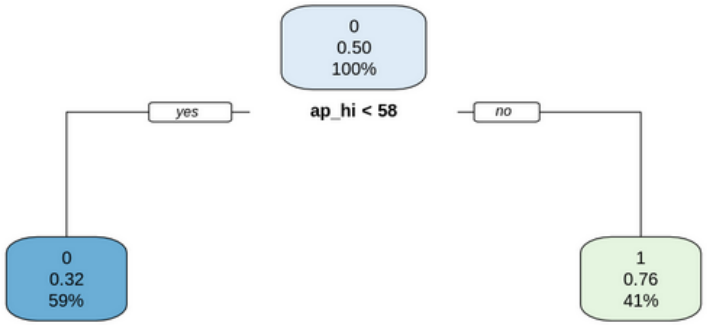

Figure B.1: CART model from Trial 1

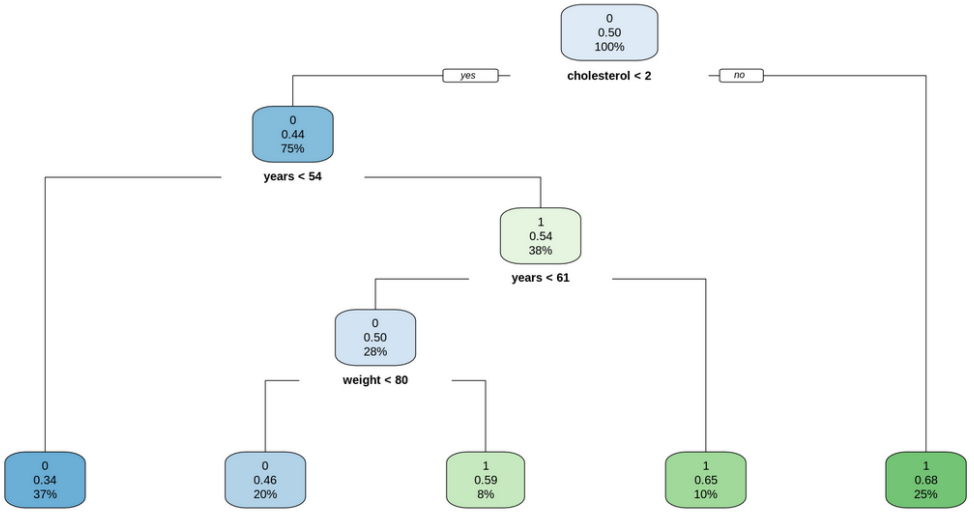

Figure B.2: CART model from Trial 2