



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

BC2406 ANALYTICS I VISUAL & PREDICTIVE TECHNIQUES

Analytical model for Early Intervention of Cardiovascular Disease

AY22/23 Sem 1 | Seminar 7, Team 6

NAME	MATRICULATION NUMBER
Bryan Lim Kai Wen	U2121763H
Gerald Ong Jon Kai	U2122864K
Hong Zhi Hao	U2121820G
Jiang Lei	U2121557B
Toh Jing Qiang	U2121442H
Zainab Zuvairiyah D/O V M Saifudeen	U2010049J

Table of Contents

Executive Summary	4
1. Introduction	5
1.1 Background of CVD in Singapore	5
1.2 Early Intervention of CVD	5
1.3 Business Problem	5
1.4 Existing Solution	5
1.5 Proposed Solutions.....	5
o Stage 1: Individual Empowerment	6
o Stage 2: Prediction at Primary Care	6
2. Stage 1: Individual Empowerment.....	7
2.1 Data Cleaning & Pre-processing.....	7
2.1.1 Handle missing values	8
2.1.2 Filter outlier in continuous variables.....	8
2.1.3 Clean up categorical variables	8
2.1.4 Encode categorical variables	8
2.2 Data Exploration	8
2.3 Modelling.....	9
2.3.1 Pre-modelling preparations	9
2.3.2 Logistic Regression	9
2.3.3 Gradient Boosting Classifier (GBC)	10
2.3.4 Random Forest.....	11
2.3.5 Model Selection.....	11
2.3.6 Model Explanation.....	11
3. Stage 2: Prediction at Primary Care	13
3.1 Data Cleaning and Pre-Processing	13
3.1.1 Remove irrelevant variables (columns).....	14
3.1.2 Handle missing values	14
3.1.3 Filter outlier in continuous variables.....	14
3.1.4 Clean up categorical variables	14
3.1.5 Convert categorical values to meaningful string	14
3.2 Data Exploration	15
3.3 Modelling.....	15
3.3.1 Pre-modelling preparations	15
3.3.2 CART – Decision Tree (dectree_m1 and dectree_m2)	15
3.3.3 Random Forest	17
3.3.4 Conclusion on Models.....	18
3.3.5 Model Explanation.....	19
4. Business Implementation.....	20
4.1 Summary and Flow chart	20
4.2 Expected Outcomes	21

4.2.1 Raise Public Awareness	21
4.2.2 Provide Timely Alert	21
4.2.3 Optimise Medical Procedures	22
4.3 Comparison with Latest Solution	22
4.4 Limitations and Future Improvements	23
4.4.1 Data Origin and Imbalance	23
4.4.2 Other risks factors not considered	24
5. Conclusion.....	24
References.....	25
Appendices.....	27
Appendix A: Stage 1 – Data Exploration.....	28
Continuous Variables	28
Categorical Variables	30
Appendix B: Stage 1 – 17 Questions Individuals Can Answer	36
Appendix C: Stage 1 – What at-risk Individuals can do by themselves.....	37
Appendix D: Stage 1 Logistic Regression Models.....	38
Model Training Procedures	38
Performance of 3 Logistic Regression models	39
Appendix E: Stage 1 Gradient Boosting Classifier (GBC)	40
Performance of Gradient Boosting Classifier Model (gbc_m1).....	40
Appendix F: Stage 1 Random Forest Classifier	41
Performance of Random Forest Model (random_forest_m1)	41
Feature Importance	42
Appendix G: Stage 2 – Data Exploration.....	44
Continuous Variables	44
Categorical Variables	48
Overall Correlation	50
Appendix H: Stage 2 CART and Random Forests	52
CART.....	52
Random Forest.....	52
Advantages and Disadvantages of Random Forest compared to CART	53
Appendix I: Stage 2 CART and Random Forests Full Performance Results	54
CART Model Performance Results	54
Random Forest Model Performance Results	56

Executive Summary

This report aims to deploy data analytics to solve the business problem for National Heart Centre Singapore (NHCS). Given the increasing incidence of reported cases of cardiovascular disease (CVD) in Singapore, NHCS handles more than 120,000 outpatient consultations each year. The sudden onset of heart disease is severe and expensive to treat. Therefore, NHCS can shift the focus to early prevention rather than treating post-diagnosis.

To increase the involvement of individuals and primary care sectors in the prevention of heart disease, our team proposes a 2-step solution – HeartDetect. The first stage is to raise individuals' awareness and manage their heart health regularly. The second stage is to enable the prediction of heart disease risk in the primary care sector to provide timely prevention.

The first stage enables individuals to predict their risk of heart disease using personal metrics easily. We discovered through the dataset analysis that adults over 60 have a significant prevalence of heart disease. Moreover, variables such as "stroke", "diabetes", "kidney disease", and "difficulty walking" were strongly associated with heart disease. To estimate the risk of heart disease accurately, our team selected a random forest model that outperformed other models, namely the logistic regression model and gradient boosting classifier, in terms of overall accuracy, false negative score and ROC-AUC score. It utilizes 17 variables of personal attributes for which data could be collected through a simple questionnaire. To make it simpler for patients to understand the prediction result, we used a locally interpretable model-diagnostic interpretation (LIME) to interpret the model behaviours in terms of contributing factors.

The second stage is designed to provide a decision support tool to help clinicians from primary care identify patients at high risk of a heart attack. Our data analysis revealed a significant correlation between chest pain and heart disease and a higher maximum heart rate. An optimized random forest model was finally selected because of its excellent performance in accuracy and low false negative rate. It uses 11 medical attributes collected from primary check-ups for classifying patients with high or low risk of heart disease. Shapley Additive exPlanations (SHAP) - a model explainer - was employed to better assist physicians in understanding the predicted results. Providing physicians with the impact of each variable on the level of risk not only builds physician confidence in using the model. It also gives physicians insights into patients' conditions to plan prevention or treatment accordingly.

The two stages mentioned above are complementary to each other. Suppose an individual is at high risk in stage 1; they are instructed to use medical indicators for further screening, as done in stage 2. Alternatively, an automated cardiac risk report is generated when the individual undergoes a regular physical examination, as in stage 2. He/she will be encouraged to monitor their heart health in their daily life through the stage 1 protocol.

Overall, this solution aims to raise awareness among Singaporeans on the prevention of heart disease and reduce the severe life-threatening effects of heart disease through early detection, timely intervention, and optimal treatment. It also allows NHCS to make the best use of available resources to provide priority and timely treatment for high-risk patients by avoiding unnecessary treatment for those with a low chance of developing heart disease.

1. Introduction

1.1 Background of CVD in Singapore

Cardiovascular disease (CVD) is consistently ranked among Singapore's top three causes of hospitalization and death. One in five Singaporeans has one or more risk factors for CVD, such as diabetes, high blood pressure, high cholesterol, smoking and physical inactivity. (Khoo, 2022) Currently, CVD accounts for one-third of deaths in Singapore (Singapore Heart Foundation, 2022); hence there is a pressing demand to reduce the prevalence of CVD among Singaporeans.

Besides that, many Singaporeans are still unaware of CVD risk factors (Anthony, 2020). This lack of awareness hampers efforts to encourage early intervention, resulting in unhealthy lifestyles, which increase the risk of CVD. Furthermore, manual assessment of CVD risk levels significantly affects human resources, with NHCS handling more than 120,000 outpatient consultations yearly. (National Heart Centre Singapore, n.d.)

1.2 Early Intervention of CVD

Given the severity of CVD, early intervention is crucial to preventing its onset and increasing the life expectancy of Singaporeans (Chrysant, 2011). A European study concluded that the elimination of health risk behaviours could prevent 80% of CVD (Piepoli, et al., 2016). With the slow onset of CVD and long incubation period, there is a large window of opportunity for effective early intervention instead of treating post-diagnosis, which it will generally be more serious (Qian, et al., 2022). Therefore, the early detection of high-risk individuals for CVD is critical for the prevention of CVD (Liu, et al., 2019).

1.3 Business Problem

With the rising prevalence of CVD in Singapore and the low awareness of Singaporeans towards its risk factors, there is a need for a data-analytics based predictive tool to help in the early intervention of CVDs among individuals and be utilized by medical professionals at the primary care level.

1.4 Existing Solution

PRECISE (Predictive Risk Score for CAD In Southeast Asians with Chest Pain) is Singapore's first diagnostic risk calculator for coronary artery disease (CAD). It estimates the likelihood of developing CAD in patients with no known history by using variables such as age, gender, underlying chronic conditions, smoking status, type of chest pain, pain radiating to the neck, and electrocardiogram (ECG) changes. According to Dr Chee Fang Yee (National Heart Centre of Singapore, 2021), PRECISE is particularly helpful for patients with chest pain to monitor their health status and focus on identifying specific CAD.

1.5 Proposed Solutions

The existing solution PRECISE affirms the promise of data analysis for risk prediction in the clinical setting. However, PRECISE only targets people presenting with chest pain and exclusively addresses CAD, a subset of cardiovascular disease, which leaves early intervention coverage for CVDs inadequate.

Our team, therefore, proposes HeartDetect, a two-step prevention approach that will enable more Singapore residents, regardless of their signs of discomfort or history of the disease, to monitor their heart disease risk conveniently and accurately. Such a solution will raise the quality of healthcare provided to Singaporeans through identifying heart disease risk factors and timely intervention, raising awareness of risk factors and hence promoting a healthier lifestyle, and

maximising medical resource allocation. Focusing on heart disease would cover a broader scope towards CVDs as compared to only CADs.

- **Stage 1: Individual Empowerment**

The first phase enables individuals to monitor their risk of having heart disease through data analysis using relevant biological, medical, behavioural, genetic, and environmental information about the individual. Users can predict their risk of heart disease by answering questions anytime, anywhere. At the same time, it increases their awareness and prevention understanding of heart disease. This phase aligns with the RIE2025 initiative, Singapore's 5-year research and development strategy. The initiative empowers individuals to better manage their health and chronic conditions through data-driven and patient-centric solutions to promote wellness and prevent disease.

- **Stage 2: Prediction at Primary Care**

Individuals in the high-risk category for heart disease can visit a primary care physician for screening. Here is when our stage 2 solution comes into play. We provide a data analytics model as a decision-making support tool for physicians. The tool will identify people at high risk for heart disease using vital signs from a basic physical exam. With it, physicians from primary care can better assess a patient's heart disease risk index so that the patient can receive appropriate and timely care.

2. Stage 1: Individual Empowerment

The first stage of our proposed solution is to allow individuals to quickly assess the risk of heart disease using their medical history, lifestyle, current health condition, etc. Suppose individuals find they are in a high-risk group using the prediction model. In that case, they can monitor their health more closely and seek timely medical treatment.

2.1 Data Cleaning & Pre-processing

The dataset was obtained from [Kaggle](#). The dataset contains 18 variables (14 categorical variables and 4 continuous variables), and the overview of the original dataset is summarized in the Table 1.

Variables	Datatype	Description
Smoking	Boolean [Yes / No]	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
AlcoholDrinking	Boolean [Yes / No]	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
Stroke	Boolean [Yes / No]	(Ever told) (you had) a stroke?
DiffWalking	Boolean [Yes / No]	Do you have serious difficulty walking or climbing stairs?
Sex	Boolean [Male / Female]	Are you male or female?
AgeCategory	Categorical Variable From 18 to 80+, divided into 13 groups	Thirteen-level age category
Race	Categorical Variable [White / Hispanic / Black / Asian / Other]	Imputed race/ethnicity value
Diabetic	Boolean [Yes / No]	(Ever told) (you had) diabetes?
PhysicalActivity	Boolean [Yes / No]	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
GenHealth	Categorical Variable [Very Good / Good / Excellent / Fair / Poor]	Would you say that in general your health is...
Asthma	Boolean [Yes / No]	(Ever told) (you had) asthma?
KidneyDisease	Boolean [Yes / No]	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
SkinCancer	Boolean [Yes / No]	(Ever told) (you had) skin cancer?
BMI	Continuous Variable	Body Mass Index (BMI)
PhysicalHealth	Continuous Variable [0-30]	Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 was your physical health not good?
MentalHealth	Continuous Variable [0-30]	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?
SleepTime	Continuous Variable	On average, how many hours of sleep do you get in a 24-hour period?
HeartDisease	Boolean [Yes / No]	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

Table 1: Overview of Stage 1 Dataset

2.1.1 Handle missing values

There are no NA entries in this dataset.

2.1.2 Filter outlier in continuous variables

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319795.000000	319795.000000	319795.000000	319795.000000
mean	28.325399	3.37171	3.898366	7.097075
std	6.356100	7.95085	7.955235	1.436007
min	12.020000	0.00000	0.000000	1.000000
25%	24.030000	0.00000	0.000000	6.000000
50%	27.340000	0.00000	0.000000	7.000000
75%	31.420000	2.00000	3.000000	8.000000
max	94.850000	30.00000	30.000000	24.000000

Figure 1: Summary of Continuous Variables

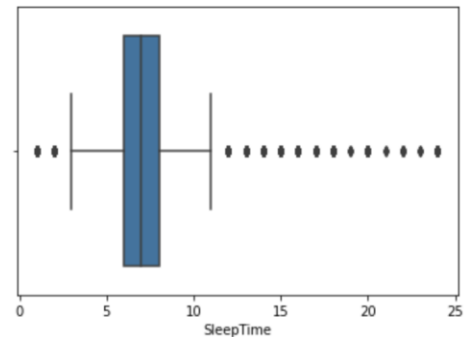


Figure 2: Barplot of SleepTime

From the summary of continuous variables, the maximum value of the SleepTime variable is 24 hours, indicating the presence of outliers in SleepTime. The confidence interval is defined by $(\mu - 3\sigma, \mu + 3\sigma)$, which is 2.79-11.41, and values outside this confidence interval can be considered outliers and are excluded.

2.1.3 Clean up categorical variables

The values of all categorical variables are consistent, and there are no missing values. No data cleaning is required for them.

2.1.4 Encode categorical variables

For categorical variables, one hot encoding, which encodes them into numerical form, is required to facilitate training models. Encoding results in a total of 46 categorical columns, and with 4 continuous columns, the cleaned dataset has 50 columns and 315252 rows.

2.2 Data Exploration

After cleaning the data, data visualization tools from ggplot2 package were used to derive insights. The key findings are listed below, and all the plots and insights are listed in [Appendix A](#).

Unbalanced Distribution of HeartDisease

The distribution of the predictor variable HeartDisease is highly unbalanced: 91% of rows with class "No" – Low risk; and 9% of rows with class "1" – high risk. Hence, the trainset should be oversampled until the even distribution among 2 classes before training the models.

Comparison of individual characteristics between groups with and without heart disease

Heart disease rates are high among people **over 65** ($> 10\%$), and more than 30% of people with **poor** overall health status suffer from heart disease. Of those with a history of **stroke** or diagnosed with **diabetes** or **kidney** disease, 35%, 21%, and 29% had heart disease, respectively. Among people with **difficulty walking** or **climbing stairs**, 22% have heart disease, in line with the research that found that people that walk fast (Yates, et al., 2017) and have no difficulty in climbing stairs (Vazquez, Bouzas-Mosquera, Rivadulla-Varela, Barbeito-Caamano, & Vazquez-Rodriguez, 2021) show better cardiac status. People who **smoke** are at double the risk of heart disease.

The proportion of people suffering from heart disease is similar (the difference is less than 5%) across different **rac**es, **g**enders, **physical-active**, **people that drink alcohol**, **people with asthma**, and **skin cancer**. The same applies to **BMI**, **MentalHealth** and **SleepTime**.

2.3 Modelling

The objectives of the stage 1 modelling included building the optimal model for predicting heart attack risk based on essential personal attributes and discovering the important personal indicators for prediction.

2.3.1 Pre-modelling preparations

a) Train-Test Split

The dataset was randomly divided into training-test dataset with a ratio of 7:3.

b) Oversampling for trainset

Since the distribution of predictor `HeartDisease` was unbalanced (91% low risk & 9% high risk), SMOTE (Synthetic Minority Oversampling Technique) was employed to oversample the trainset to make the distribution of heart disease be 1 (with heart disease): 1 (without heart disease).

c) Metrics to Measure Performance

The performance of the model will be measured primarily from the following 4 dimensions:

- 1 Classification Accuracy
- 2 Confusion Matrix with true positive rate and false negative rate
- 3 Precision / Recall / F-score
- 4 ROC AUC Curve score

Of these, the ROC AUC score will be used as the primary metric because true-positive and true-negative classes are equally cared for, and AUC calibrates the trade-off between sensitivity and specificity at the best-chosen threshold. While the overall accuracy measures the performance of a single model, the AUC compares two models. It evaluates the performance of the same model at different thresholds.

The False Negative Rate (FNR) indicates the possibility of misclassifying high-risk populations as low-risk (prediction 0, truth 1). A higher false negative rate than a false positive rate is a more severe problem because it may delay users from seeking timely treatment. Therefore, the lower the FNR, the better the model performs.

Therefore, given this medical background, the best model selection will be based on **classification accuracy**, **ROC AUC score** and **false negative rate** as the primary measures.

2.3.2 Logistic Regression

Logistic regression is one of the most efficient machine learning classification algorithms when the different outcomes or distinctions represented by the data are linearly separable. It is widely used in the medical field, such as Trauma and Injury Severity Score (TRISS), for predicting mortality in injured patients (Boyd, Tolson, & Copes, 1987).

a) Model Building with Feature Engineering Using Recursive Feature Elimination (RFE)

This dataset contains up to 50 variables after one-time coding. Backward feature elimination is performed until the desired number of features is reached. RFE is employed to select features by recursively considering smaller and smaller sets of features. For this dataset, 3 different features will be selected for RFE, and the performances of resulting models are compared in [Appendix D](#).

b) Model Evaluation

Model logreg_m21 with 21 variables is selected as the optimal logistic regression model. However, it does not meet the pre-determined benchmark for prediction, as shown in the Table 2.

	logreg_m21 [Selected Model]	Benchmark
Overall Accuracy	74.22%	>80%
False Negative Rate	78.46%	<20%
ROC-AUC Score	0.75	>0.7

Table 2: logreg_m21 performance

The selected best logistic regression model did not meet most of the predetermined benchmarks, i.e., overall accuracy and false negative rate. Therefore, a better model with higher predictive power is desired to be trained.

2.3.3 Gradient Boosting Classifier (GBC)

To obtain higher prediction accuracy and a lower false negative rate than Logistic Regression, GBC was chosen to build a more robust model which learns iteratively from each weak learner.

a) Model Building and Hyperparameter Tuning

The hyperparameter tuning is conducted on these 2 parameters, n_estimators (the number of boosting stages to perform) and learning_rate, to obtain the best parameters for the model to perform well.

From the plot, the best value of learning_rate and n_estimator combination is (1, 100).

Model Fitting with best hyperparameters

- n_estimators = 100
- learning_rate = 1
- max_features = 2
- max_depth = 2
- random_state = 0

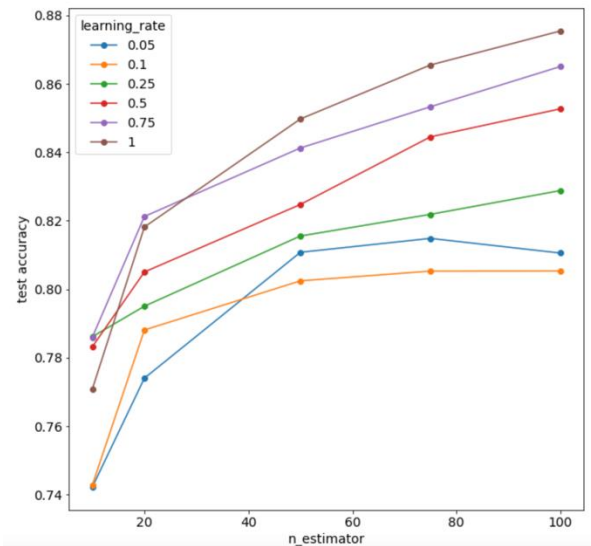


Figure 3: learning_rate and n_estimator against test accuracy

b) Model Evaluation

	GBC Model	Benchmark
Overall Accuracy	87.54%	> 80%
False Negative Rate	72.28%	< 20%
ROC-AUC Score	0.63	> 0.7

Table 3: Gradient Boosting Classifier performance

The Gradient Boosting Classifier model is better than the best Logistic Regression model, and there is a significant boost in the overall accuracy from 74% to 87.54%, as seen in Table 3. However, the improvement in accuracy came at the expense of false negative rates, with more than 72.28% of high-risk patients being misclassified as low risk, which is undesirable. Therefore, a model with better predictive power for positive cases is demanded.

2.3.4 Random Forest

Random forest is an ensemble of decision trees. It randomly selects observations and features to build several decision trees and then averages the results. Thus, it helps reduce overfitting and is expected to reduce the false negative rate in the test dataset.

Model Evaluation

	Random Forest Model	Benchmark
Overall Accuracy	99.62%	> 80%
False Negative Rate	1.70%	< 20%
ROC-AUC Score	0.99	> 0.7

Table 4: Random Forest Model Performance

The random forest classifier model performs satisfactorily on the test set and meets all predefined benchmarks, as shown in Table 4. While achieving ultra-high accuracy, it ensures a very low false negative rate, making it a superior model.

2.3.5 Model Selection

Optimal Model

	Logistic Regression	Gradient Boosting Classifier	Random Forest
Overall Accuracy	74.22%	87.54%	99.62%
False Negative Rate	78.46%	72.28%	1.70%
ROC-AUC Score	0.75	0.63	0.99

Table 5: Comparison of models

The Random Forest Classifier model (random_forest_m1) performs the best among all models with the highest accuracy, 99.62%, the highest ROC-AUC Score of 0.99 and the lowest false negative rate of 1.70%, as shown in Table 5. Hence, it is selected as the best model for stage 1. The optimal model (random_forest_m1) reveals that all **17 features** are important in making an accurate prediction of heart disease. The rank of their importance are shown from the list on the right.

Additionally, `BMI` is found to be the top important feature in predicting heart disease. Hence, individuals should monitor their weight in a healthy BMI range to maintain heart health.

2.3.6 Model Explanation

The model evaluation above selected random forest as the best model based on predefined metrics. However, an accuracy number may mean little to the user. In addition, random forest is an ensemble model, which means that the algorithm behind it is complex and difficult to understand intuitively. In this case, a model explainer can be employed to interpret the prediction results to build end-user trust.

Local Interpretable Model-agnostic Explanations (LIME) is employed to explain predictions of random forest model for individual record data. Figure 5 explains the prediction results for randomly selected records from the test set in terms of each contributing variable.

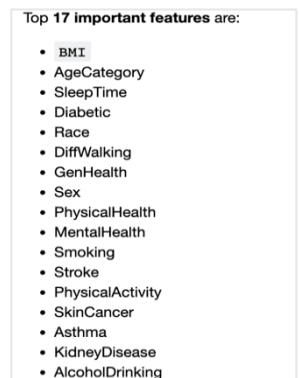


Figure 4: Top 17 important features of Random Forrest Model

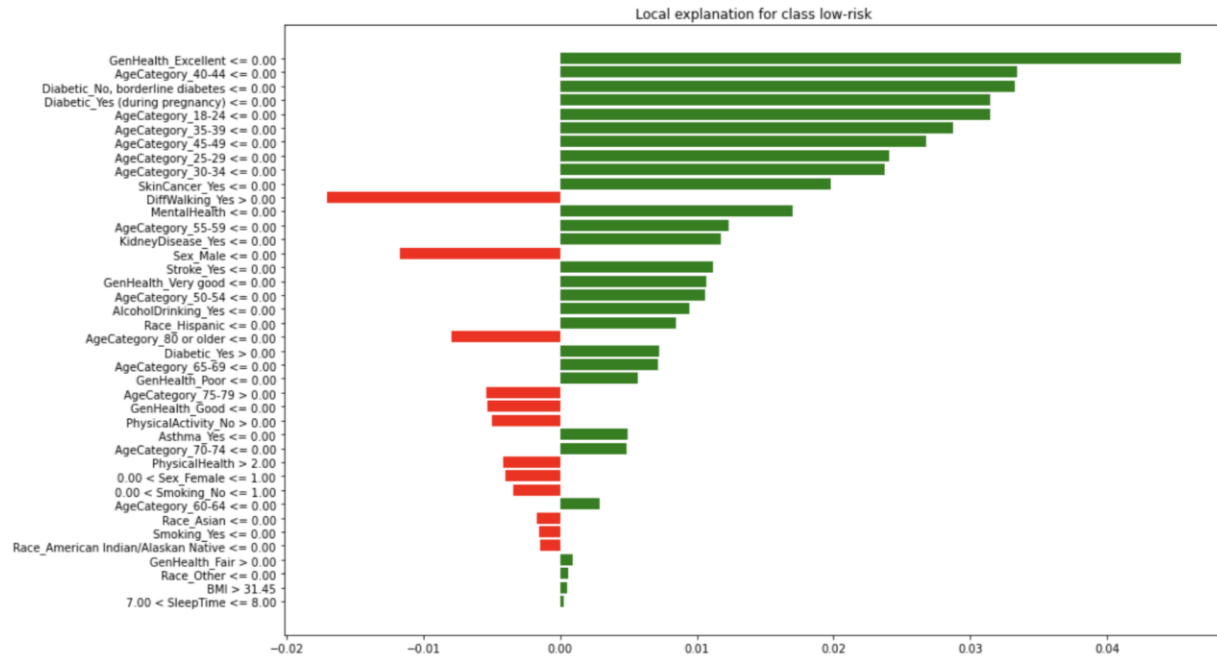


Figure 5: LIME explanation for a randomly selected record

Hence, each data record can be well explained using LIME algorithm, which provides a qualitative understanding between the input variables and the response to facilitate understanding and decision making.

3. Stage 2: Prediction at Primary Care

The second stage is to provide a decision-making tool to assist doctors in identifying patients with high chances of heart attack using the patient's vitals from a check-up. Once the doctors have identified that the patient is at high risk of a heart attack, they can work together with the patient to take preventive measures for the heart attack.

3.1 Data Cleaning and Pre-Processing

The dataset for this stage was obtained from [Kaggle](#). The steps described in this section can be found in 'data-cleaning-preprocessing'. The cleaned dataset contains 12 variables (7 categorical and 5 continuous variables), and the overview of the original dataset is summarized in Table 6.

**Variables were renamed to more meaningful names*

Variables	Datatype	Description
age	Continuous Variable	Age of the patient
sex	Boolean [Male / Female]	Sex of the patient
chest_pain	Categorical Variable [No Chest Pain / atypical angina / non-anginal pain / typical angina]	Chest Pain type
resting_blood_pressure	Continuous Variable	Resting blood pressure (in mm Hg)
chol	Continuous Variable	Cholesterol in mg/dl fetched via BMI sensor
fasting_blood_sugar	Boolean [True / False]	Fasting blood sugar > 120 mg/dl
rest_ecg	Categorical Variable [having ST-T wave abnormality / normal / showing probable or definite left ventricular hypertrophy Estes' criteria]	Resting electrocardiographic results
max_heart_rate	Continuous Variable	Maximum heart rate achieved
exercise_induced_angina	Boolean [Yes / No]	Exercise induced angina
num_of_major_vessels	Categorical Variable [0 / 1 / 2 / 3 / 4]	Number of major vessels colored by fluoroscopy
o2_saturation	Continuous Variable	Saturation Level
oldpeak	Continuous Variable	Previous Peak
slp	Categorical Variable [0, 1, 2]	Slope
thall	Categorical Variable [0, 1, 2, 3]	Thal rate
heart_attack_chance	Boolean [Less Chance / More Chance]	Chance of heart attack

Table 6: Overview of variables

3.1.1 Remove irrelevant variables (columns)

The variables oldpeak, slp, and thall were removed from the analysis because they were not explained, and the collection method was not clear.

3.1.2 Handle missing values

There are no NA entries in this dataset.

3.1.3 Filter outlier in continuous variables

	age	trtbps	chol	thalachh	o2_saturation
count	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.324503	131.662252	246.294702	149.907285	97.484106
std	9.067887	17.554429	51.914022	22.489378	0.342667
min	29.000000	94.000000	126.000000	88.000000	96.500000
25%	47.250000	120.000000	211.000000	134.500000	97.500000
50%	55.000000	130.000000	240.500000	153.000000	97.500000
75%	61.000000	140.000000	274.750000	166.000000	97.500000
max	77.000000	200.000000	564.000000	202.000000	98.600000

Figure 6: Summary of Continuous Variable

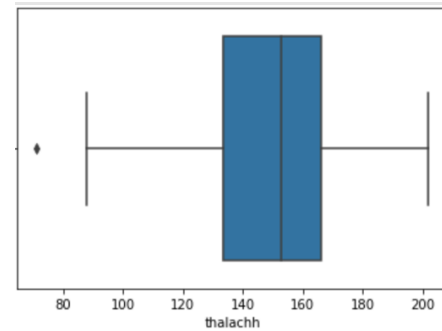


Figure 7: Boxplot of thalachh

From the summary of continuous variables, the minimum value of the max_heart_rate variable is 71 bpm, indicating the presence of outliers in max_heart_rate (American Heart Association, 2021). The confidence interval is defined by $(\mu - 3\sigma, \mu + 3\sigma)$, which is 81.04 - 218.25, and values outside this confidence interval can be considered outliers and are excluded.

3.1.4 Clean up categorical variables

The values of all categorical variables are within a reasonable range and there are no missing values. No data cleaning is required for them.

3.1.5 Convert categorical values to meaningful string

For categorical variables, we converted the values into meaningful strings for EDA and visualization purposes.

- sex: 1 = male, 0 = female
- exercise_induced_angina: 1 = yes; 0 = no
- chest_pain: Chest Pain type chest pain type
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic
- fasting_blood_sugar: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg: resting electrocardiographic results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- heart_attack_chance: 0= less chance of heart attack 1= more chance of heart attack

3.2 Data Exploration

After cleaning the data, data visualization tools from ggplot2 package were used to derive insights.

The key findings are listed below, and all the plots and insights are listed in [Appendix G](#).

The chances of heart disease increases when one has a **higher max heart rate**, which is in line with research (Perret-Guillaume, Joly, & Benetos, 2009). Besides that, **females** are shown to be at a **higher risk of a heart attack** (~5%), which also follows research on the relationship between gender and heart disease (Carauna, 2018).

The presence of chest pain is a sign of a heart attack (Singapore General Hospital, n.d.), which follows the trend shown as all those with **chest pain** have a greater than 69.6% higher chance of heart attack. It also applies to **rest ECG** (Beckerman, et al., 2005) as those with ST-T wave abnormality have 17.3% more chance of higher risk of heart disease. In addition, the more **major vessels** (not blocked or interrupted by fatty substances) **detected by fluoroscopy**, the lower the chance of having a high risk of heart disease.

The proportion of people suffering from heart disease is similar (the difference is less than 5%) across different **resting_blood_pressure, chol, o2_saturation, fasting_blood_sugar**.

Some anomalies that we notice that are different from current research and findings would be for **age** and **exercise-induced angina**. The younger the individual is, the higher the risk of getting a heart disease, which contradicts existing research (Rodgers, et al., 2019). Also, exercise-induced angina is a common complaint of cardiac patients (Brown & Oldridge, 1985). However, our exploration shows that not having exercise-induced angina means one is at a higher risk of heart disease.

3.3 Modelling

The goals of the second phase of modelling include building the best model for predicting heart attack risk based on medical attributes, discovering important variables for prediction, and mapping the analysis for end users.

3.3.1 Pre-modelling preparations

a) Train-Test Split

The dataset was randomly divided into training-test dataset with a ratio of 7:3.

b) Metrics to Measure Performance

The performance of the model will be measured in the following 5 dimensions:

- 1 Classification Accuracy (5-fold cross validation with ROC-AUC-score)
- 2 Confusion Matrix with true positive rate and false negative rate
- 3 Precision / Recall / F-score
- 4 Out-of-bag (OOB) score (for random forest)
- 5 ROC AUC Curve score.

With the same reasons stated in [2.3.1 \(c\)](#), the decision-making process will be focused based on **classification accuracy, ROC AUC score**, and **false negative rate** given this medical context.

3.3.2 CART – Decision Tree (dectree_m1 and dectree_m2)

CART is one of the most common and powerful models in machine learning. It is inexpensive to process and transparent, different from the "black box" of neural network models, so the decision tree used for classification can be quickly built and understood.

a) Tree Growing and Pruning

First, grow the tree to the maximum depth – 9. The decrease in accuracy on the test set (67.25%) compared to that on trainset (76.89%) indicates that this tree is overfitted on the trainset. Therefore, pruning is performed to reduce the size of the decision tree by removing the branches that do not provide greater power to classify instances.

To find the point where the tree will be pruned, 2 plots were generated at different α levels.



Figure 8: Alpha vs Tree Size

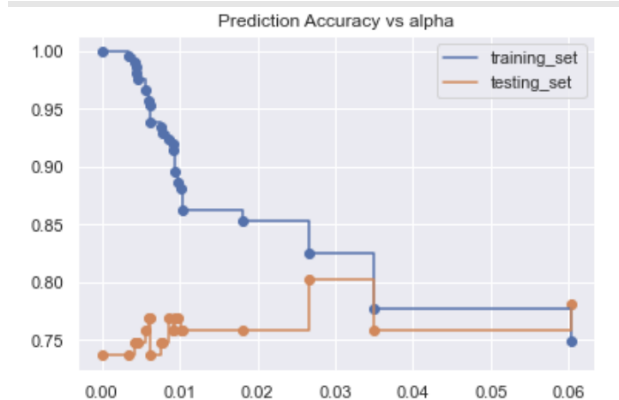


Figure 9: Prediction Accuracy vs Alpha

As seen from the plot on the left, the size of the tree decreases as α increases. This decreasing rate abruptly slows down when α exceeds 0.01. From the figure on the right, there is a trade-off when maximising the accuracy for the training set and the test set. To maximise both accuracies, that point between the alpha value of 0.02-0.03 is spotted. Therefore, 0.03 is chosen as the α value where the tree is pruned.

b) Model Evaluation & Important Features

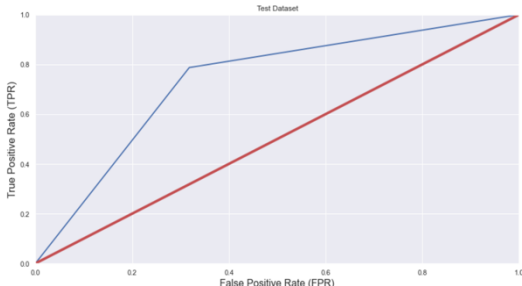
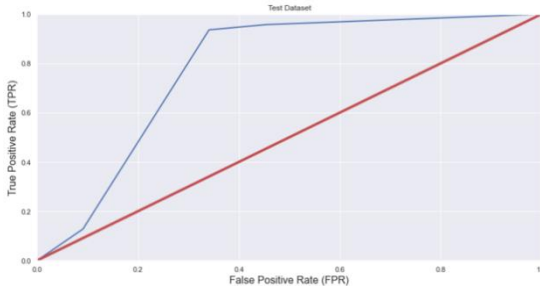
	Pre-pruned Tree [Test set]	Pruned Tree [Test set]
Classification Accuracy	67.25%	74.41%
True Positive Rate	66%	69.57%
False Negative Rate	34%	30.43%
F1 Score	68.04%	68.82%
ROC & ROC AUC Score	 <p>ROC AUC Score of pre-pruned tree: <u>0.7345</u>.</p>	 <p>ROC AUC Score of pre-pruned tree: <u>0.7802</u>.</p>

Table 7: Evaluation of Models

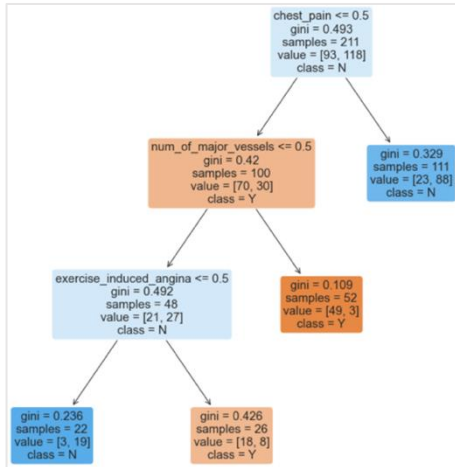


Figure 10: Pruned Decision Tree

The pruned decision tree is shown below, which uses `chest_pain`, `num_of_major_vessels`, `exercise_induced_angina` as 3 key classification factors. After pruning, the performance of CART model is greatly improved in all the above 6 metrics. However, the predictions from this stage will be used by physicians to classify a patient's risk of heart disease. Therefore, a higher accuracy (>85%) in prediction is expected. In addition, the type 2 error (False Negative Rate) should be further reduced so that fewer patients at high risk are misclassified as low risk.

3.3.3 Random Forest

To obtain a better performance model in terms of overall accuracy and false negative rate, we built a random forest classifier, which is essentially a CART algorithm but an aggregate of many trees instead of just one, improving performance through the wisdom of crowds and less likely to be overfitted.

a) Model Building (random_forest_m1)

Build a random forest classifier model with all features and default hyperparameters.

b) Model Optimization – Feature Importance (random_forest_m2)

The random forest model is further optimized by considering the feature importance. The ranking of feature importance for the random forest model is as follows:

	feature	importance
0	max_heart_rate	0.158
1	num_of_major_vessels	0.157
2	chest_pain	0.136
3	age	0.108
4	exercise_induced_angina	0.105
5	chol	0.104
6	resting_blood_pressure	0.091
7	sex	0.064
8	o2_saturation	0.040
9	rest_ecg	0.024
10	fasting_blood_sugar	0.013

Table 8: Ranking of Feature by Importance

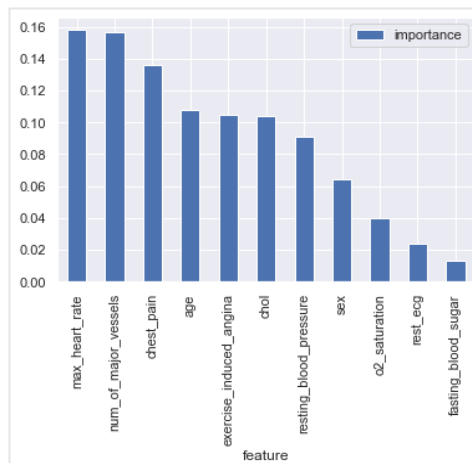


Figure 11: Ranking of Feature by importance

The feature, fasting_blood_sugar, has a very low importance score of 0.13 compared to other features.

This feature was then dropped to examine whether there would be a boost in performance.

c) Model Optimization – Hyperparameter Tuning (random_forest_m3)

By undergoing hyperparameter tuning, the model's behaviour can be better controlled to maximize model performance. If the hyperparameters are not correctly tuned, the model will produce suboptimal results as it will not minimize the loss function, resulting in more errors.

The Hyperparameter Tuning outcome from GridSearchCV is shown on the right.

```
{
  'criterion': 'entropy',
  'min_samples_leaf': 1,
  'min_samples_split': 4,
  'n_estimators': 100
}
```

Figure 12: Hyperparameter Tuning Outcome

d) Model Evaluation (on test dataset)

Metrics	random_forest_m1	random_forest_m2	random_forest_m3
Classification Accuracy	0.8431	0.8171	0.8506
True Positive Rate	0.7954	0.8	0.7907
False Negative Rate	0.2045	0.2	0.2093
F1 Score	0.7692	0.7826	0.7555
ROC-AUC Score	0.8902	0.8871	0.8917
Out-of-bag (oob) score	83.89%	84.83%	84.36%

Table 9: Random Forest Model Evaluation

The 3rd model, random_forest_m3 with hyperparameter tuning, performs best among all three random forest models, with the highest classification accuracy of 85%, a ROC AUC score of 0.89, and a similarly low false negative rate of 0.2093 as shown in Table 8.

* See [Appendix I](#) for detailed evaluations of each model.

3.3.4 Conclusion on Models

Models / Metric	Classification Accuracy	False Negative Rate	ROC-AUC Score
dectree_m2 (CART)	74.41%	30.43%	78.02%
random_forest_m3	85.06% - (increase by 10.65%)	20.93% - (decrease by 9.5%)	84.36% - (increase by 6.34%)

Table 10: Comparison of Models

From Table 9, the optimal random forest classifier model performs way better than the optimal CART model for all 3 important metrics. See [Appendix H](#) for details regarding the differences between CART and Random Forest.

Thus, the optimal random forest classifier model is selected as decision-making assistant to be used by physicians in stage 2 with the following important features.

1. max_heart_rate
2. num_of_major_vessels
3. chest_pain
4. age
5. exercise_induced_angina
6. chol
7. resting_blood_pressure
8. sex
9. o2_saturation
10. rest_ecg
11. fasting_blood_sugar

3.3.5 Model Explanation

The above model analysis selected random forest as the best model based on predefined metrics. As explained in [Section 2.3.6](#), the accuracy number means little to the user.

Therefore, **SHapley Additive exPlanations** (SHAP) is employed to explain predictions of the random forest model for both individual record data and the weighted overall importance of each variable.

Hence, each data record can be well explained using SHAP algorithm, which provides a qualitative understanding between the input variables and the response to facilitate understanding and decision-making.

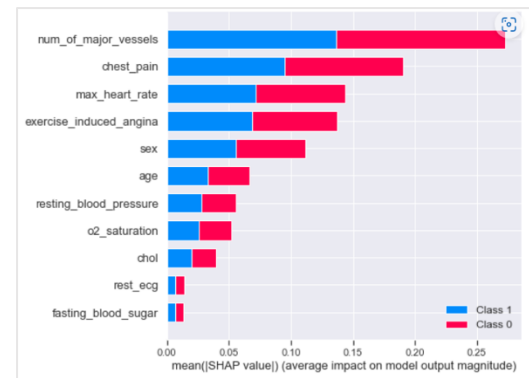


Figure 13: Average Positive/Negative Impact of each variable

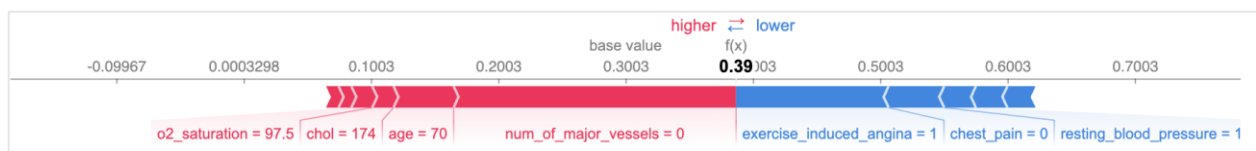


Figure 14: Example of SHAP explanation for a randomly selected record

The figure above explains the prediction results for randomly selected record from the test set in terms of each contributing variable.

4. Business Implementation

4.1 Summary and Flow chart

Our 2-stage solution, HeartDetect, is designed to arm the NHCS with the ability to identify individuals at risk for heart disease at an early stage so that intervention can be provided promptly.

The first stage enables individuals to easily predict their risk of heart disease using personal metrics, while the second phase is designed to provide a decision support tool to help clinicians identify patients at high risk of heart attack using their vital signs during check-ups.

The diagram below illustrates the procedures of our solution.

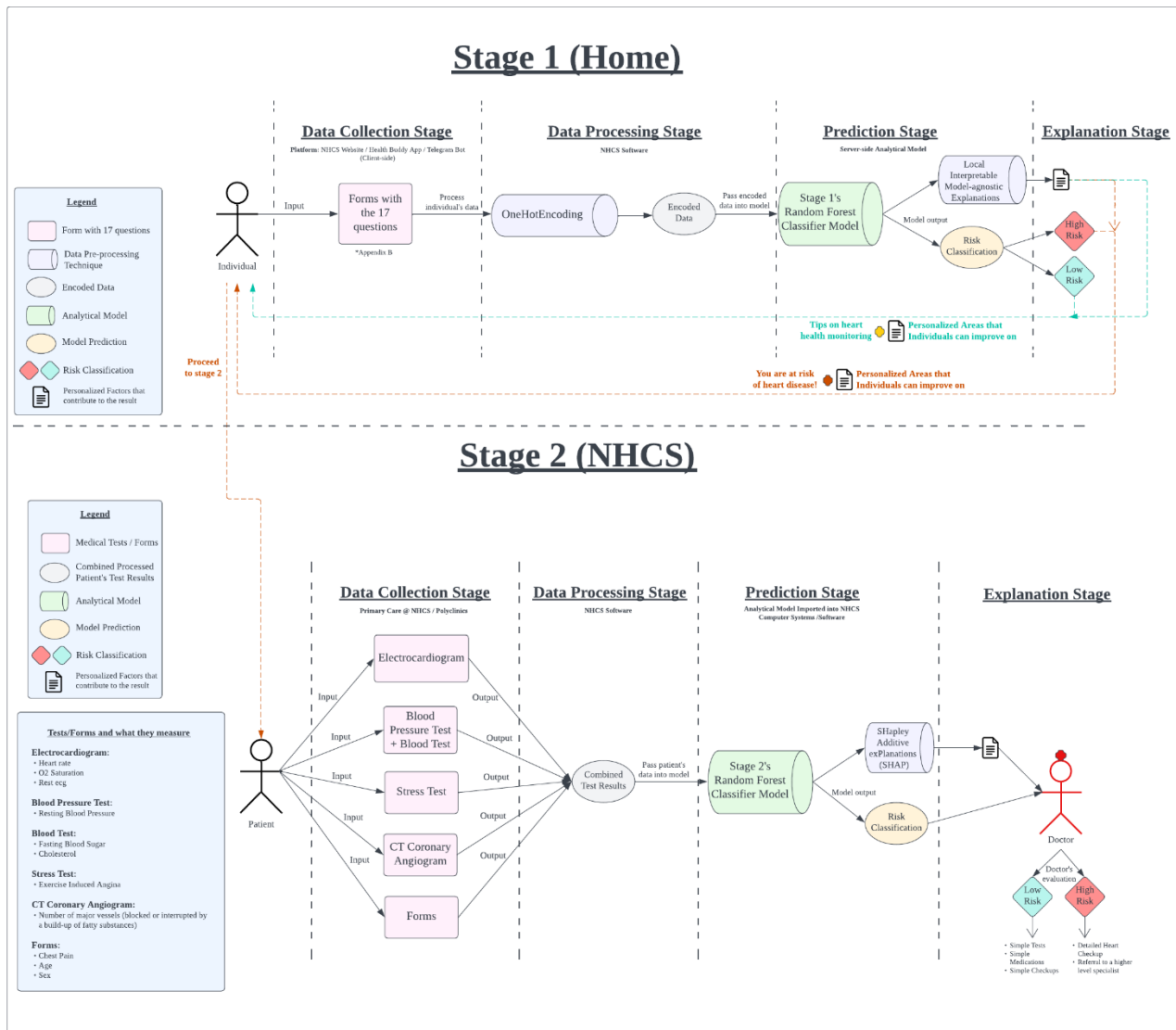


Figure 15: Flowchart of HeartDetect

Stage 1

Individuals can predict their likelihood of developing heart disease in this stage, which can be done anytime, anywhere through accessible platforms such as the NHCS website, Health Buddy mobile app, etc. Social media applications such as Telegram bot can also be utilised to extend coverage to a young population,

Individuals will answer 17 easy-to-understand questions (see [Appendix B](#)) on these platforms. The answers will be validated and processed through one-hot coding before sending for prediction.

The individual's report of potential heart disease risk, predicted by the random forest model, will be fed back to the user along with the attribute analysis generated by the LIME model, such as "low sleep duration increases your risk of heart disease". Users will be instructed to maintain heart health even in the low-risk group. Otherwise, if the user is found to be in the high-risk group, he/she will be directed to consult a general doctor or specialist for screening.

In addition, the same questionnaire will be sent periodically to NHCS visitors who have symptoms of heart disease or a history of related diseases (such as diabetes) to remind them to monitor their heart health regularly.

Stage 2

After a person is predicted to be at high risk in Stage 1, he/she will be directed to a primary care centre, such as a polyclinic, for a general examination.

The clinician will perform primary tests such as ECG, blood pressure test, blood test, stress test and coronary CT ECG. These data will be processed and sent to an optimal random forest model. The output will be interpreted through SHAP, providing insights on which medical data would significantly impact risk levels and to what extent so that clinicians can better understand the model results and the meaning behind them. These insights introduce a greater explainability to the heart disease prediction process.

The physician will utilize the prediction result and indicator analysis report and apply his or her expertise to make the most appropriate treatment or prevention plan. If the risk is low, the next course of action would be simple testing with medications and screening, while if the patient is predicted in high risk, a detailed cardiac workup and referral to a secondary hospital would be provided.

4.2 Expected Outcomes

4.2.1 Raise Public Awareness

HeartDetect will raise heart disease awareness in Singapore. By answering the questions required for stage 1, the individual gains awareness of the risk factors for heart disease regardless of whether they have a high or low risk. Raising awareness is a step towards preventing heart disease, as everyone knows his risk factors for heart disease from the analytic report and will be encouraged to engage in primordial prevention (actions such as eating more healthy foods), which has been found to aid in mitigating risk factors in later years by avoiding risk factors altogether (Gilman, 2015).

4.2.2 Provide Timely Alert

Apart from raising awareness of heart disease risk amongst individuals, HeartDetect also provides a timely alert to individuals by allowing all residents to monitor their heart disease risks with an insightful analysis report. With the risk classification they received from HeartDetect, the individuals can seek timely intervention before the heart disease materialises, such as regular check-ups, adopting a healthier lifestyle, monitoring personal attributes such as weight, etc.

The timely alert is expected to engage individuals in primary intervention on their part by monitoring the risk factor of heart disease in this case (Kisling & Das, 2022). Hence, the standard

of care for heart disease will improve significantly as the overall focus can be shifted to prevention instead of attempting to treat the symptoms.

4.2.3 Optimise Medical Procedures

With HeartDetect, specialist consultation slots can be allocated based on the patient's level of risk. Optimised allocation of medical resources ensures that the patients who need it most are treated and helps improve outcomes by conducting timely interventions.

Suppose the stage 1 analytical model determines that individuals are at low risk. They will be instructed to follow the NHCS' preventive measures (see [Appendix C](#)) and continue monitoring their heart disease regularly from home. However, if the model determines they are at risk, they must go for a check-up at NHCS.

At stage 2, when the Stage 1 high-risk patient goes through primary check-ups, the doctor will get a prediction result and an analysis report. This report breakdowns the contributing factors and allows doctors to better determine the next course of action. For instance, if the variable of exercise-induced chest pain contributes significantly to the patient's high risk, the physician can decide on the next test (with his medical expertise), such as a coronary artery disease screening.

HeartDetect would thus help the patient save medical costs as he/she will be directed to go through more optimised actions to tackle the high risk rather than trial and error. The cost savings are a big help for treating heart disease as patients with heart disease incur more than twice the medical costs compared to those without heart disease (Kumar, Siddharth, Singh, & Narang, 2022).

4.3 Comparison with Latest Solution

For this portion, we would be primarily comparing HeartDetect with PRECISE, which is the latest solution to tackle the problem of heart disease and the comparison is summarised in Table 10.

	HeartDetect		PRECISE
	Stage 1	Stage 2	
Purpose	Allow people to self-detect their risk of having heart disease	Allow doctors to detect the risk of having heart disease	Allow patients/doctors to self-detect/detect the risk of having coronary artery disease
Variables	17	11	7
Prediction Model used	Random Forest	Random Forest	Logistic Regression
Explainer Model	LIME	SHAP	
Output	High Risk (1) or Low Risk (0) (Categorical) with Explanation of factors	High Risk (1) or Low Risk (0) (Categorical) with Explanation of factors	Probability of having CAD (Continuous)

Table 11: Comparison of HeartDetect and PRECISE

HeartDetect and PRECISE work towards the same goal, which is risk detection. However, PRECISE is only for CAD, while HeartDetect is for the whole category of heart disease. By

tackling the whole category of heart diseases, we are helping reduce the impact of CVD more. In addition, our model caters to a larger group of people in Singapore as PRECISE is a tool designed specifically for people aged 30 years and above with stable chest pain. Whereas HeartDetect is meant for people of all ages and whether they have chest pain.

Besides the different types of diseases, the solutions are meant for, both solutions also employ different models and have different outcomes.

For HeartDetect, the model we decided on was a Random Forest model that helps tell us whether individuals are at High Risk or Low Risk of having heart disease. The chosen model differs from PRECISE, which uses a Logistic Regression model that helps tell the individual their probability of getting a CAD. We chose a categorical variable as our output variable to ensure there is no subjectivity to the user of what percentage is high risk.

On top of the difference in output variables, HeartDetect explains how the factors contribute to the outcome using LIME and SHAP. Explaining allows the user to have a clearer understanding of the variables that contribute to the model's outcome. This explanation would help us achieve Explainable AI instead of a black box model.

Apart from the difference in output variables, HeartDetect uses more variables than PRECISE for both stages. The choice of some variables varies at each stage to offer a better prediction. Although PRECISE does offer the option of adding additional variables (Significant Q Waves and ST-T Abnormalities) for those with resting ECG available (probably doing the test in clinics), it is only partially specialised for personal and clinical usage. On the other hand, for PRECISE, we use easily collectable variables for Stage 1 and, subsequently, more medical variables for Stage 2.

In conclusion, HeartDetect is better than PRECISE in terms of purpose, outcome, and explanation of how the model produces the prediction and customisation for the respective stages of usage.

4.4 Limitations and Future Improvements

4.4.1 Data Origin and Imbalance

The dataset used in our model is not based on data from Singaporeans but from Americans. As there are differences in demographics and lifestyles across different countries, this may lead to inaccuracies in local implementation.

There may also be an impact due to the imbalance in the dataset - 91% low risk and 9% high risk in Phase 1. This imbalanced dataset will generate models skewed towards classification under low risk. It will have a higher rate of falsely classifying individuals as low risk, leading to a higher false negative rate, which is undesirable. While this imbalance is dealt with by oversampling the training set using SMOTE, the data may not simulate the real-world data accurately. In the context of the rising incidence of cardiovascular disease in Singapore (Singapore Heart Foundation, 2022), this limitation of not using local and imbalanced data may be magnified.

To overcome this limitation, we will start by collecting more cardiovascular disease data from Singaporeans. This will allow the dataset to be more consistent with the local situation, thus optimizing the model to make it more reliable and applicable.

4.4.2 Other risks factors not considered

The data set could also perform better with other risk factors not considered, namely genetics, which has been found to play a role in detecting heart diseases (Knowles & Ashley, 2018)

We could research or consult experts to determine which genetic indicators, such as the family history of heart disease, have a qualitative impact on predicting the risk of heart disease. Future models can be augmented by research and updated to include these potentially helpful indicators, which would help improve the prediction model's accuracy.

5. Conclusion

To conclude, our solution, HeartDetect aims to provide NHCS with a comprehensive analytical model and a unique approach to tackling the pertinent issue of rising heart disease death rates. In HeartDetect, each phase is customised to be suitable for the relevant demographics. To recap, stage 1 makes it simple and convenient for all Singaporeans to assess their heart health and risk level. With a readily accessible self-monitoring tool, we can raise the awareness of heart disease risk factors as well as the steps taken for prevention. On the other hand, stage 2 streamlines primary care for medical professionals, introducing explainability into prediction results and gives more direction to the treatment process through SHAP.

Overall, HeartDetect will help to maximize resource allocation for NHCS, allowing them to deal with heart disease more efficiently in Singapore in 2 ways – by ensuring that individuals in need are prioritised, and shifting the focus from treatment to prevention.

References

- Anthony, R. C. (2020, Feb 10). *MIMS News*. Retrieved from <https://specialty.mims.com/topic/more-than-one-third-of-singaporeans-may-be-unaware-of-cvd-risk-factors>
- Chrysant, S. G. (2011). A new paradigm in the treatment of the cardiovascular disease continuum: focus on prevention . *Hippokratia*, 7-11.
- Khoo, B. (2022, September 12). *Farrer Park Hospital*. Retrieved from Connection between High Cholesterol and High Blood Pressure : <https://www.farrerpark.com/farrerhealth/articles/detail.html?id=84>
- National Heart Centre Singapore. (n.d.). *National Heart Centre Singapore*. Retrieved from Overview – National Heart Centre Singapore: <https://www.nhcs.com.sg/about-us>
- Singapore Heart Foundation. (2022). *Singapore Heart Foundation*. Retrieved from Heart Disease Statistics: <https://www.myheart.org.sg/health/heart-disease-statistics/>
- Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., . . . Løchen, M.-L. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice . *EAS Updates*, 207-274.
- Qian, X., Li, Y., Zhang, X., Guo, H., He, J., & Wang, X. (2022). A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study . *Front. Cardiovasc. Med.* .
- Liu, S., Li, Y., Zeng, X., Wang, H., Yin, P., Wang, L., . . . Liu, J. (2019). Burden of Cardiovascular Diseases in China, 1990-2016: Findings From the 2016 Global Burden of Disease Study . *JAMA Cardiol*, 342-352.
- Perret-Guillaume, C., Joly, L., & Benetos, A. (2009). Heart rate as a risk factor for cardiovascular disease . *Prog Cardiovasc Dis.* , 6-10.
- Carauna, C. (2018 , December 14). *SciDev.Net*. Retrieved from Lifestyle diseases swamp Asia's healthcare systems : <https://www.scidev.net/asia-pacific/news/lifestyle-diseases-swamp-asia-s-healthcare-systems/>
- Singapore General Hospital. (n.d.). *Singapore General Hospital*. Retrieved from Chest Pain: <https://www.sgh.com.sg/patient-care/conditions-treatments/heart-chest-pain>
- Beckerman, J., Yamazaki, T., Myers, J., Boyle, C., Chun, S., Wang, P., & Froelicher, V. (2005). T-wave abnormalities are a better predictor of cardiovascular mortality than ST depression on the resting electrocardiogram . *Ann Noninvasive Electrocardiol*, 146-151.
- Boyd, C. R., Tolson, M. A., & Copes, W. S. (1987). Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score . *J Trauma*, 370-378.
- Yates, T., Zaccardi, F., Dhalwani, N. N., Davies, M. J., Bakrania, K., Celis-Morales, C. A., . . . Khunti, K. (2017). Association of walking pace and handgrip strength with all-cause, cardiovascular, and cancer mortality: a UK Biobank observational study . *European Heart Journal*, 3232-3240.
- Vazquez, J. P., Bouzas-Mosquera, A., Rivadulla-Varela, C., Barbeito-Caamano, C., & Vazquez-Rodriguez, J. (2021). Time to step up 4 flights of stairs gives relevant information on

- exercise testing performance and results. *European Heart Journal - Cardiovascular Imaging*.
- Kumar, A., Siddharth, V., Singh, S. I., & Narang, R. (2022). Cost analysis of treating cardiovascular diseases in a super-specialty hospital . *PLoS One*.
- Brown, C. F., & Oldridge, N. B. (1985). Exercise-induced angina in the cold. *Med Sci Sports Exerc*, 607-12.
- Rodgers, J. L., Jones, J., Bolleddu, S. L., Vanthenapalli, S., Rodgers, L. E., Shah, K., . . . Panguluri, S. K. (2019). Cardiovascular Risks Associated with Gender and Aging . *J Cardiovasc Dev Dis*, 19.
- Kisling, L. A., & Das, J. M. (2022, May 8). *StatPearls [Internet]*. Retrieved from Prevention Strategies: <https://www.ncbi.nlm.nih.gov/books/NBK537222/>
- American Heart Association. (2021, March 9). *American Heart Association*. Retrieved from Target Heart Rates Chart : <https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates>
- Knowles, J. W., & Ashley, E. A. (2018). Cardiovascular disease: The rise of the genetic risk score . *PLoS Med.*, 15.
- Gilman, M. W. (2015). Primordial Prevention of Cardiovascular Disease. *Circulation*, 599-601.
- National Heart Centre of Singapore. (2021, November 24). *Singhealth Academy*. Retrieved from Predicting Risk of Coronary Artery Disease: <https://www.singhealthacademy.edu.sg/residency/news/murmurs/predicting-risk-of-coronary-artery-disease>

Appendices

[Appendix A: Stage 1 – Data Exploration](#)

[Appendix B: Stage 1 – 17 Questions Individuals Can Answer](#)

[Appendix C: Stage 1 – What at-risk Individuals can do by themselves](#)

[Appendix D: Stage 1 Logistic Regression Models](#)

[Appendix E: Stage 1 Gradient Boosting Classifier \(GBC\)](#)

[Appendix F: Stage 1 Random Forest Classifier](#)

[Appendix G: Stage 2 – Data Exploration](#)

[Appendix H: Stage 2 CART and Random Forests](#)

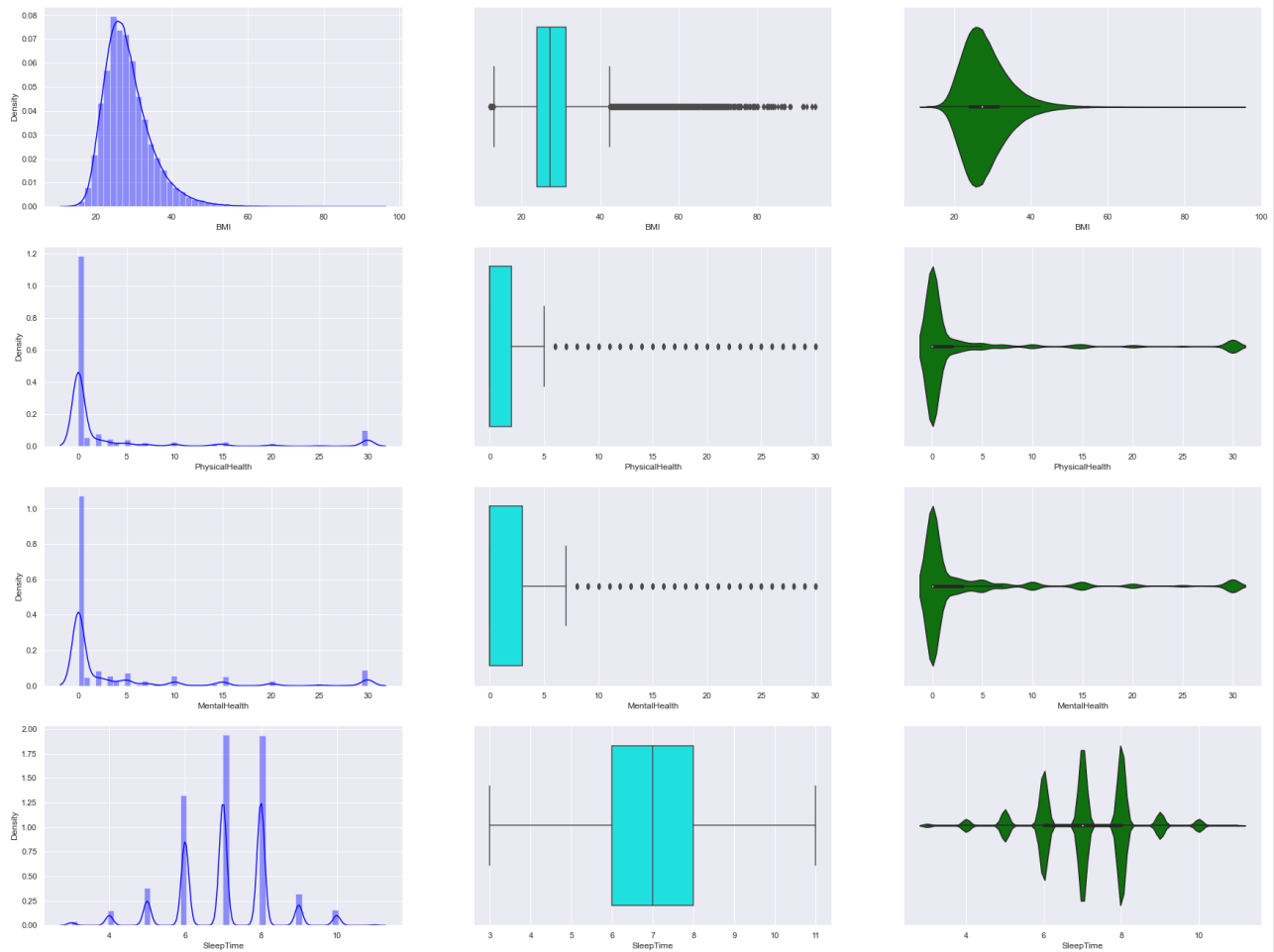
[Appendix I: Stage 2 CART and Random Forests Full Performance Results](#)

Appendix A: Stage 1 – Data Exploration

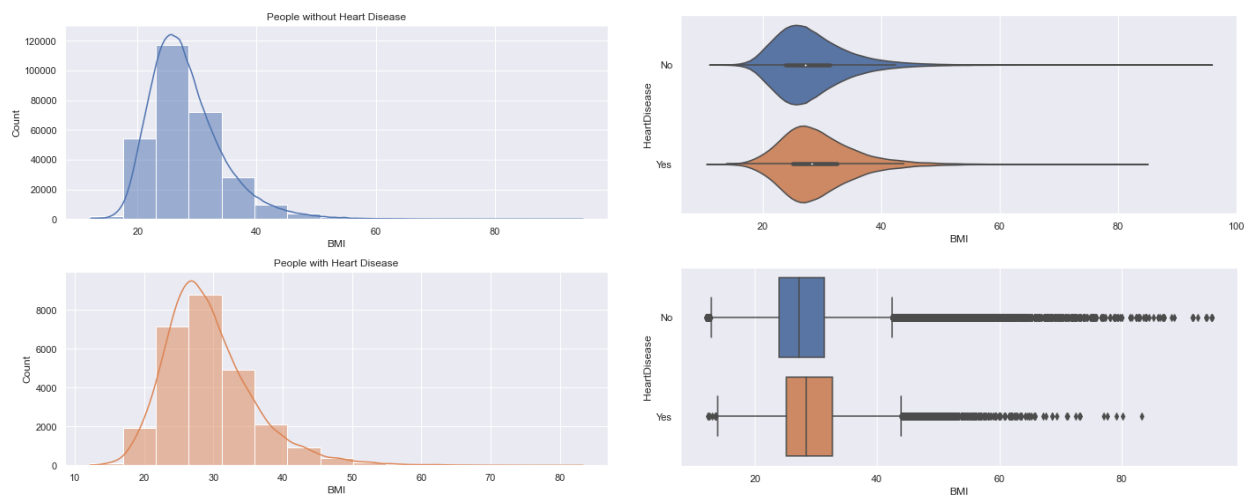
Continuous Variables

Distribution

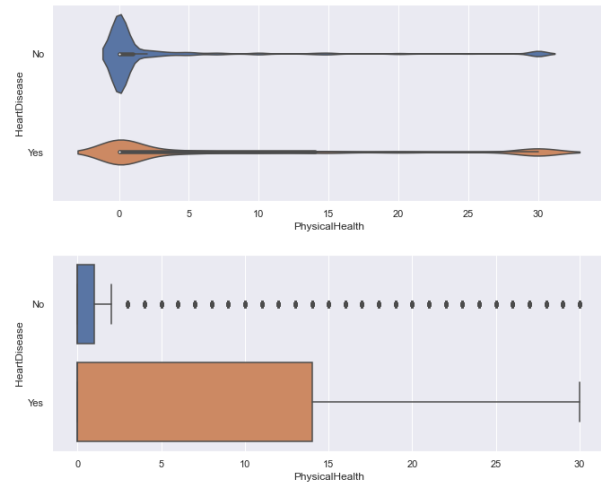
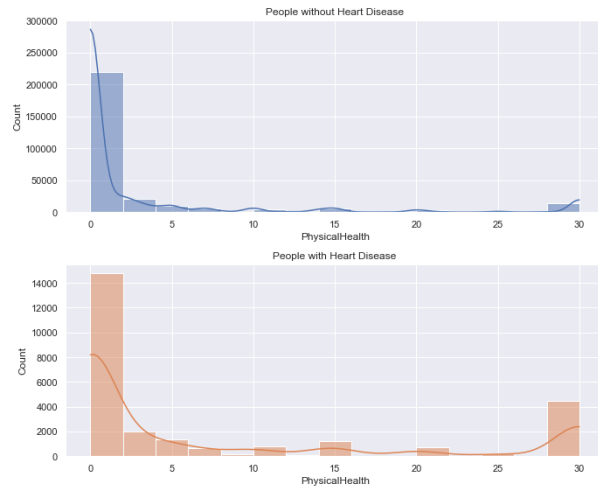
Distribution of Continuous Variables



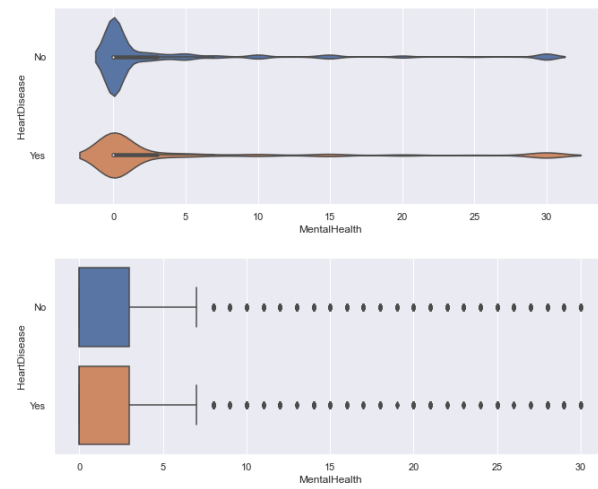
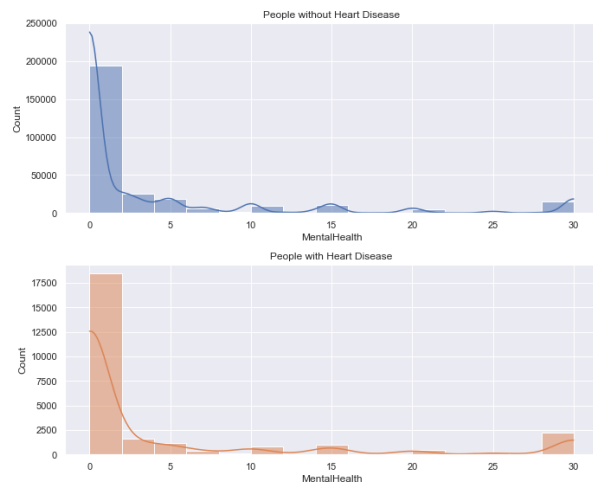
BMI



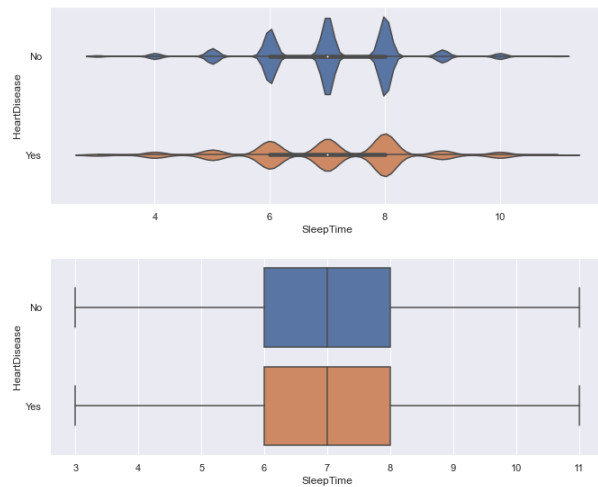
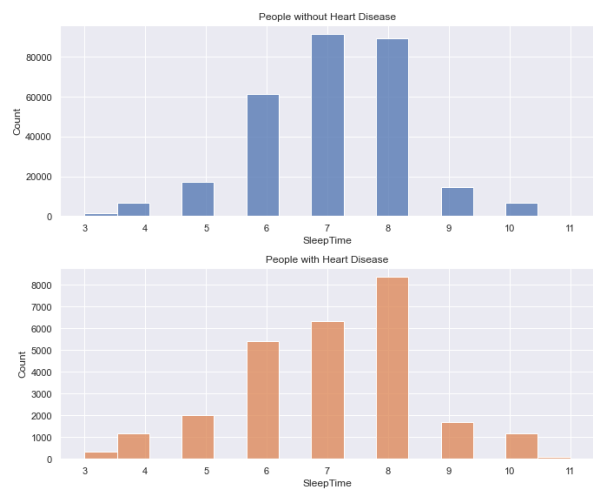
PhysicalHealth



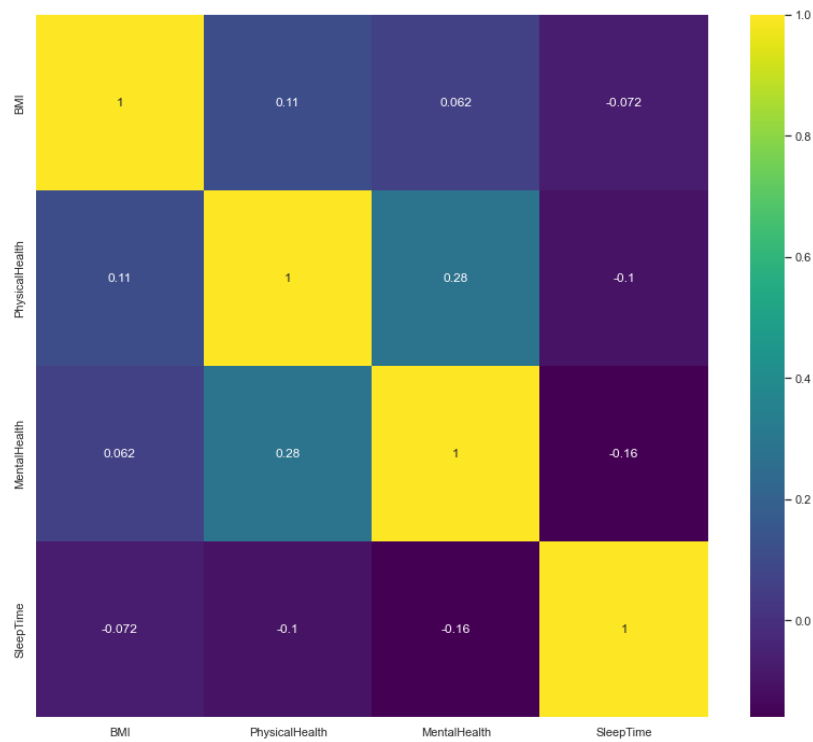
MentalHealth



SleepTime

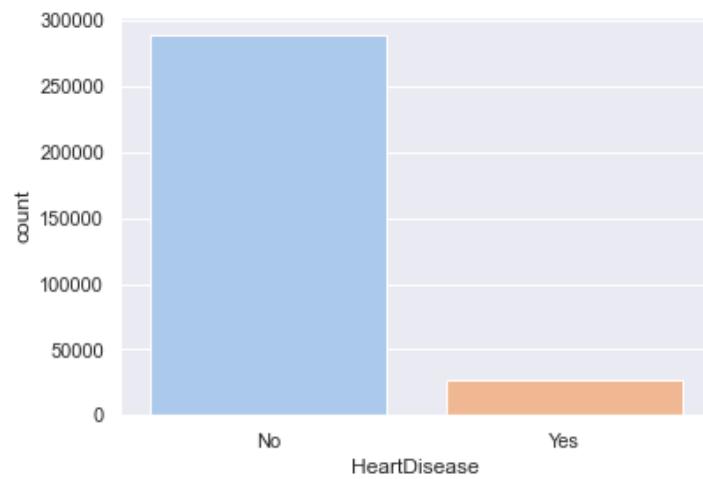


Correlation

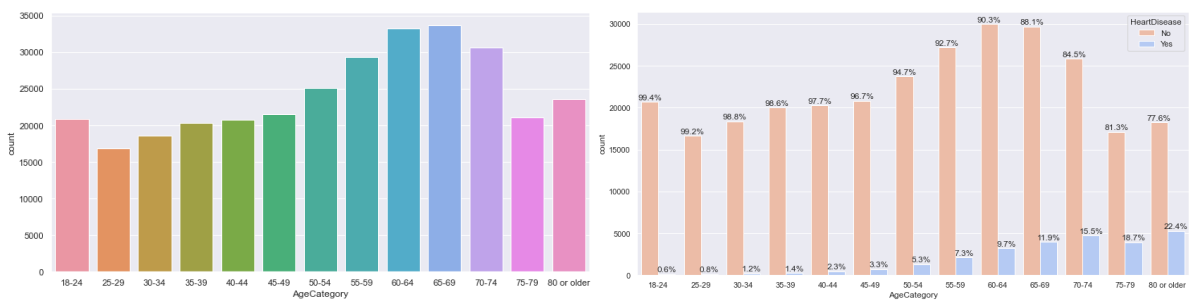


Categorical Variables

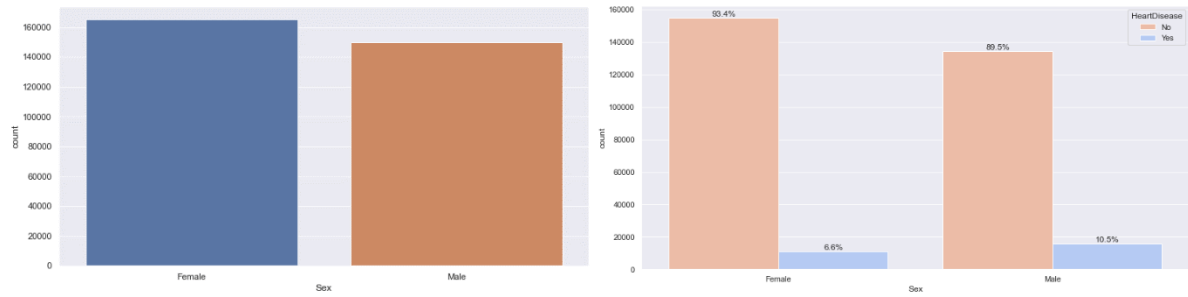
HeartDisease



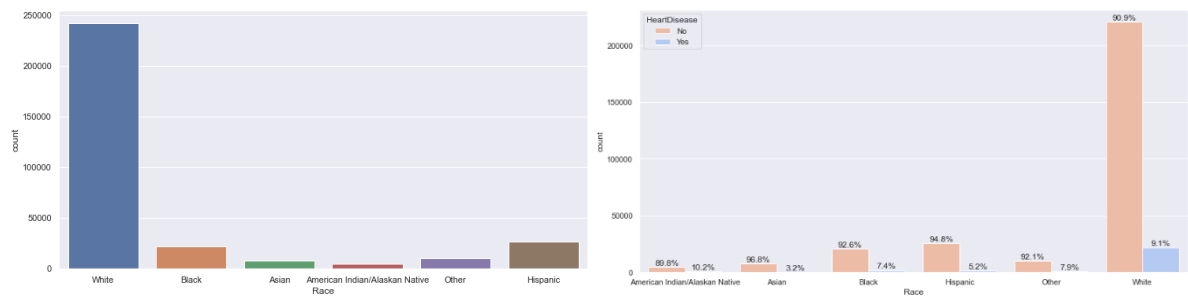
AgeCategory



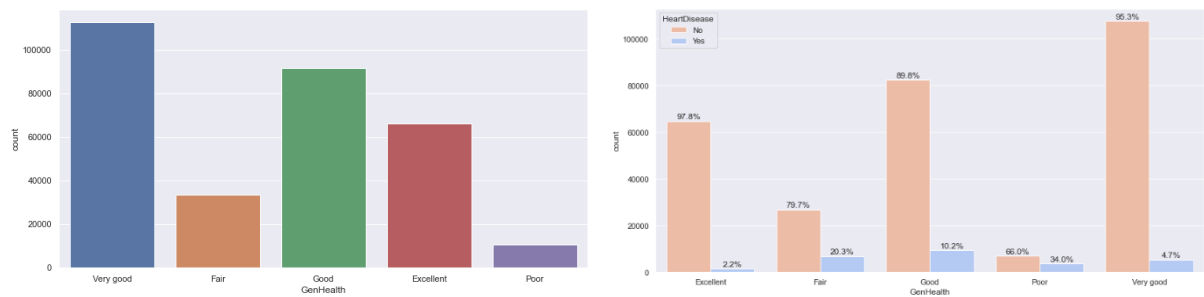
Sex



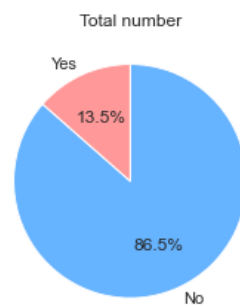
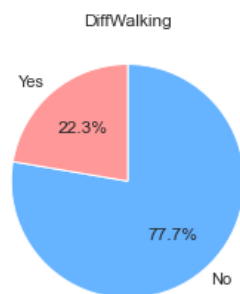
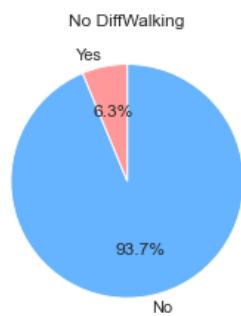
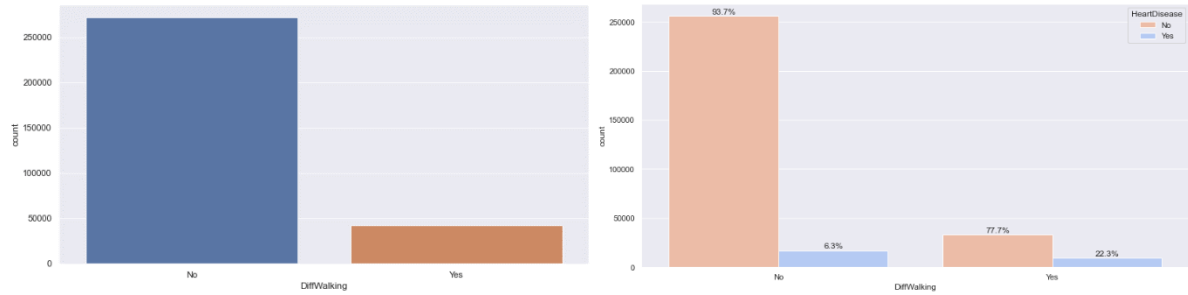
Race



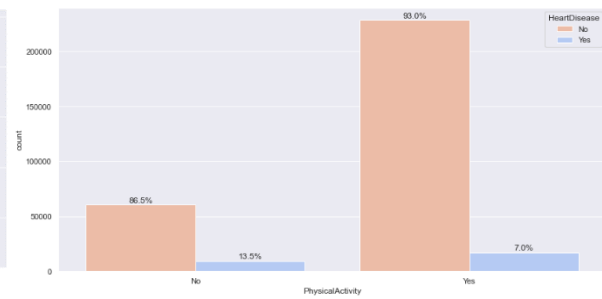
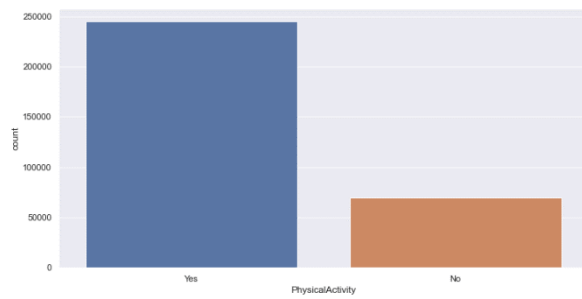
GenHealth



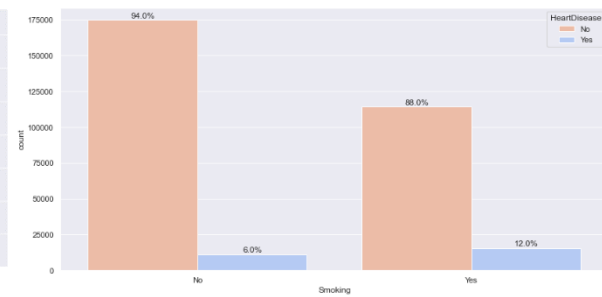
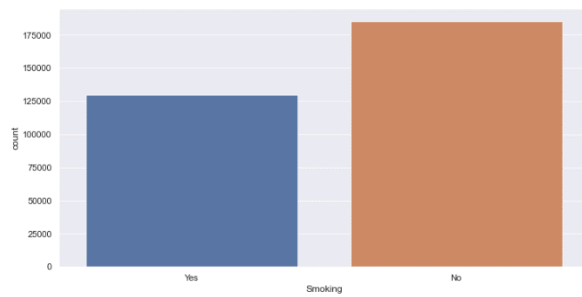
DiffWalking



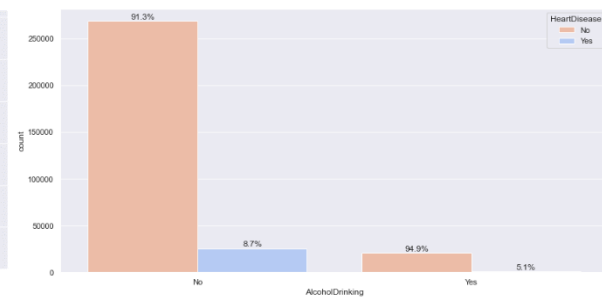
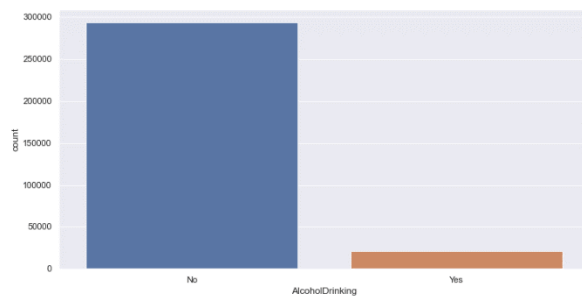
PhysicalActivity



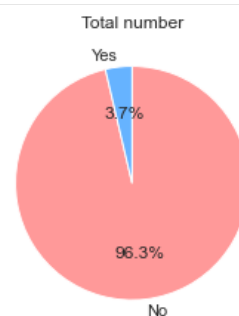
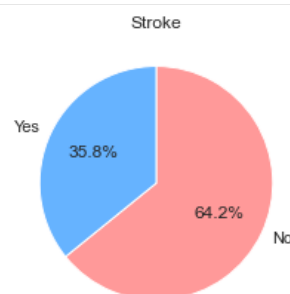
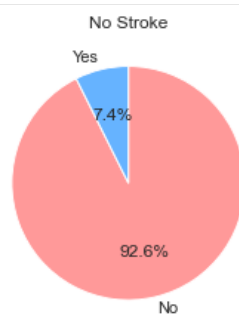
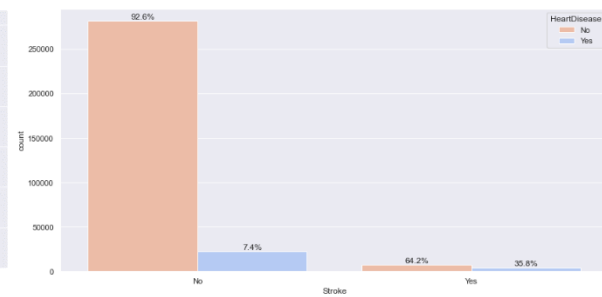
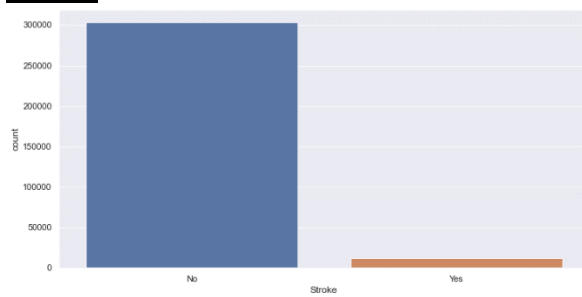
Smoking



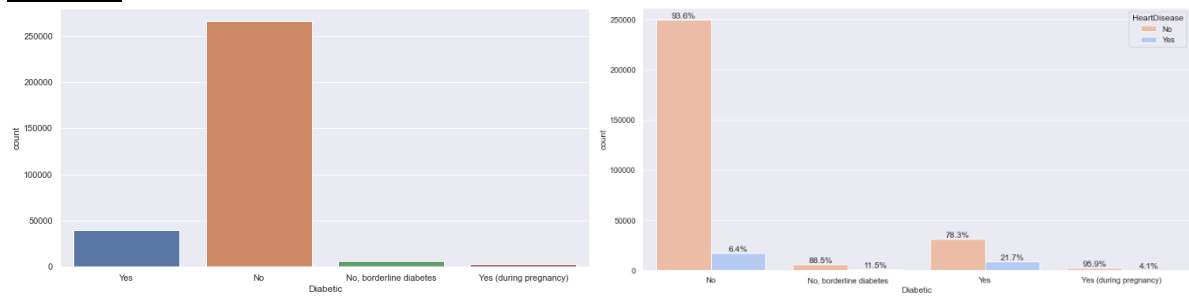
AlcoholDrinking



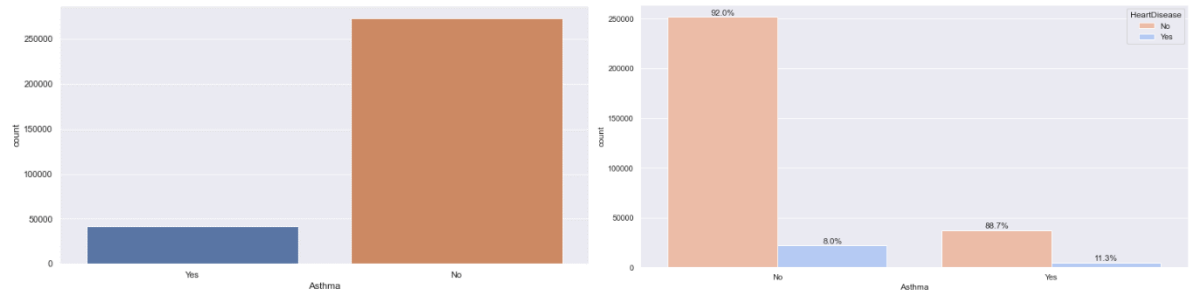
Stroke



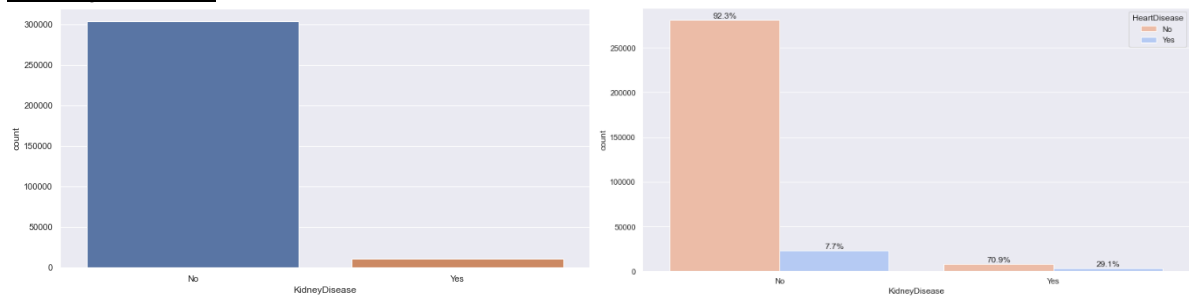
Diabetic



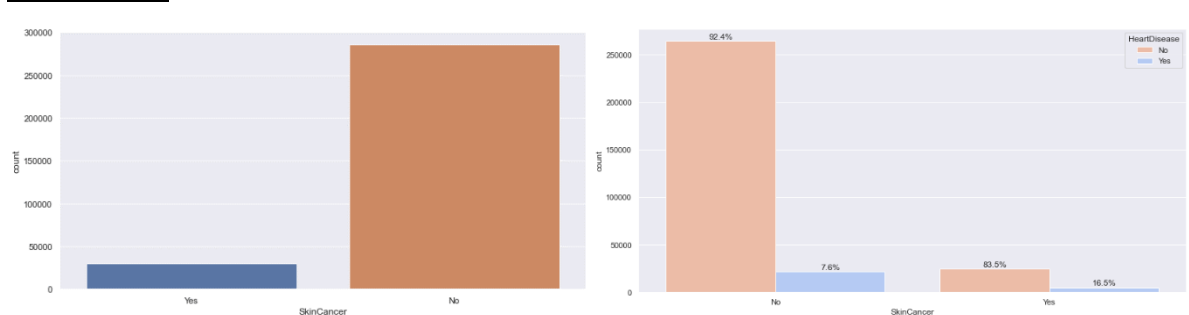
Asthma



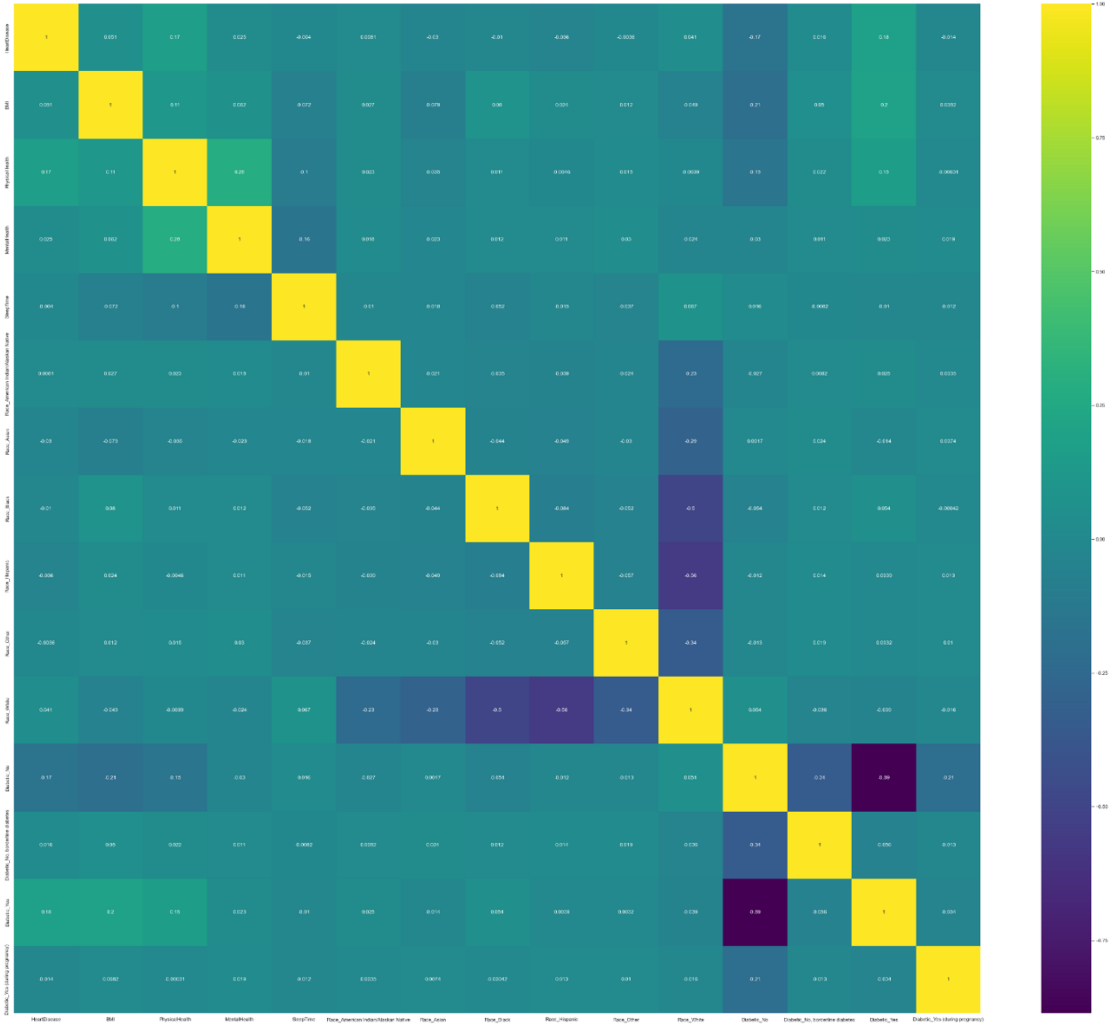
KidneyDisease

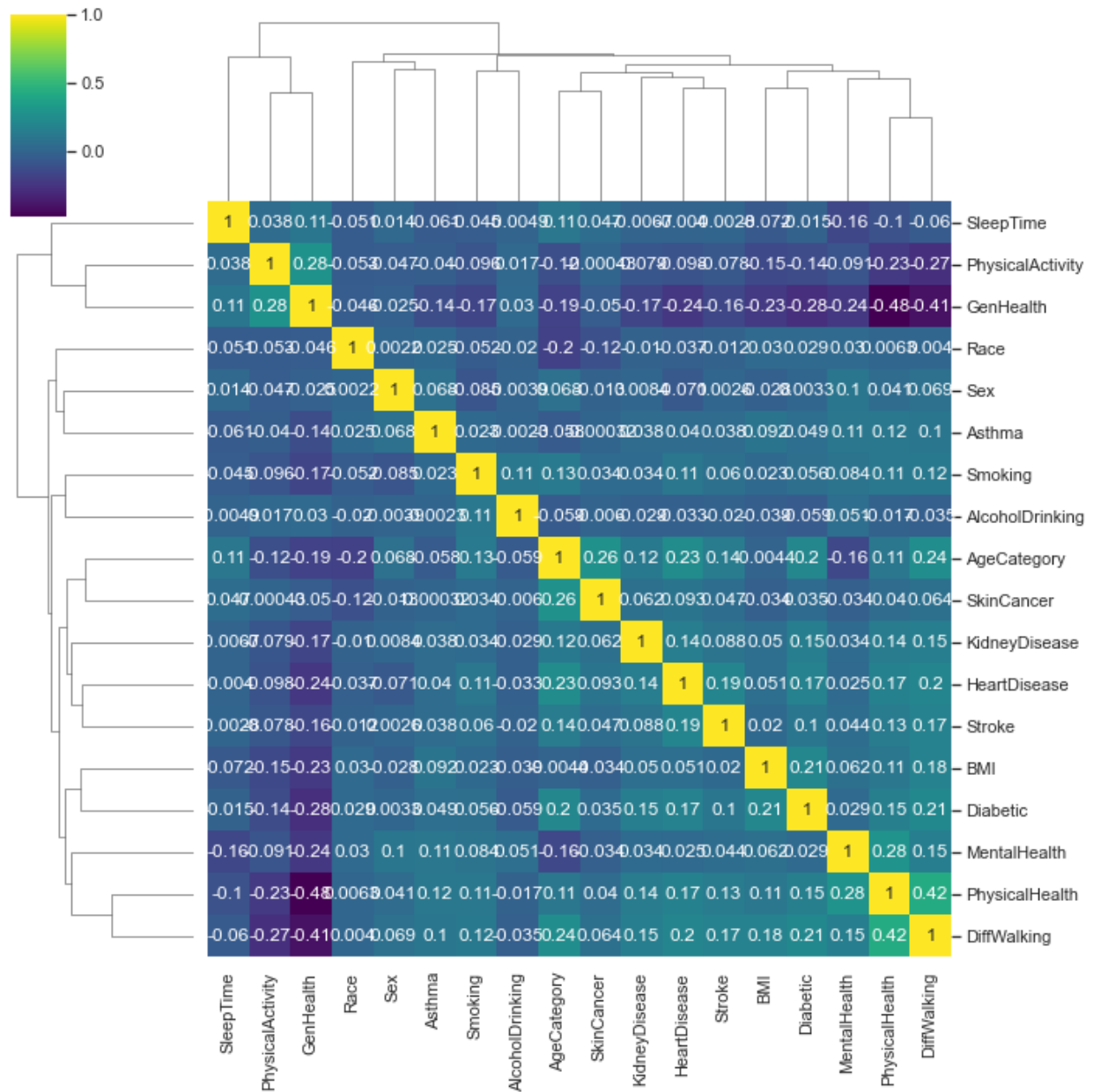


SkinCancer



Overall Correlation





[Click here to go back to the main report content](#)

Appendix B: Stage 1 – 17 Questions Individuals Can Answer

1. What is your BMI?
2. (Ever told) (you had) a stroke?
3. What is your gender?
4. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 are you in good physical condition?
5. Now thinking about your mental health, for how many days during the past 30 was your mental health not good?
6. Would you say that in general your health is excellent, very good, good, fair, or poor?
7. Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
8. Have more than 7 drinks per week?
9. (Ever told) (you had) skin cancer?
10. What is your race?
11. (Ever told) (you had) diabetes?
12. (Ever told) (you had) asthma?
13. Do you have serious difficulty walking or climbing stairs?
14. Which of the fourteen-level age category do you fall into?
15. Would you say that in general your health is good?
16. On average, how many hours of sleep do you get in a 24-hour period?
17. Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?

[Click here to go back to the main report content](#)

Appendix C: Stage 1 – What at-risk Individuals can do by themselves



(National Heart Centre of Singapore, 2021)

[Click here to go back to the main report content](#)

Appendix D: Stage 1 Logistic Regression Models

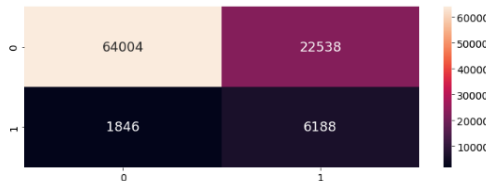
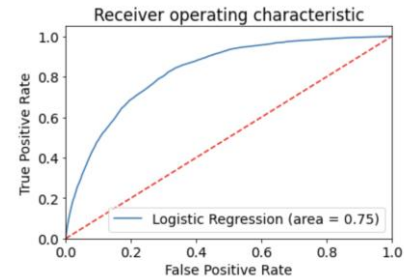
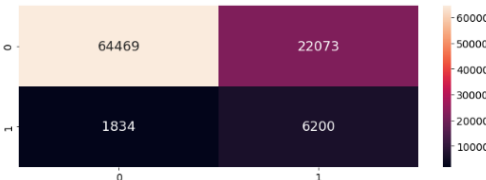
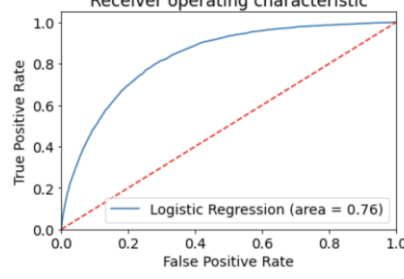
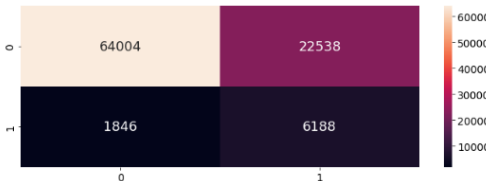
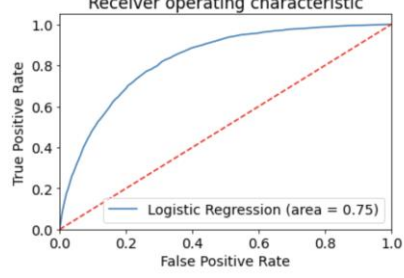
Model Training Procedures

Based on REF, 3 models were trained and logreg_m21 was found to be the optimal model

	logreg_m20 [Model 1]	logreg_m25 [Model 2]	logreg_m21 [Model 3]
Description	Top 20 features selected by RFE Model	Top 25 features selected by RFE Model	Top 21 features selected by RFE Model
Findings	<ul style="list-style-type: none">- Pseudo r-squared is 0.297, which is between 0.2 and 0.4, indicating excellent fit- p-values of all 20 important variables are close to 0, signifying that they are all statistically significant	<ul style="list-style-type: none">- Pseudo r-squared is 0.305, which is between 0.2 and 0.4, indicating excellent fit- p-values of 4 variables are NaN (Stroke_No, Stroke_Yes, Sex_Female, Sex_Male), signifying that they are all not statistically significant	<ul style="list-style-type: none">- Pseudo r-squared is 0.297, between 0.2 and 0.4, indicating excellent fit- p-values of all 21 important variables are close to 0, signifying that they are all statistically significant (better than logreg_m25)

[Click here to go back to the main report content](#)

Performance of 3 Logistic Regression models

	Confusion Matrix	Accuracy	Precision, Recall, F1-score and Support	ROC-AUC Curve
Model 1		Overall Accuracy: <u>74.22%</u> True Positive Rate: <u>21.54%</u> False Negative Rate: <u>78.46%</u>	<pre>===== logreg_m20 ===== precision recall f1-score support 0 0.97 0.74 0.84 86542 1 0.22 0.77 0.34 8034 accuracy 0.74 94576 macro avg 0.59 94576 weighted avg 0.91 94576</pre>	
Model 2		Overall Accuracy: <u>74.72%</u> True Positive Rate: <u>21.93%</u> False Negative Rate: <u>78.07%</u>	<pre>===== logreg_m25 ===== precision recall f1-score support 0 0.97 0.74 0.84 86542 1 0.22 0.77 0.34 8034 accuracy 0.75 94576 macro avg 0.60 94576 weighted avg 0.91 94576</pre>	
Model 3		Overall Accuracy: <u>74.22%</u> True Positive Rate: <u>21.54%</u> False Negative Rate: <u>78.46%</u>	<pre>===== logreg_m21 ===== precision recall f1-score support 0 0.97 0.74 0.84 86542 1 0.22 0.77 0.34 8034 accuracy 0.74 94576 macro avg 0.59 94576 weighted avg 0.91 94576</pre>	
Findings	The accuracy is quite decent for all models. However, the false negative rate for all models is very bad.		For all models, the weighted average of precision and recall are quite decent. The f1-score is also quite decent (around 0.8), which shows that the model performs quite well.	<ul style="list-style-type: none">- The ROC AUC score (~0.75) indicates that all the 3 models are good classifier for the dataset- The best ROC AUC score belongs to logreg_m25- The model logreg_m21 has a ROC AUC score of 0.75 which is almost the same to that of logreg_25, even though it has less variables.- Taking the model complexity into account, logreg_m21 is the better model of the 3

Appendix E: Stage 1 Gradient Boosting Classifier (GBC)

Performance of Gradient Boosting Classifier Model (gbc_m1)

	Train Dataset	Test Dataset
Classification Accuracy	90.81%	87.54%
True Positive Rate	88.64%	30.84%
True Negative Rate	92.98%	92.81%
False Positive Rate	10.88%	6.18%
False Negative Rate	8.71%	72.28%

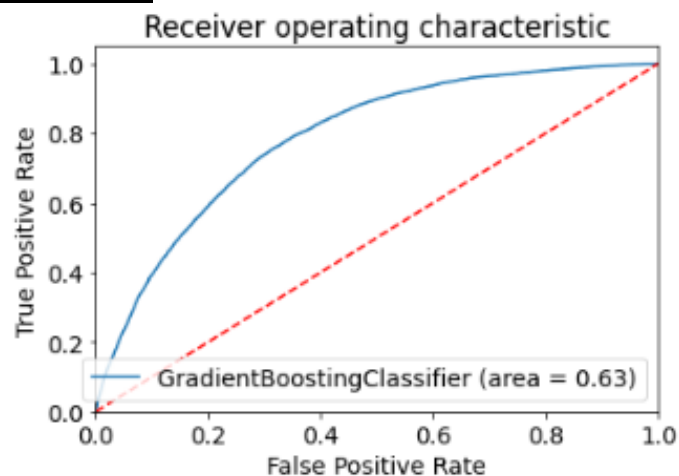
- High classification accuracy on test dataset (87.54%)
- However, low true positive rate and high false negative rate on the test dataset, which is very bad.

Precision, Recall, F1-score and Support

GBC has excellent weighted average for precision, recall, and f1-score, which shows that the model performs well, and better than logistic regression models. However, it has very bad precision, recall, and f1-score for predicting a '1'.

	precision	recall	f1-score	support
0	0.94	0.92	0.93	86542
1	0.28	0.35	0.31	8034
accuracy			0.87	94576
macro avg	0.61	0.63	0.62	94576
weighted avg	0.88	0.87	0.87	94576

ROC Curve



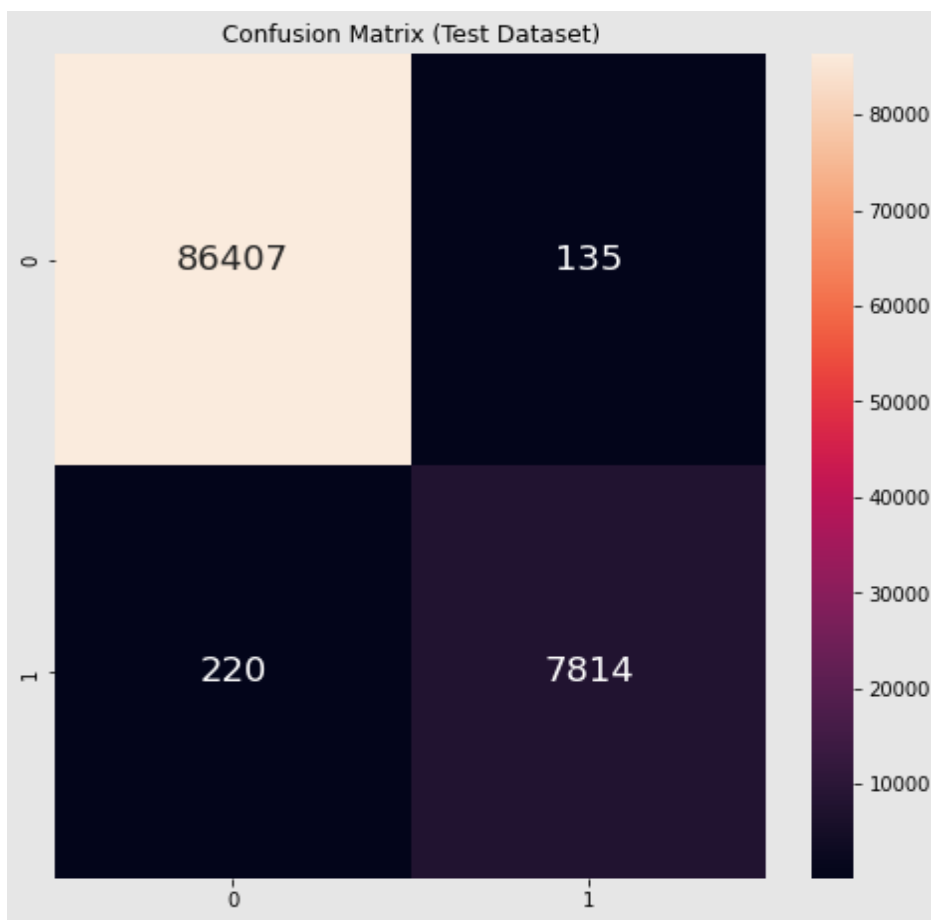
ROC AUC score of 0.63, which is worse than that of Logistic Regression's.

Appendix F: Stage 1 Random Forest Classifier

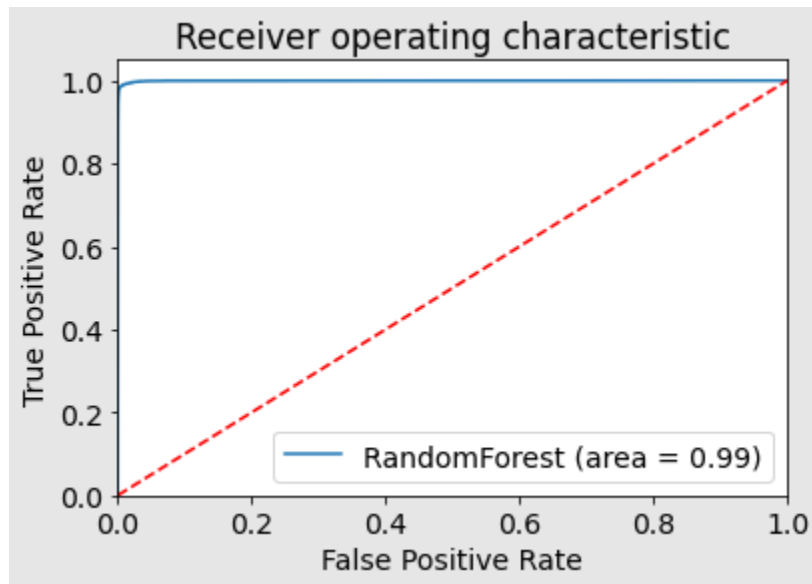
Performance of Random Forest Model (random_forest_m1)

	Test Dataset
Classification Accuracy	99.62%
True Positive Rate	98.30%
True Negative Rate	99.75%
False Positive Rate	0.25%
False Negative Rate	1.70%

Confusion Matrix



ROC Curve



ROC AUC score of 0.99 which is the best among all models in stage 1.

Feature Importance

- BMI is the most important feature for the random forest classifier model

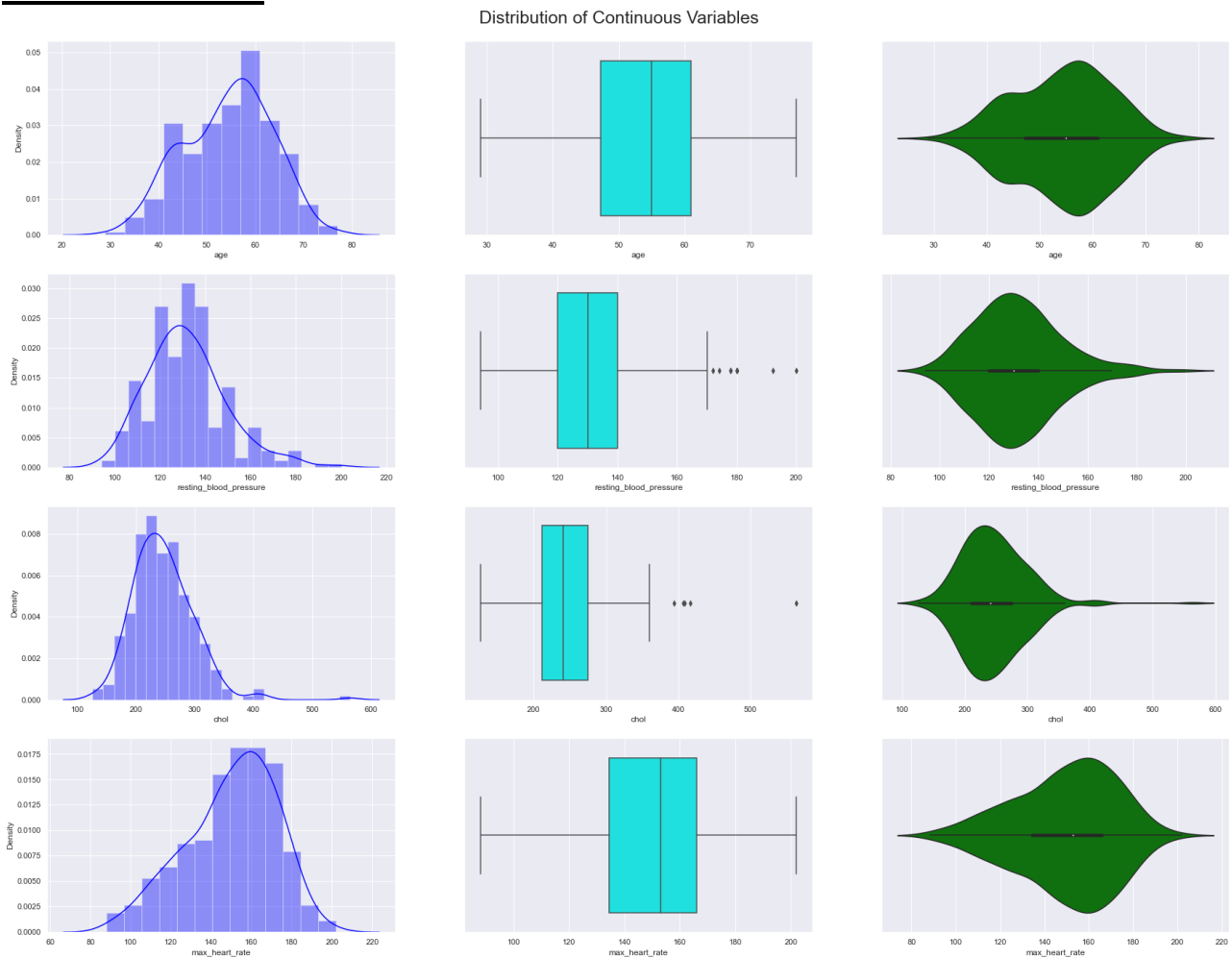
feature		importance			
0	BMI	0.123	27	PhysicalActivity_Yes	0.010
1	AgeCategory_80 or older	0.056	28	SkinCancer_No	0.009
2	SleepTime	0.054	29	SkinCancer_Yes	0.009
3	Diabetic_Yes	0.041	30	AgeCategory_50-54	0.009
4	Race_White	0.039	31	Asthma_Yes	0.008
5	DiffWalking_No	0.038	32	Asthma_No	0.008
6	AgeCategory_70-74	0.038	33	AgeCategory_45-49	0.007
7	DiffWalking_Yes	0.035	34	KidneyDisease_Yes	0.007
8	AgeCategory_75-79	0.035	35	KidneyDisease_No	0.007
9	GenHealth_Excellent	0.034	36	Race_Hispanic	0.006
10	Diabetic_No	0.030	37	AgeCategory_40-44	0.006
11	Sex_Female	0.030	38	AgeCategory_35-39	0.006
12	GenHealth_Very good	0.029	39	AgeCategory_30-34	0.005
13	GenHealth_Fair	0.029	40	AlcoholDrinking_No	0.005

14	Sex_Male	0.027			
15	GenHealth_Good	0.026			
16	PhysicalHealth	0.026			
17	MentalHealth	0.026			
18	Smoking_Yes	0.025	41	AlcoholDrinking_Yes	0.005
19	AgeCategory_65-69	0.025	42	Race_Black	0.005
20	Smoking_No	0.023	43	AgeCategory_25-29	0.004
21	AgeCategory_60-64	0.017	44	AgeCategory_18-24	0.004
22	Stroke_Yes	0.017	45	Diabetic_No, borderline diabetes	0.004
23	Stroke_No	0.014	46	Race_Other	0.002
24	AgeCategory_55-59	0.011	47	Diabetic_Yes (during pregnancy)	0.001
25	PhysicalActivity_No	0.010	48	Race_American Indian/Alaskan Native	0.001
26	GenHealth_Poor	0.010	49	Race_Asian	0.001

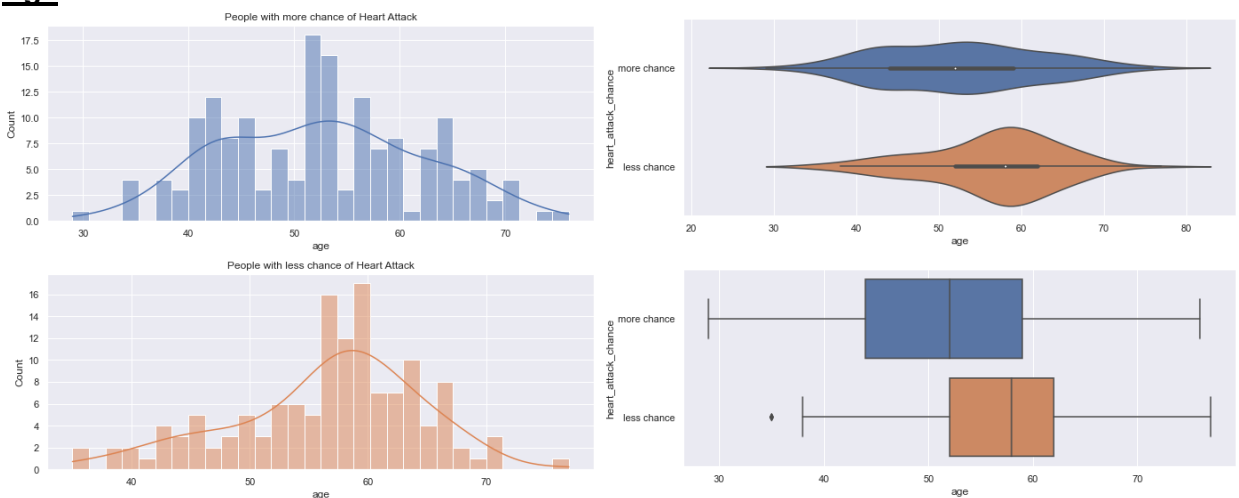
Appendix G: Stage 2 – Data Exploration

Continuous Variables

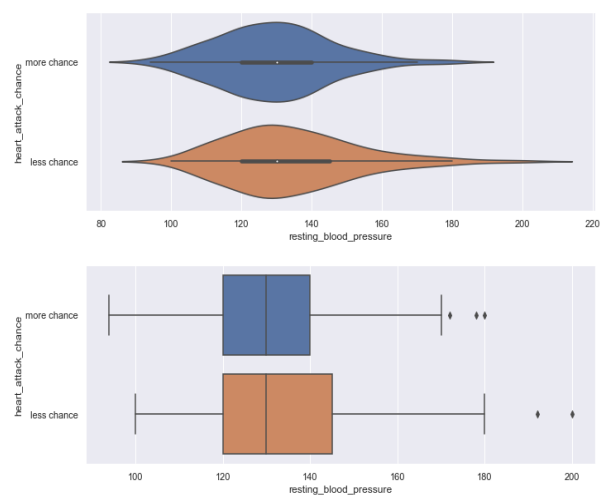
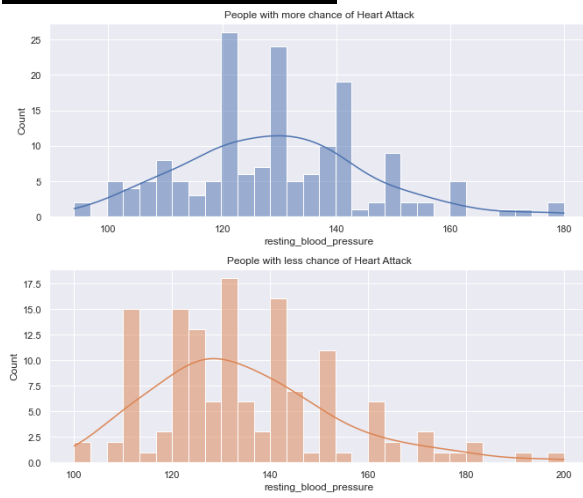
Overall Distribution



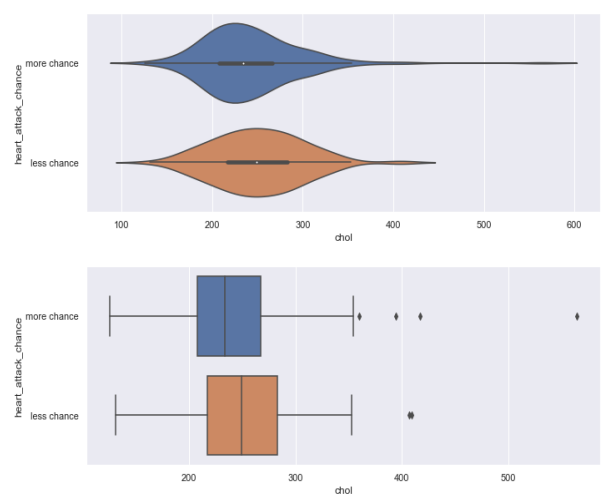
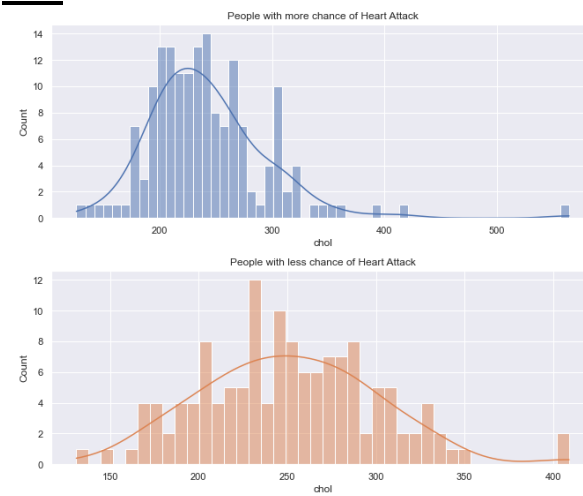
Age



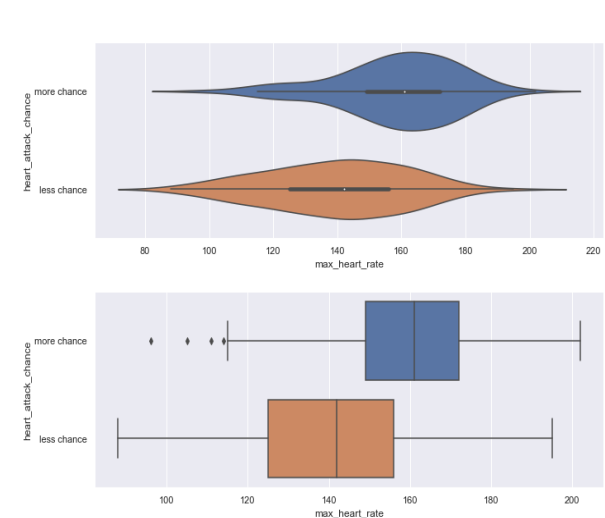
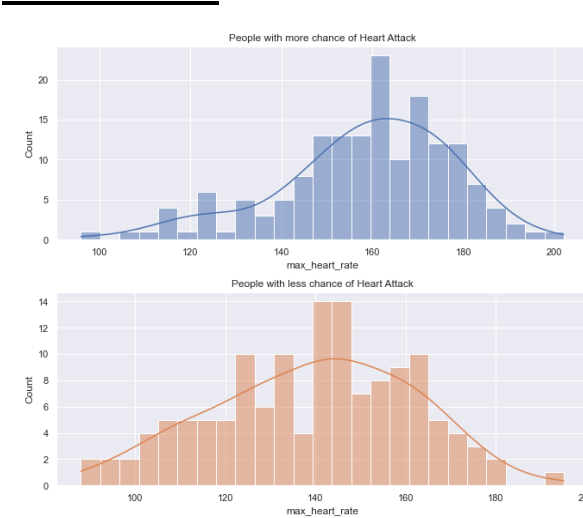
resting blood pressure



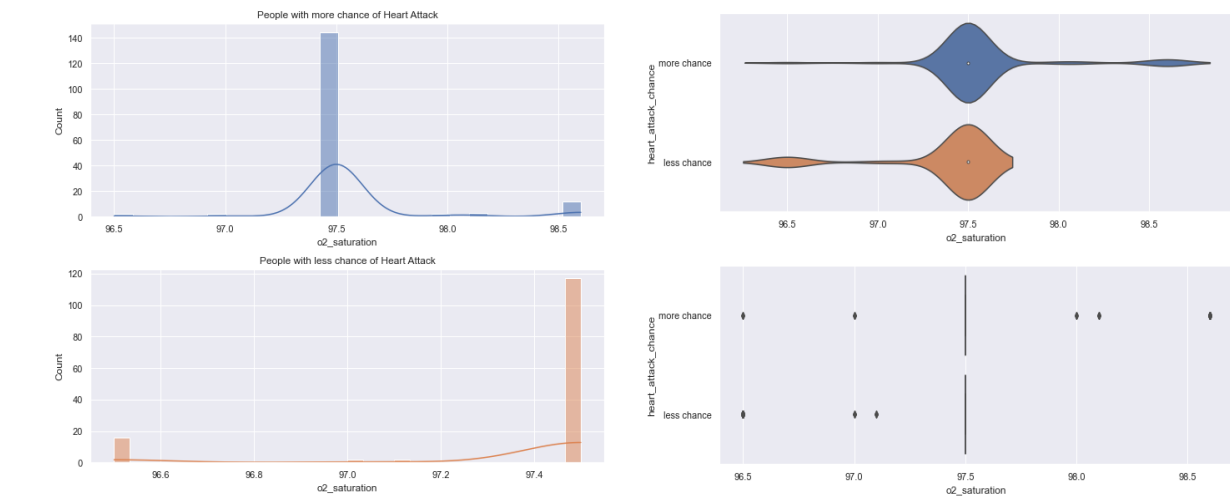
chol



max heart rate

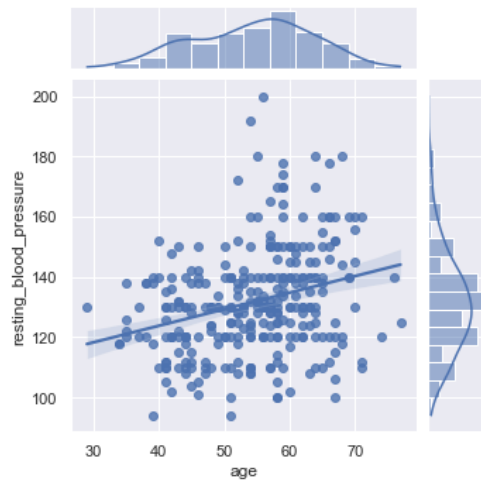


o2 saturation

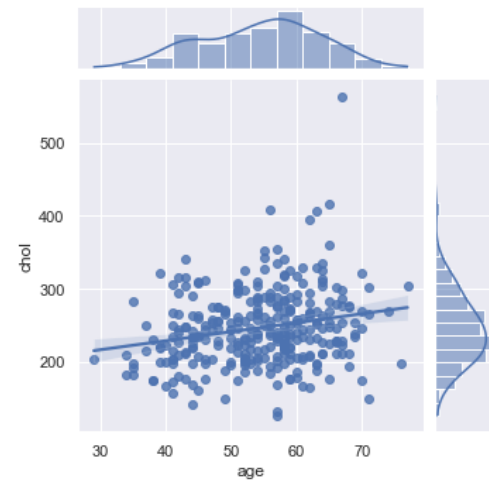


Correlation

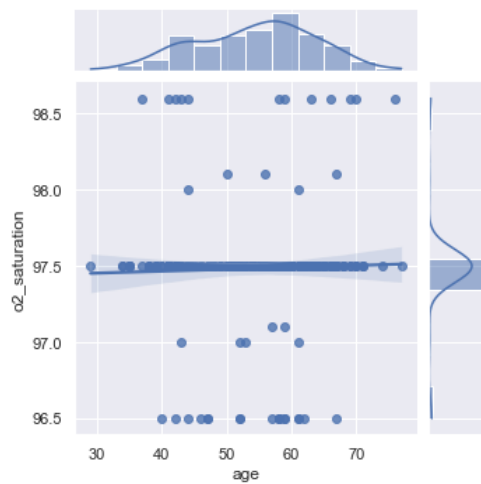
resting blood pressure VS age



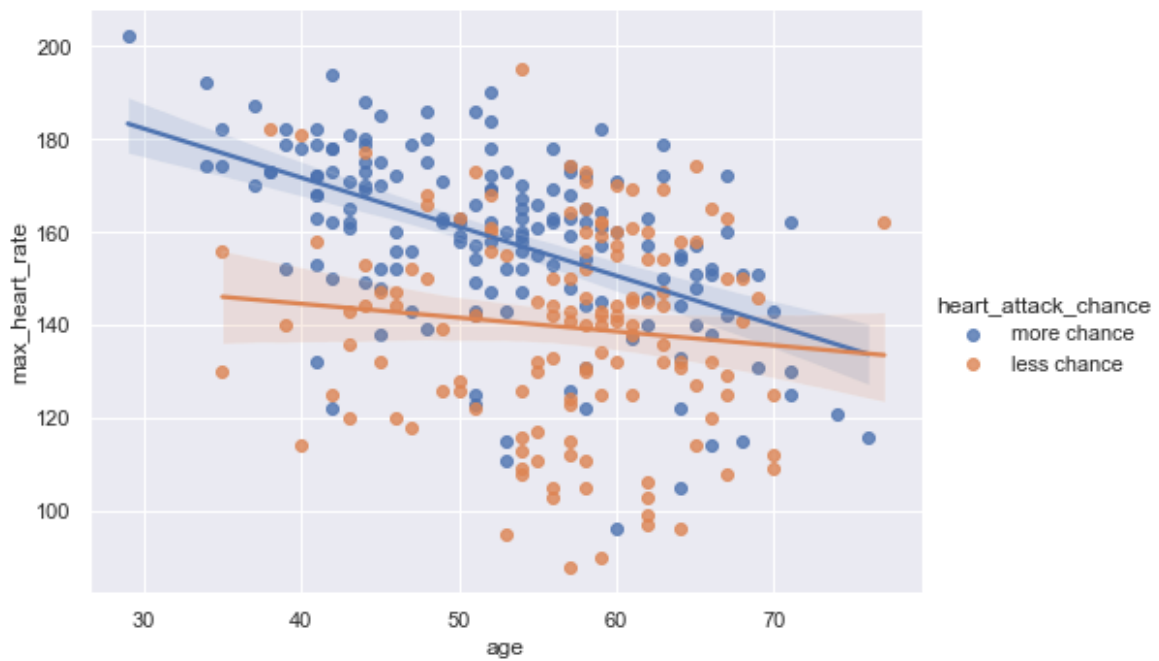
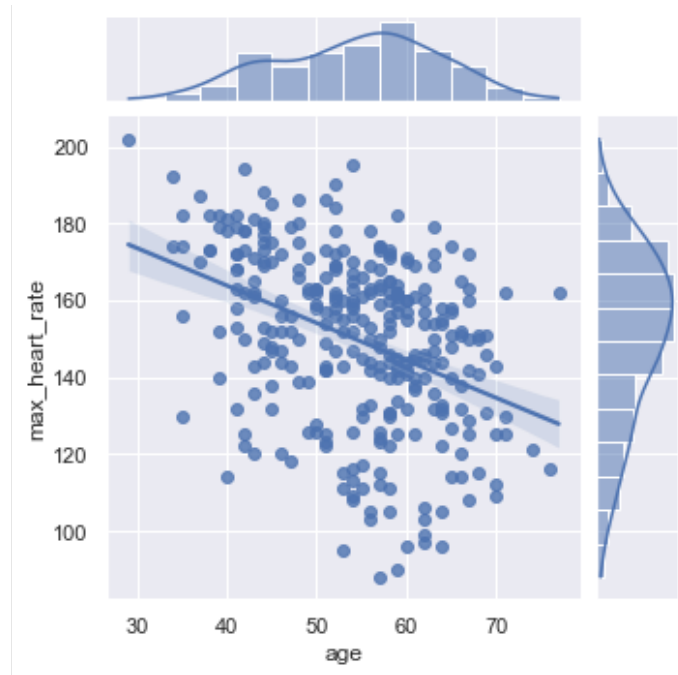
chol VS age



o2 saturation VS age

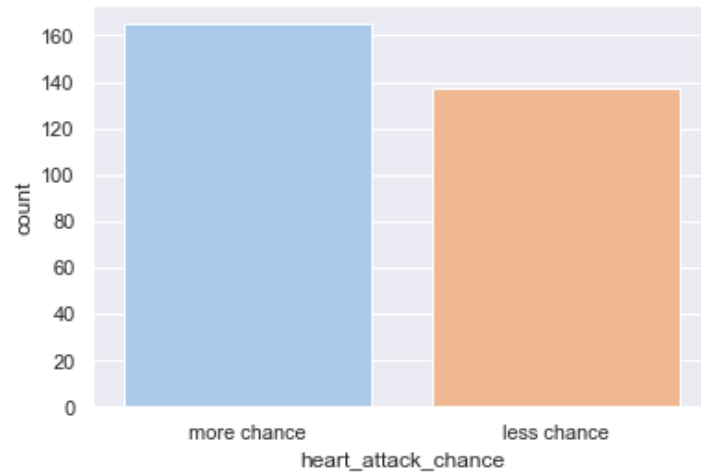


max_heart_rate VS age

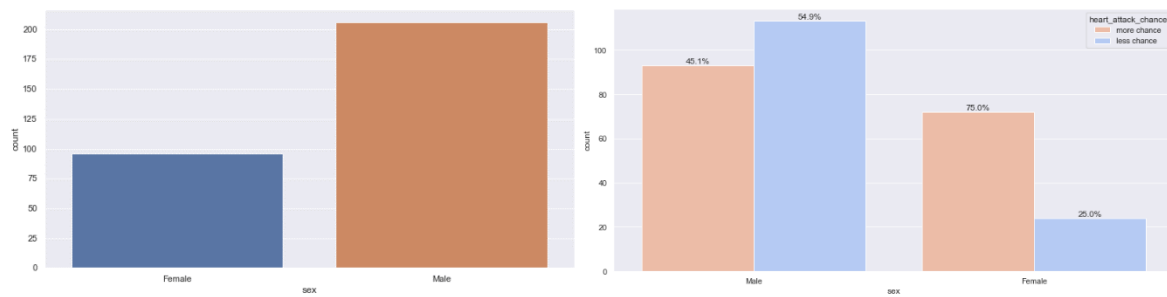


Categorical Variables

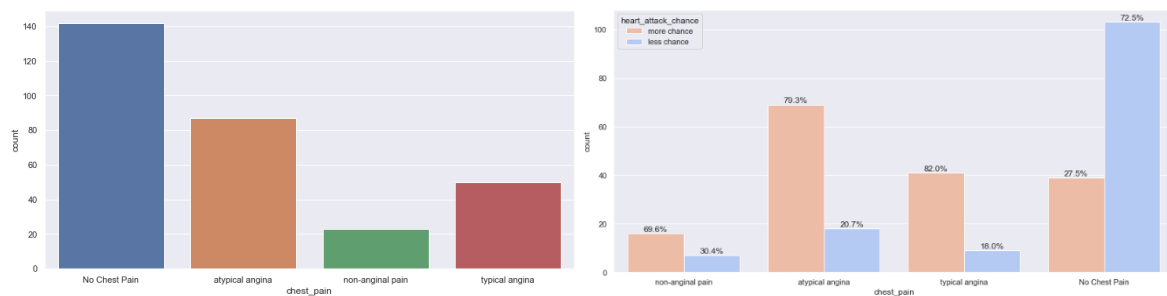
heart_attack_chance



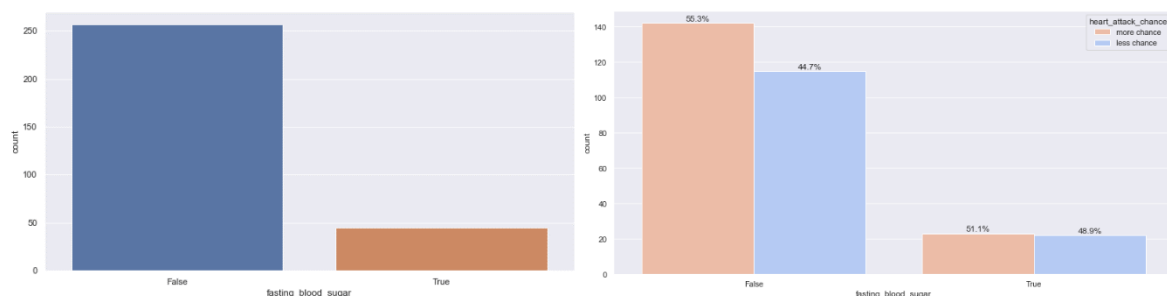
sex



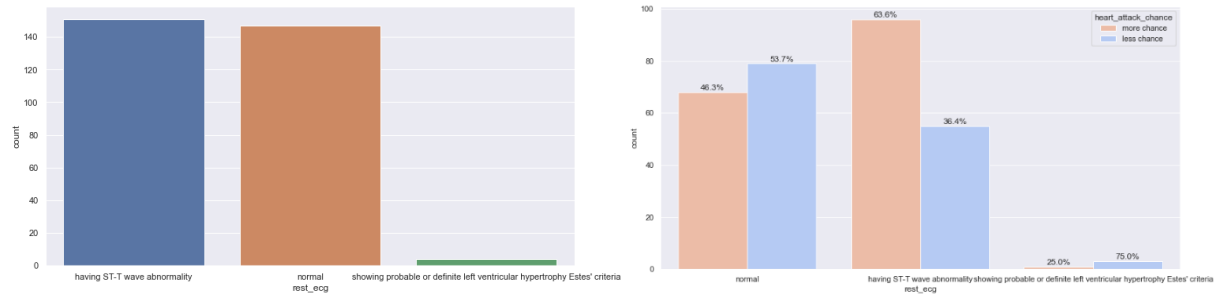
chest_pain



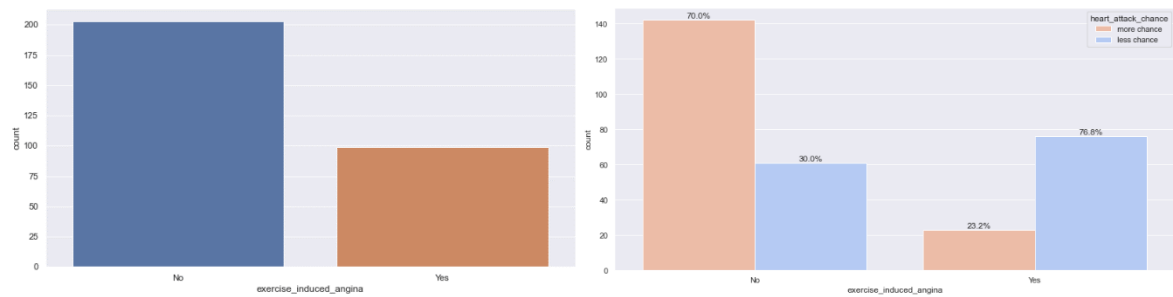
fasting_blood_sugar



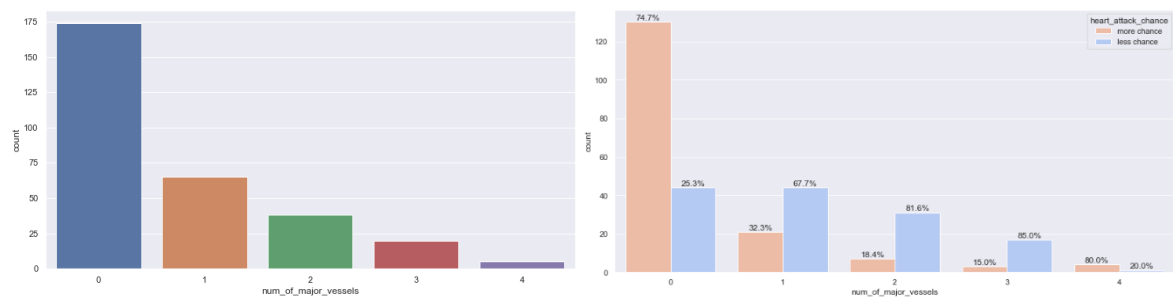
rest_ecg



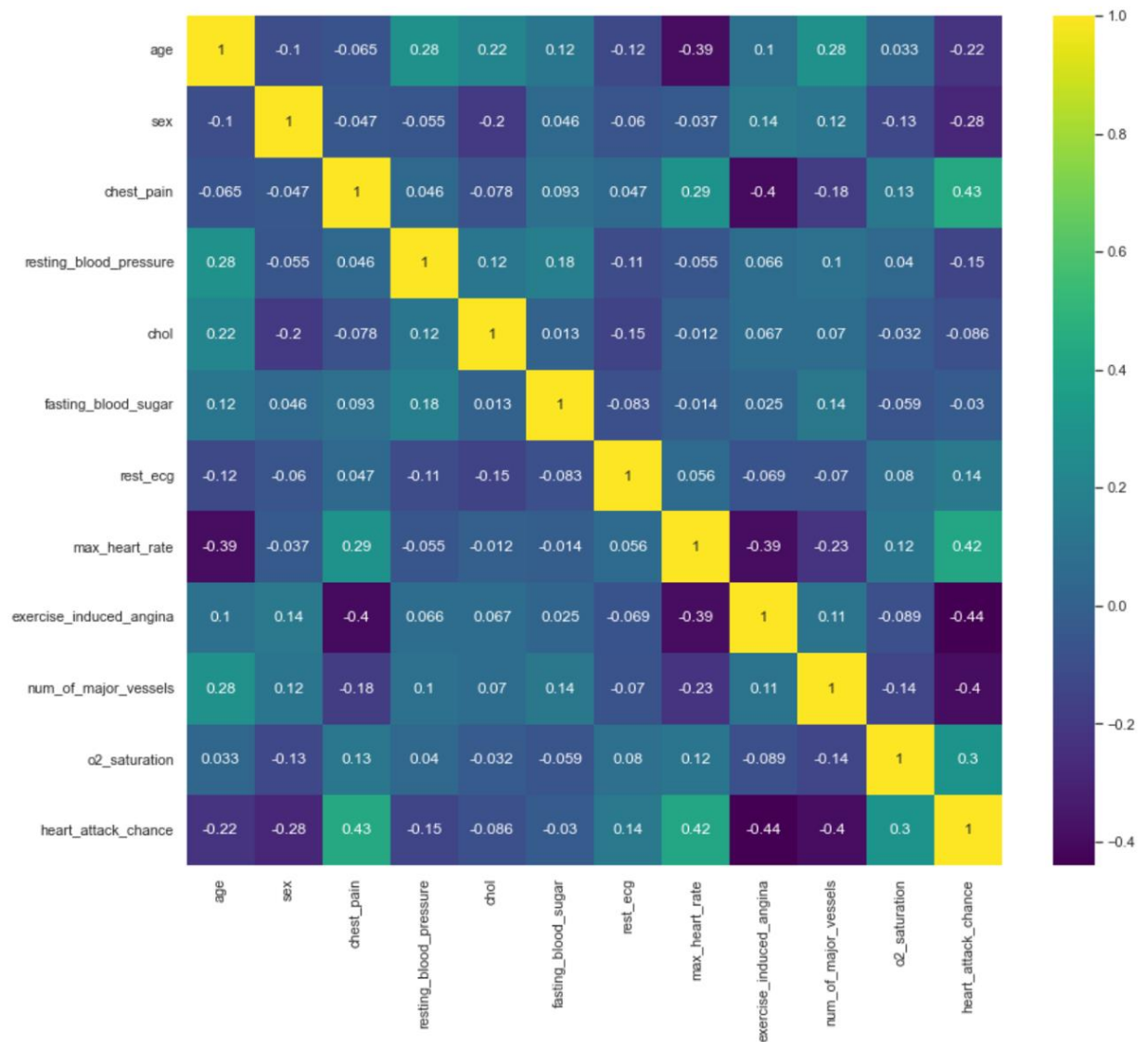
exercise_induced_angina

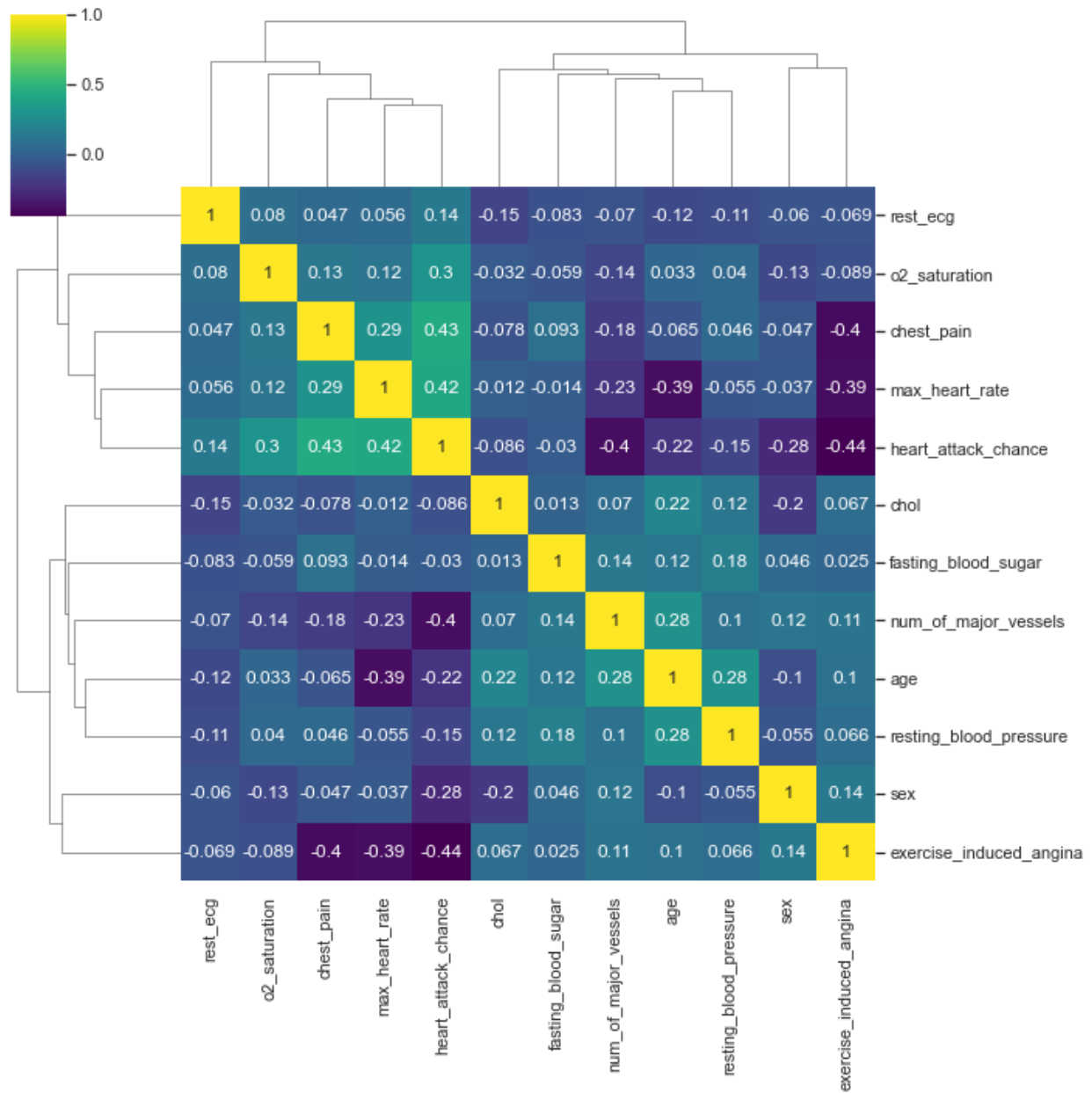


num_of_major_vesels



Overall Correlation





[Click here to go back to the main report content](#)

Appendix H: Stage 2 CART and Random Forests

CART

CART is one of the most common and powerful models in machine learning. It is inexpensive to process and transparent, different from the "black box" of neural network models, so that decision tree used for classification can be built and understood easily.

Random Forest

a) How it works

- An ensemble of decision trees
- Building multiple decision trees and merging them together to get a more accurate and stable prediction
- Random Forest adds additional randomness to the model, while growing the trees
- Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.
- Only a random subset of the features is taken into consideration by the algorithm for splitting a node

b) Important Hyperparameters

The hyperparameters in random forest are either used to increase the predictive power of the model or to make the model faster. Let's look at the hyperparameters of sklearn's built-in random forest function.

Increasing the predictive power

n_estimators hyperparameter

- Which is just the number of trees the algorithm builds before taking the maximum vote or taking the averages of predictions.
- In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.

max_features hyperparameter

- Which is the **maximum number of features** random forest considering to **split a node**.
- Sklearn provides several options, all described in the documentation.

min_sample_leaf hyperparameter

- This determines the **minimum number of leafs** required to **split an internal node**.

Increasing the model speed

n_jobs hyperparameter

- Tells the engine **how many processors** it is allowed to use.
- If it has a value of 1, it can only use one processor. A value of "-1" means that there is no limit.

random_state hyperparameter

- Makes the **model's output replicable**.
- The model will always produce the same results when it has a definite value of random_state and if it has been given the same hyperparameters and the same training data.

oob_score hyperparameter (also called oob sampling)

- Which is a random forest **cross-validation method**.
- In this sampling, about one-third of the data is not used to train the model and can be used to evaluate its performance. These samples are called the out-of-bag samples. It's very similar

to the leave-one-out-cross-validation method, but almost no additional computational burden goes along with it.

Advantages and Disadvantages of Random Forest compared to CART

a) Advantages

- Versatility, as it can be used for both regression and classification, and it's easy to view the relative importance it assigns to the input features
- Straight forward and easy to understand hyperparameters
- No overfitting if there are enough trees in the forest

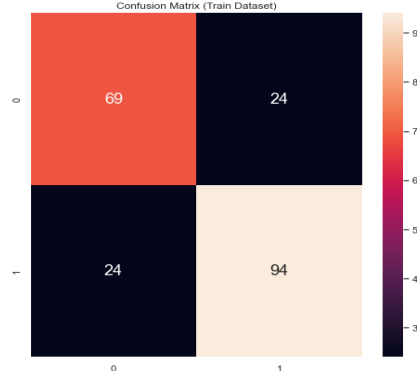
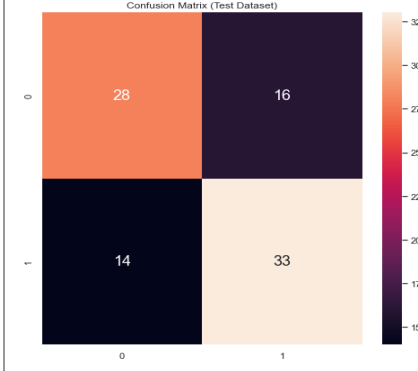
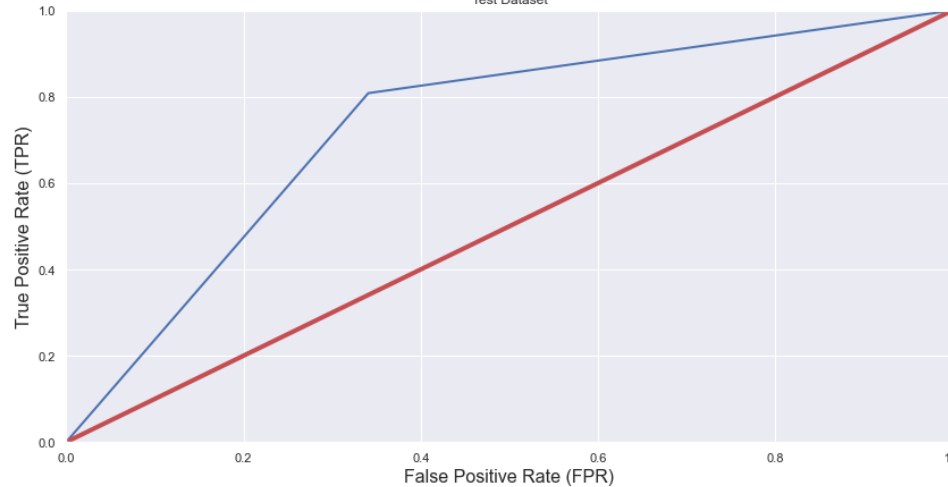
a) Disadvantages

- Large number of trees can make the algorithm too slow and ineffective for real-time predictions
- Algorithms are fast to train, but quite slow to create predictions once they are trained
- Trade-off between accuracy vs speed

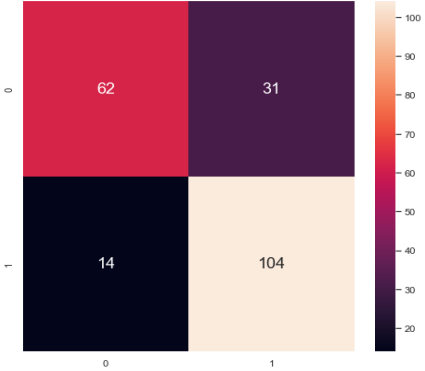
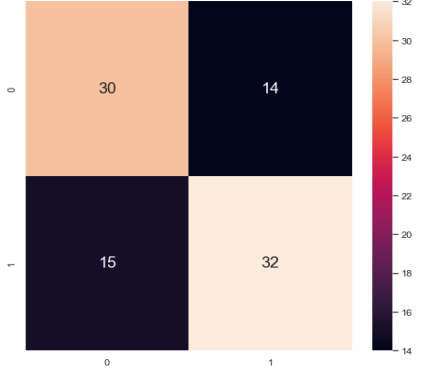
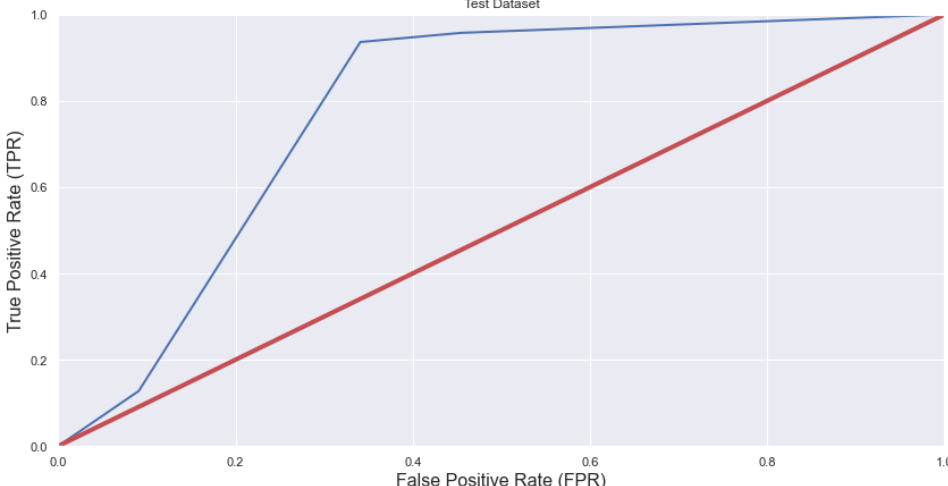
Appendix I: Stage 2 CART and Random Forests Full Performance Results

CART Model Performance Results

a) *dectree_m1* (pre-pruned tree)

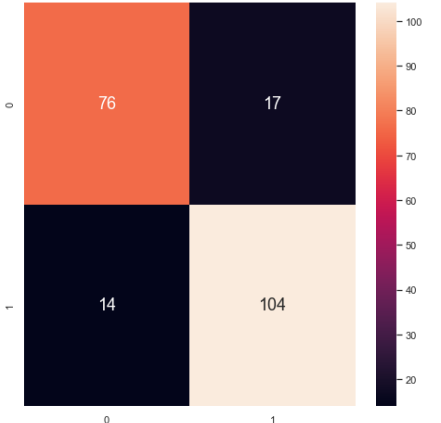
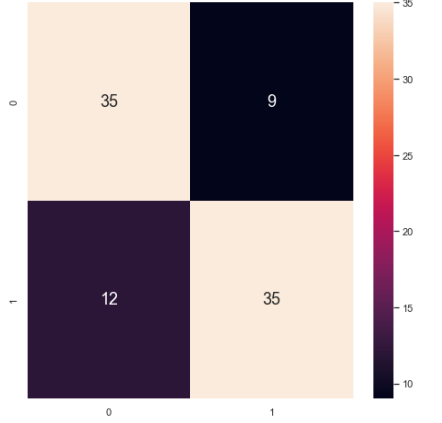
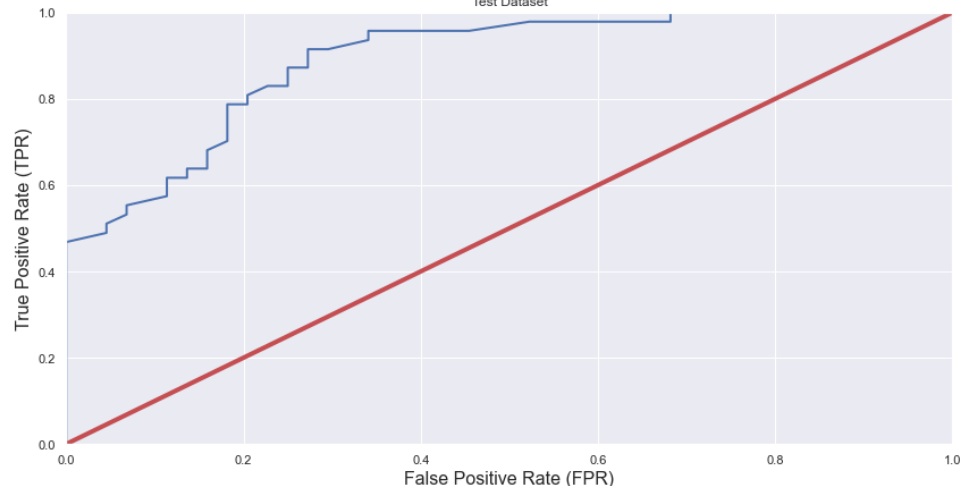
	Train dataset	Test dataset
Classification Accuracy	0.7689	0.6725
True Positive Rate	0.7966	0.6735
True Negative Rate	0.7419	0.6667
False Positive Rate	0.2581	0.3333
False Negative Rate	0.2034	0.3265
Precision	0.7966	0.6735
Recall	0.7966	0.7021
F1 Score	0.7966	0.6875
Confusion Matrix	<p>Confusion Matrix (Train Dataset)</p> 	<p>Confusion Matrix (Test Dataset)</p> 
ROC & ROC AUC Score	<p>Test Dataset</p>  <p>ROC AUC Score on test dataset: 0.7338.</p>	

b) *dectree_m2* (pruned tree – CP value = 0.03)

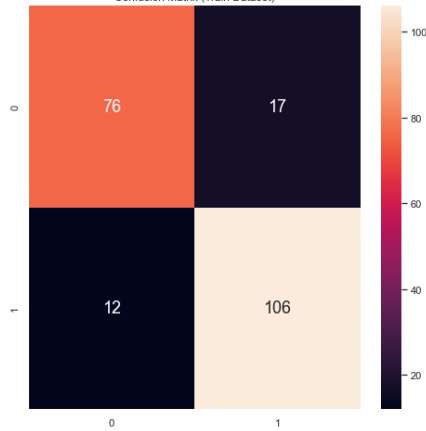
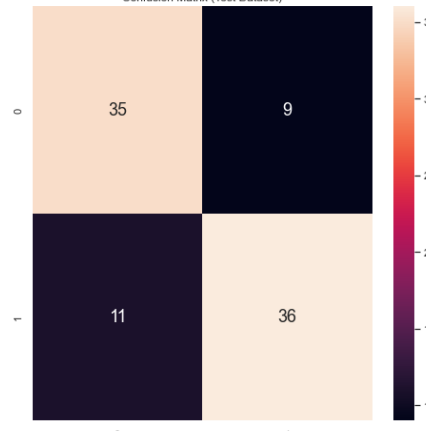
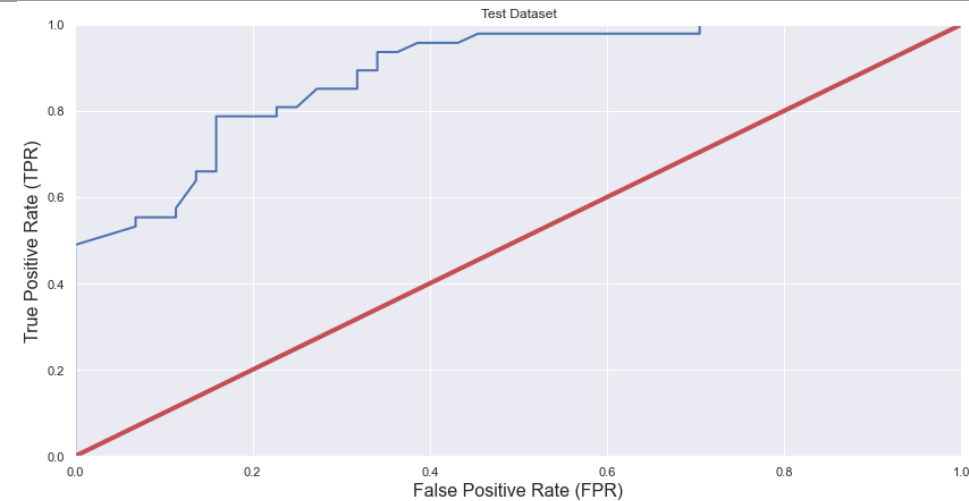
	Train dataset	Test dataset
Classification Accuracy	0.7969	0.7441
True Positive Rate	0.7704	0.6957
True Negative Rate	0.8158	0.6667
False Positive Rate	0.1842	0.3333
False Negative Rate	0.2296	0.3043
Precision	0.7704	0.6957
Recall	0.8814	0.6809
F1 Score	0.8221	0.6882
Confusion Matrix	<p>Confusion Matrix (Train Dataset)</p> 	<p>Confusion Matrix (Test Dataset)</p> 
ROC & ROC AUC Score	<p>Test Dataset</p>  <p>ROC AUC Score on test dataset: 0.7802.</p>	

Random Forest Model Performance Results

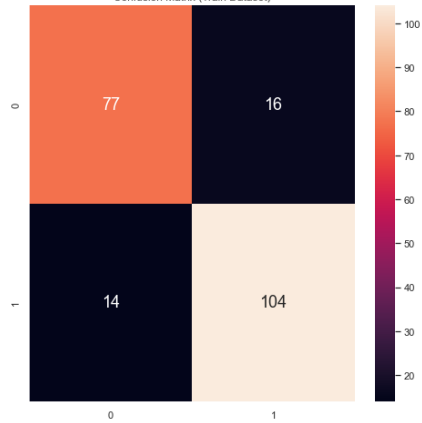
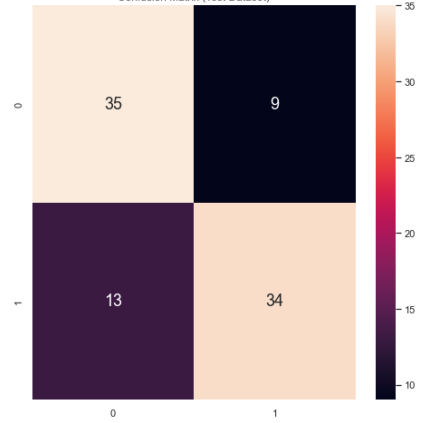
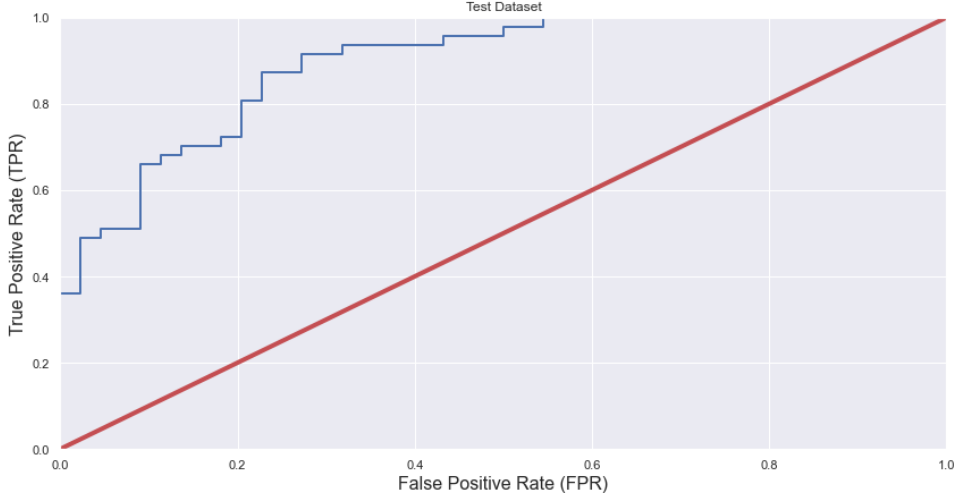
a) *random_forest_m1* (default hyperparameters and all features)

	Train dataset	Test dataset
Classification Accuracy	0.8895	0.8413
True Positive Rate	0.8595	0.7954
True Negative Rate	0.8444	0.7447
False Positive Rate	0.1555	0.2553
False Negative Rate	0.1405	0.2045
Precision	0.8595	0.7954
Recall	0.8814	0.7447
F1 Score	0.8703	0.7692
Confusion Matrix	<p>Confusion Matrix (Train Dataset)</p> 	<p>Confusion Matrix (Test Dataset)</p> 
ROC & ROC AUC Score	<p>Test Dataset</p>  <p>ROC AUC Score on test dataset: 0.8902.</p>	
Out-of-bag (oob) score	83.89%	

b) random_forest_m2 (feature importance – dropped 'fast_blood_sugar' feature)

	Train dataset	Test dataset																		
Classification Accuracy	0.8955	0.8171																		
True Positive Rate	0.8618	0.8																		
True Negative Rate	0.8636	0.7609																		
False Positive Rate	0.1364	0.2391																		
False Negative Rate	0.1382	0.2																		
Precision	0.8618	0.8																		
Recall	0.8983	0.7660																		
F1 Score	0.8797	0.7826																		
Confusion Matrix	<p>Confusion Matrix (Train Dataset)</p>  <table border="1"> <thead> <tr> <th></th> <th>Actual 0</th> <th>Actual 1</th> </tr> </thead> <tbody> <tr> <th>Predicted 0</th> <td>76</td> <td>12</td> </tr> <tr> <th>Predicted 1</th> <td>17</td> <td>106</td> </tr> </tbody> </table>		Actual 0	Actual 1	Predicted 0	76	12	Predicted 1	17	106	<p>Confusion Matrix (Test Dataset)</p>  <table border="1"> <thead> <tr> <th></th> <th>Actual 0</th> <th>Actual 1</th> </tr> </thead> <tbody> <tr> <th>Predicted 0</th> <td>35</td> <td>11</td> </tr> <tr> <th>Predicted 1</th> <td>9</td> <td>36</td> </tr> </tbody> </table>		Actual 0	Actual 1	Predicted 0	35	11	Predicted 1	9	36
	Actual 0	Actual 1																		
Predicted 0	76	12																		
Predicted 1	17	106																		
	Actual 0	Actual 1																		
Predicted 0	35	11																		
Predicted 1	9	36																		
ROC & ROC AUC Score	<p>Test Dataset</p>  <p>ROC AUC Score on test dataset: 0.8871.</p>																			
Out-of-bag (oob) score	84.83%																			

c) random_forest_m3 (Hyperparameter Tuning)

	Train dataset	Test dataset
Classification Accuracy	0.8945	0.8506
True Positive Rate	0.8667	0.7907
True Negative Rate	0.8462	0.7292
False Positive Rate	0.1538	0.2708
False Negative Rate	0.1333	0.2093
Precision	0.8667	0.7907
Recall	0.8814	0.7234
F1 Score	0.8739	0.7556
Confusion Matrix	<p>Confusion Matrix (Train Dataset)</p> 	<p>Confusion Matrix (Test Dataset)</p> 
ROC & ROC AUC Score	<p>ROC AUC Score on test dataset: 0.8917.</p> 	
Out-of-bag (oob) score	84.36%	