Name: Brandon Chen Yun Xin
Email: bran0025@e.ntu.edu.sg

Part A: Data Exploration and Preparation (30%)

Q1: Interesting findings and Data Quality Issues.

Data quality issues:

- NA Values
  There are a total of 6 NA values from 2 columns (4 from Ca and 2 from Thal). Ca refers to the number of major vessels coloured by Fluoroscopy while Thal refers to whether the patient has an inherited blood disorder that causes the body to have less haemoglobin than normal.

  Both NA values for Thal and Ca are missing at random. The proportion of these values also constitute less than 1% of the data. Therefore, the mode of these variables can be used to replace the NA values.

- Categorical Variables
  The original dataset contains categorical variables which have numerical data type. These must be converted to categorical data type using factor() so as to allow for a logical data exploration as well as model predictions.

- Outliers
  There are numerous outliers that will be highlighted in the Data Exploration below.

Note: Patients with Angiographic Heart Disease will be referred to as AHD Patients.

Data Exploration Insights

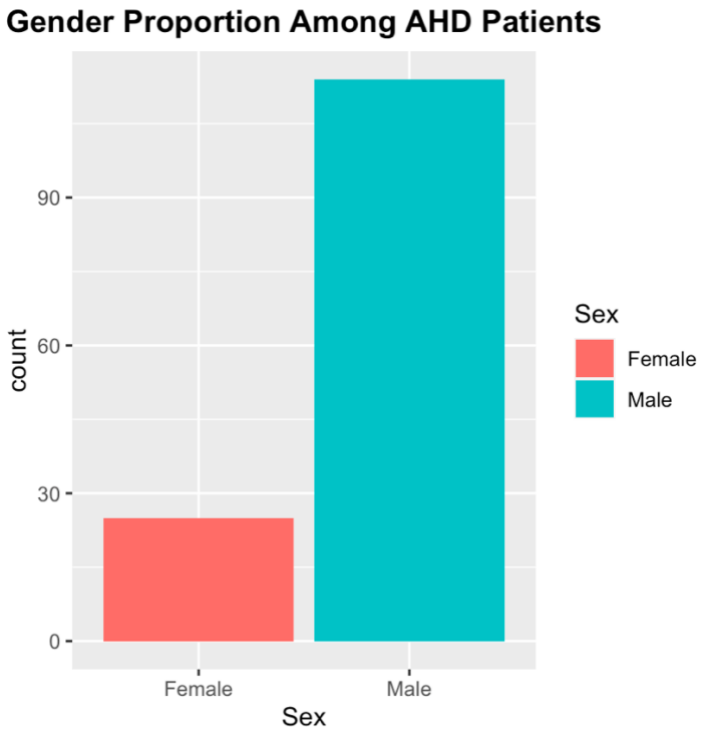

Figure 1

In general, there is a higher proportion of Males than Females among AHD patients. This will be analysed further in the summary table below.

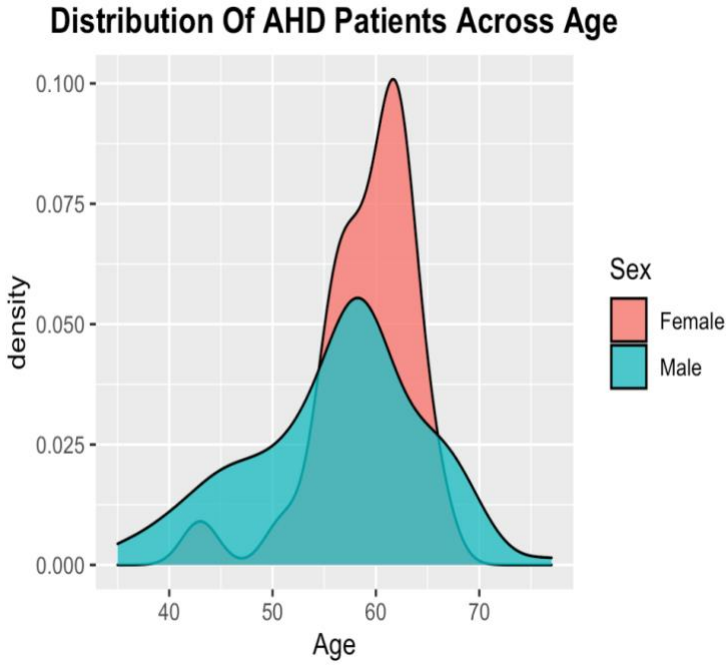

Figure 2

Males tend to have AHD at an earlier age compared to Females.

As seen from the density plot, below the age of 54, the number of Males are higher compared to Females.

Interestingly, between the ages 54 and 65, there are a greater proportion of Females than Males. This will be analysed further in the summary table below.

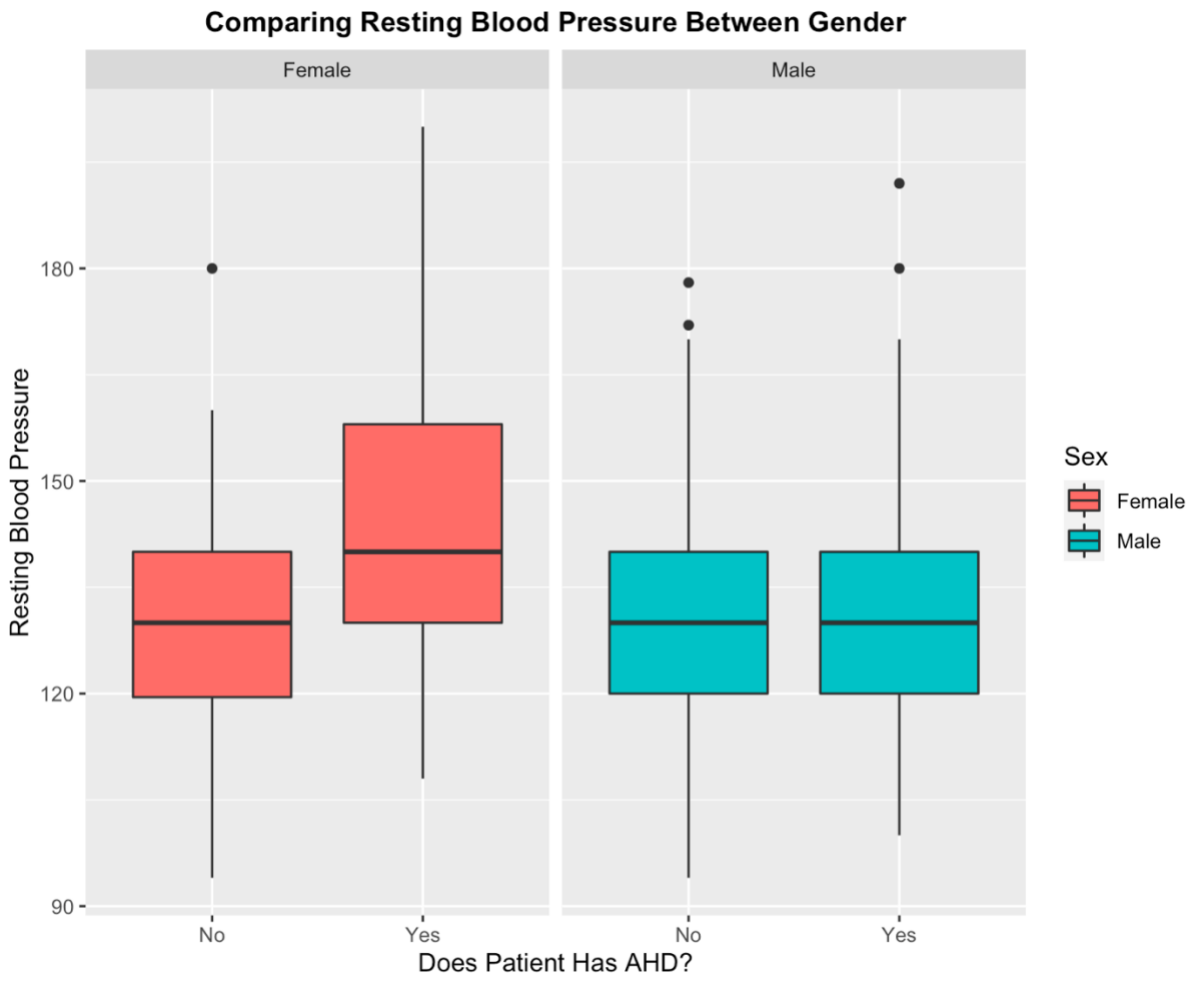

Figure 3

Resting Blood Pressure (RBP) is generally similar amongst Males regardless of whether they have AHD or not.

However, for Females, the RBP of an AHD patient tends to be **higher** than a non-AHD patient. As seen from *Figure 3*, the median RBP of Female AHD patients is higher than Female non-AHD patients as well as Males. The fact that Female non-AHD patients have similar RBP to Males also further suggests that there might be some association between high RBP of Females and AHD. This will be further analysed in the summary table below.

Data Quality: It is also noted that there are some patients who have much higher RBP than normal. This can be observed from the outliers seen in males as well as non-AHD Females and must be taken into consideration when performing regression analysis.

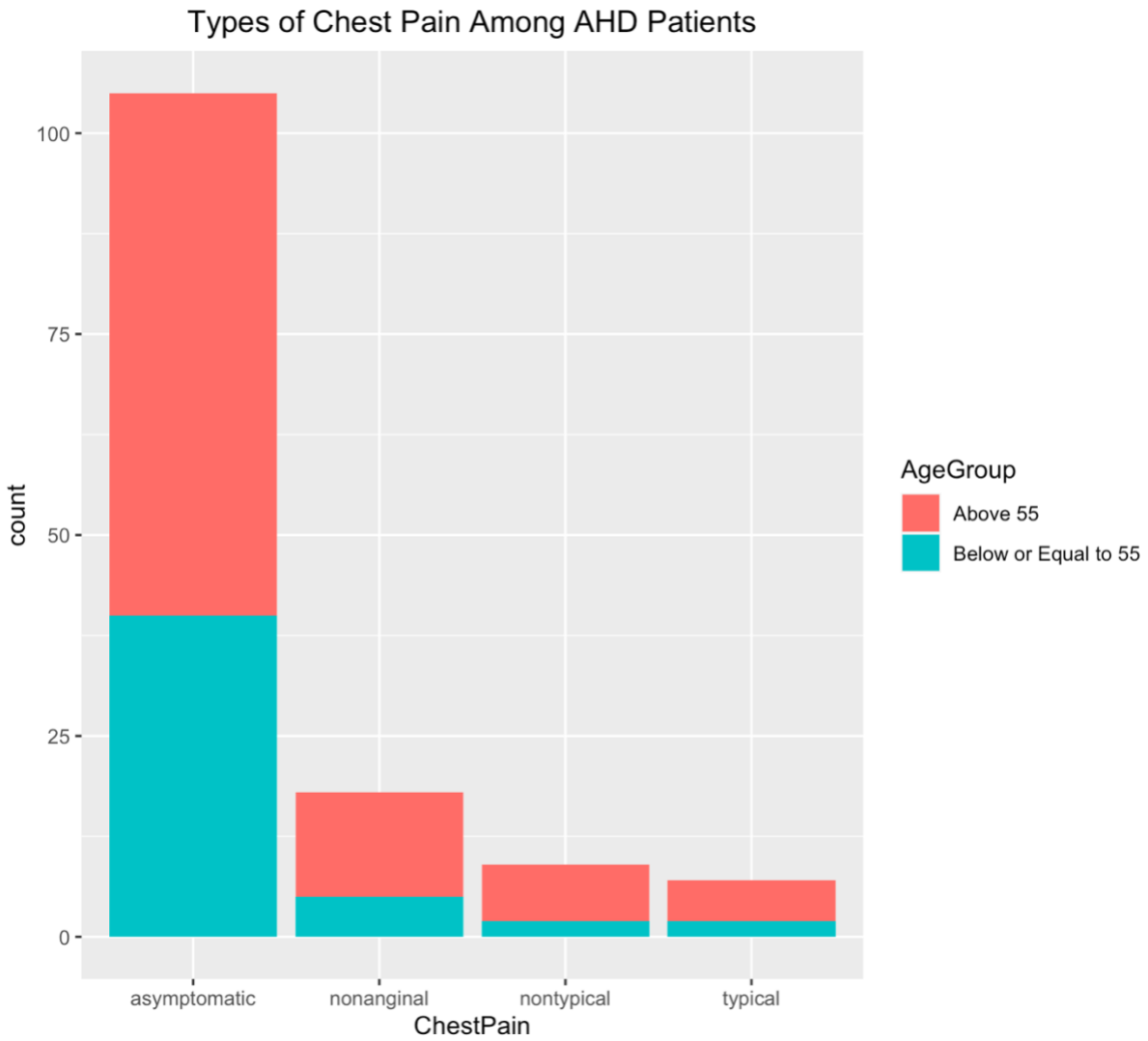

Figure 4

Most of the AHD patients show no sign of Chest Pain as seen from the vast difference in numbers between Asymptomatic patients and those with various types of Chest Pain.

While it is clear from *Figure* 4 that for every type of Chest Pain, there is a greater proportion of patients above the age of 55, it can be noted that for asymptomatic patients, there is a relatively large proportion of patients below or equal to the age of 55. This might suggest an increase in AHD patients among the younger age group who show no clear indication of the disease itself. This will be further analysed in the summary table below.

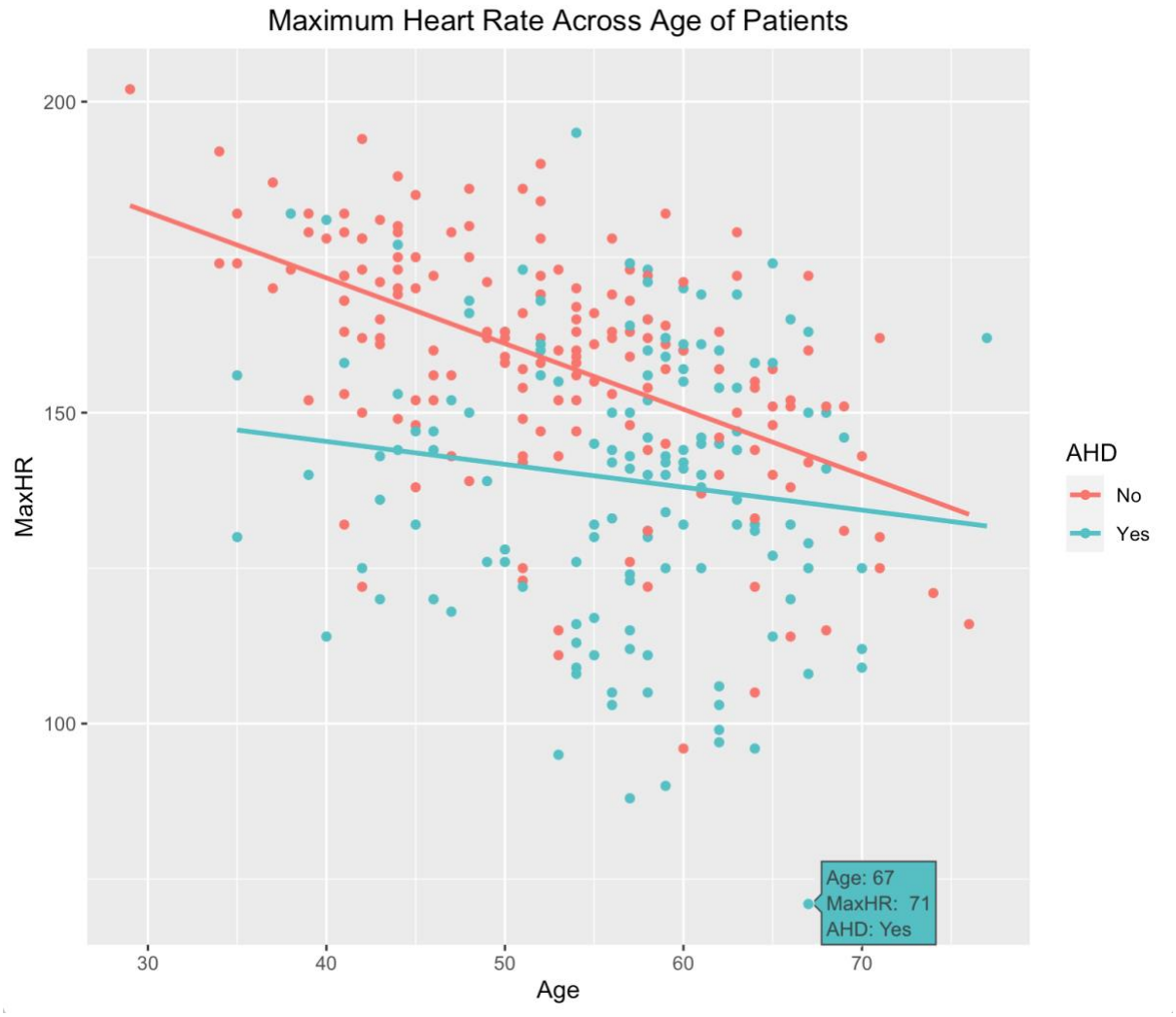

Figure 5

As expected, **patients' Maximum Heart Rate decreases as they become older**. Based on observations, **AHD patients also tend to have a lower Maximum Heart Rate compared to non-AHD patients**. This might suggest some association between having a low Maximum Heart Rate and having AHD. It is also noted that there is a larger variance in Maximum Heart Rate amongst AHD patients. These will be further analysed in the summary table below.

Data quality: As seen from the bottom right in *Figure 5*, there is an AHD patient with a significantly lower Maximum Heart Rate of 71. This should be taken into account when performing regression analysis since outliers might have a significant influence on the model.

Q2: Summary of Key Findings

Variables	Key Findings	Explanation
Sex, Age, AHD	In *Figure 1 & 2,* there are more Males who have AHD compared to Females. In addition, these Males tend to contract AHD at an earlier age compared to Females. Number of AHD Patients Males: 114 Females: 25	This might be due to hormonal differences between Males and Females. A research article shown that women tend to have lower likelihood of contracting heart disease compared to men. The article further mentioned that this remains true until after menopause as their oestrogen levels decrease, explaining the sudden increase in Female patients between the ages 54 and 65 as seen in *Figure 2*.
Sex, RestBP, AHD	As seen in *Figure 3*, female AHD patients tend to have higher RestBP compared to non-AHD patients as well as Males. Median RBP Female AHD patients: 140 All other categories: 130 Max RBP Female AHD patients: 200 All other categories: 178 - 192	In addition to the above finding, hormonal differences between Males and Females might also account for the higher RestBP in Female AHD patients. In the same research article, it is also suggested that Resting Blood Pressure might increase in female patients with cardiovascular disease due to hormonal differences. Thus, a Female might have a higher RestBP if they have AHD. In comparison, Males tend to have relatively unchanged RestBP regardless of whether they have AHD.

AgeGroup, ChestPain, Age	In *Figure* 4, most of AHD patients are above the age 55 and asymptomatic. Also, there is a significant proportion of asymptomatic AHD patients below the age of 55. Number of AHD Patients Asymptomatic: 105 Non-Asymptomatic: 34 Proportion of Asymptomatic AHD Above Age 55: 65 Age 55 and Below: 40	Asymptomatic cardiovascular disease is not an uncommon observation. As noted in a research journal from British Medical Bulletin, these might be due to many lifestyle factors. There is a rising trend of asymptomatic AHD patients being of a younger age group. The American College of Cardiology reported that 1 in 5 people who suffered from heart attack are below 40 years old. This is also associated with the rising rates of obesity which might be caused due to lack of regular exercise amongst the younger generation.
MaxHR, Age, AHD	All patients' Maximum Heart Rate decreases as they become older. AHD patients also tend to have a lower Maximum Heart Rate compared to non-AHD patients as shown in *Figure 5.*	Based on domain knowledge, the heart muscles tend to become weaker with age. Therefore, this might explain why the general Maximum Heart Rate decreases as age increases. Patients with AHD also tend to have lower Maximum Heart Rate. However, based on a Healthline article, different heart diseases have different impacts on a patient's Maximum Heart Rate. Therefore, the association between AHD and Maximum Heart Rate depends on the type of AHD. **This might explain why there is a large variance in Maximum Heart Rate amongst AHD patients**.

Part B: Analytics and Model Based Insight (40%)

Q3. Executing CART and logistic regression model. Comparison of Performance.

CART performs slightly better in terms of overall accuracy, sensitivity and specificity. It has an overall accuracy of 82% while Logistic Regression model has an overall accuracy of 77%.

More importantly, it has a higher sensitivity at 81% compared to Logistic Regression's 73% sensitivity, **a difference of 7%.** Sensitivity is the ability of the model to correctly identify people who have AHD. This means that the CART model was better at correctly predicting patients with AHD.

In terms of specificity, CART also performed better at 84% while Logistic Regression obtained 80% specificity. Specificity is the ability of the model to correctly exclude individuals who do not have AHD. In other words, Logistic Regression predicted more number of people having AHD when they do not actually have them.

While these two models are done under the exact same conditions, both having NA values replaced by the variable's mode, it can be seen in an additional CART model (With Surrogates) that the performance of the CART model is greater when it handles the NA values through surrogates. This additional feature of CART model handles the NA values automatically and is able to yield greater accuracy, sensitivity and specificity in comparison to Logistic Regression model.

	Logistic Regression	**CART**	**CART (With Surrogates)**
Overall Accuracy	77%	82%	86%
Sensitivity	73%	81%	87%
Specificity	80%	84%	85%

Performance Metrics

Q4. Recommend which model to the hospital.

The CART model is recommended for the hospital's use. It has **higher accuracy, specificity and most importantly, sensitivity**. In my opinion, the ability to correctly predict AHD patients is most important to the hospital. It would mean more patients being correctly identified to have AHD when they might not show clear symptoms of having the disease.

Along with higher predictive accuracy, the CART model is also **able to handle NA values via surrogates**. This will reduce the need to manually replace or remove NA values as seen when carrying out Logistic Regression.

Also, there is **no need to split CART model into train and test set** since CART model automatically utilises 10-fold cross validation to do that. Hence, this minimises the loss of data used to train the model.

Additionally, CART model is able to **automatically identify features** that are important in making predictions. This will reduce the need to perform feature elimination to identify the relevant attributes to fit into the model as seen in Logistic Regression. Thus, it is more convenient and user-friendly for the hospital to use especially since the hospital's staff might not be analytically trained. A CART model will also be **easier for the hospital's staff to interpret**.

In general, since there are multiple attributes to be used in the model, CART is also recommended since CART does not require data to be linearly separable and automatically handles decision making. Meanwhile, Logistic Regression assumes the data to be linearly separable in space and a decision threshold has to be set.

In conclusion, a CART model is recommended for the hospital.

Q5. Explain key findings to Hospital.

Among the 13 attributes fit into the model, 11 of the attributes have been determined to be important in the predictions. They are summarised in the table shown below, along with the proportion of their importance. **ChestPain** has been identified to be the most important variable in predicting whether a patient has AHD.

	ChestPain	Thal2	Oldpeak	ExAng	MaxHR
Variable Importance	25.973	17.729	13.600	10.071	9.824
Proportion	26%	18%	15%	10%	10%

	numVessels	Slope	Sex	RestBP	Age	Fbs
Variable Importance	7.428	5.795	4.075	1.933	1.678	0.495
Proportion	7%	6%	4%	2%	2%	0%

Variable Importance

The derivation of the prediction can be summarised into the decision tree as shown below in *Figure 6*. From the root node, if the patient has non-anginal, non-typical or typical ChestPain, the probability of having AHD will be 0.23. For these patients, if Oldpeak is smaller than 2, then there is a 0.78 probability that they have AHD. If not, then they will have 0.18 probability of having AHD.

On the other hand, for patients with asymptomatic ChestPain, the probability of having AHD will be higher at 0.72. If these patients have Fixed or Normal Thalassemia, then there is a 0.93 probability of having AHD. If not, then they would have 0.47 probability of having AHD. For these patients with reversible Thalassemia, if they have 0 Vessels coloured during Fluoroscopy, then they would have 0.87 probability of having AHD. If not, then they would have 0.25 probability of having AHD.

In summary, **patients who have Asymptomatic ChestPain and Fixed or Normal Thalassemia have the highest probability of having AHD**. Whereas patients who have Asymptomatic ChestPain, Reversible Thalassemia and Number of Vessels Coloured by Fluoroscopy greater than 0 have the lowest probability of having AHD.

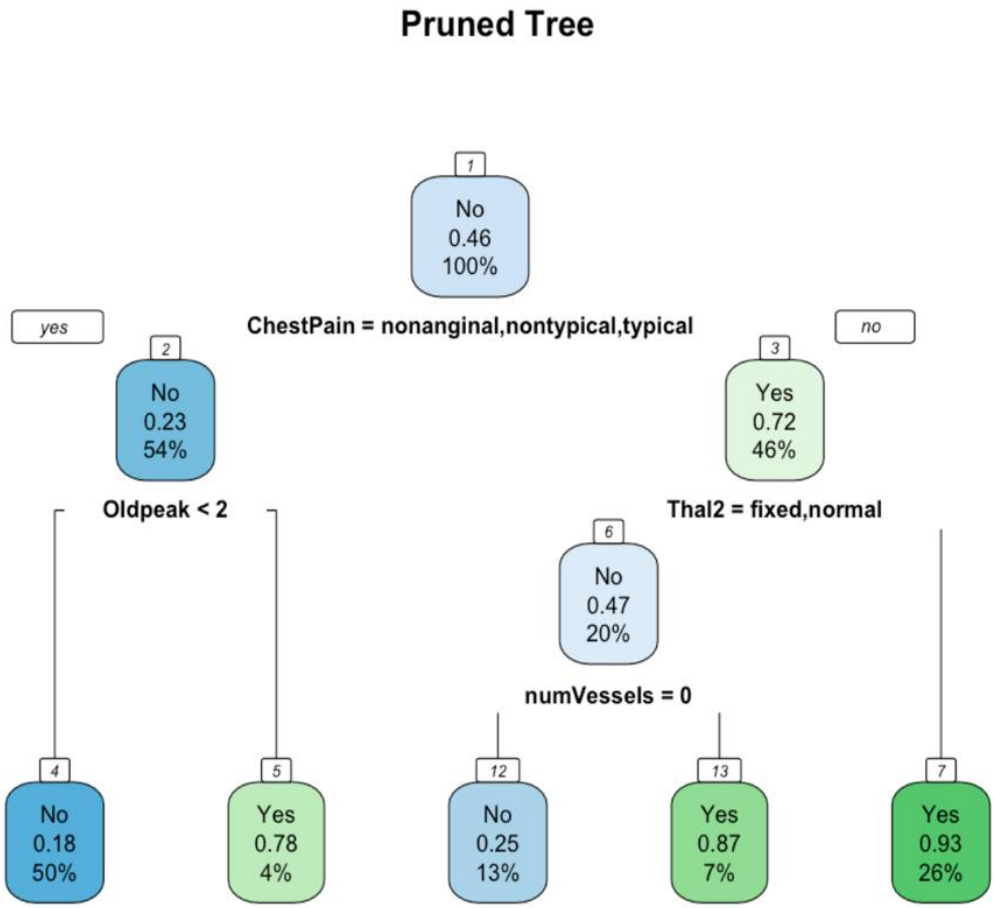

Figure 6: Pruned Decision Tree

Note:

- numVessels refer to the number of Vessels Coloured by Fluoroscopy. In the original dataset, it is the attribute Ca.
- Thal2 refer to Thalassemia. In the original dataset, it is Thal.
- These 2 variables have been added due to NA values in the original dataset.

Part C: Advanced Concepts (30%)

Q6. Cross Validation Error for Categorical Variable

If the outcome variable is continuous, the cross validation error in CART would refer to the Root Mean Squared Error (RMSE) at each node. This is in line with linear regression where RMSE is used as a metric to account for the difference in values predicted by the model and the observed values. The higher the RMSE, the greater the deviance of the predicted values from the observed values.

On the other hand, for a categorical outcome variable, the cross validation error for the CART model refers to the misclassification error. Misclassification error is calculated by subtracting the probability of the predicted category from 1. For a binary outcome variable, it is a reasonable method of measuring the accuracy of the model. However, for an outcome with multiple categories, the **cross validation error does not provide any information of the probabilities of the other individual categories** which might have a high probability but not the highest probability. Therefore, there is a need to account for these information.

This can be done by obtaining the sensitivity as well as the specificity of the model. To obtain the sensitivity, the true positives and false negatives can be acquired from the model or cp table and calculated by TP/(TP+FN). Whereas for specificity, it can be calculated by using true negatives and false positives within the formula TN/(TN+FP).

It is also important to account for sensitivity or specificity of the model because they might be the main performance metric of the model depending on the situation and objective.

For example, one might be willing to accept false-negatives in spam detection but it would not be the same acceptable set-up in detecting HIV. Thus, the objective of the model determines which metric should be prioritised. However, for the CART model, only the accuracy of the model is prioritised as the model seeks to find the optimal point to prune the tree based on the cross validation error as well as the complexity parameter. This might not necessarily translate into higher sensitivity or specificity.

Q7. Comment on Approach

Approach:
Step 1: Rank attributes according to correlation measure.
Step 2: Perform classification on known data and compare accuracy.
Step 3: Recommend relevant attributes based on accuracy of classifier model.

By performing forward selection to select relevant attributes, this approach improves the accuracy of the normal regression model. It starts with the attribute with the highest correlation measure and the other attributes are gradually added.

When a new attribute is added to the model, a decrease in accuracy suggests that the particular attribute might be a poor predictor whereas an increase in accuracy suggests that the particular attribute is a good predictor, with the difference in accuracy also assessed. This identifies how a specific attribute contributes to the model accuracy and the extent it contributes which is helpful in identifying strong attributes.

However, this method **does not evaluate multicollinearity**. The attributes are added in based on model accuracy only. If there are attributes that are collinear, then the accuracy of the model would be inflated. The newly-added attribute might then be incorrectly classified as a good predictor.

The model also uses correlation measure to classify the relevant attributes, **without considering their statistical significance**. Statistical significance can be measured in terms of *p-values*, which refer to the risk of concluding association when there are no association at all. Attributes with high correlation might have high *p-values* which would mean that these attributes have a high risk of having no association. Therefore, adding attributes starting from the highest correlation measure might not be the most ideal method.

This model is also **computationally intensive**. Three different classifiers are trained with 13 different attributes gradually added. This would mean that there would be 39 different set of predictions performed before arriving at the end-result.

With the above downsides, some improvements can be made to this approach. The attributes can be fit in based on their statistical significance. If the attributes have a low statistical significance, they should be included towards the end. At each step, the model should also be evaluated to check if certain attributes should be removed based on their statistical significance or accuracy of the classifier. The attributes currently in the model should also be measured for multicollinearity to correctly identify the relationship between attributes and the model's accuracy.

References:

1. Data Cleaning Information
   https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

2. RPart Comparison between Classification and Regression Tree
   https://www.datacamp.com/community/tutorials/decision-trees-R

3. Sensitivity Measure Implementation in R
   https://rstudio-pubs-static.s3.amazonaws.com/370944_96c386c03ac54ef3bec4535d49e92890.html

4. Model Performance Metrics
   https://labtestsonline.org.uk/articles/accuracy-precision-specificity-sensitivity

5. Research Article on Heart Disease and Blood Pressure between Genders
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283814/

6. British Medical Bulletin
   https://academic.oup.com/bmb/article/59/1/3/282080

7. American College of Cardiology Research Journal on Rise in Heart Attacks in Youths
   https://www.cardiosmart.org/news/2019/3/heart-attack-rates-on-the-rise-in-young-adults

8. Healthline Article on Maximum Heart Rate an Heart Attacks
   https://www.healthline.com/health/heart-rate-during-heart-attack#types-of-heart-attacks

9.