Name: Ernest Ang Cheng Han

Email: [ERNE0009@e.ntu.edu.sg](mailto:ERNE0009@e.ntu.edu.sg)

Matriculation Number: U1921310H

Module: BC2406

Assessment: Computer Based Assessment

1. Conduct data exploration. Show and explain interesting findings. What are the data quality issues (if any)?

Before conducting data exploration, I think it is important to first identify the data quality issues present in the data set. We will then perform data cleaning before diving into exploration, visualisation and subsequently prediction. By doing so, it ensures that subsequent analysis made are based on the most updated and correct information, improving data quality and overall productivity (Gimenez, 2018).

A picture containing text

Description automatically generated

Upon importing the data set via the fread() function from the data.table library, we observe that, “Sex”, “Fbs”, “RestECG”, “ExAng”, “Slope”, “Ca”, “Thal” are categorical variables but are read in as continuous variables. Hence we first convert these continuous variables to categorical ones via the factor() function.

Scatter chart

Description automatically generated

For the categorical variables, we observe that NA values are present in variables “Ca” and “Thal”. For the continuous variables, visualising them sequentially via a boxplot, though outlier values are present, however, they do not seem to need cleaning as these values are possible especially amongst people with severe health problems (e.g. Cholesterol levels of > 500 units are observed with people suffering from severe obesity/high blood pressure (Cleveland Clinic, 2020))

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated*Boxplot for “RestBP” Boxplot for “Age” Boxplot for “Chol”*

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated  
*Boxplot for “MaxHR” Boxplot for “Oldpeak”*

For cleaning the categorical variables “Thal” & “Ca”, there are 5 ways of cleaning it: (1) removing the NA values, (2) cleaning by mode, (3) cleaning by relevant subgroup, (4) cleaning by CART, (5) cleaning by logistic regression (Rajwanshi, 2018). Simply removing the values are possible and easy but due to the fact that this method of data cleaning should be used as a last resort and also because of the fact that our data set is already so small with just 300 observations, we will not be able to simply remove the observations with NA values (AI Multiple, 2020). Cleaning the categorical variables by CART and Logistic regression is quite expensive and time consuming while producing barely accurate results as shown:

Graphical user interface, text

Description automatically generated Graphical user interface, application

Description automatically generated  
*Results for CART Results for Logistic Regression*

Cleaning the variables by column mode and relevant subgroup of observed NA rows produced the same results. Hence, we will clean the data set via replacing all NA values with the respective column mode:

A screenshot of a cell phone

Description automatically generated  
*Cleaning by column mode*

A screenshot of a cell phone

Description automatically generated  
*Cleaning by relevant subgroup mode*

After cleaning the data set, we can move on the exploration analysis and subsequently predictive analysis in the subsequent sections.

**Categorical Variables vs AHD status:**

|  |  |
| --- | --- |
| Figure 1Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = Sex\_New, y = frequency(AHD), fill = factor(AHD)))+geom\_bar(stat = “identity”, position = “fill”)+ggtitle(“Stacked bar chart: AHD Status against Gender”)+labs(x = “Gender”, y = “Frequency” , fill = “AHD”)+scale\_y\_continuous(labels = function(y)paste0(y\*100, “%”))+scale\_fill\_manual(values = c(“light blue”, “dark blue”)) | Figure 2Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = Fbs\_New, y = frequency(AHD), fill = factor(AHD)))+geom\_bar(stat = “identity”, position = “fill”)+ggtitle(“Stacked bar chart: AHD Status against Fbs”)+labs(x = “Fbs”, y = “Frequency” , fill = “AHD”)+scale\_y\_continuous(labels = function(y)paste0(y\*100, “%”))+scale\_fill\_manual(values = c(“light blue”, “dark blue”)) |
| Figure 3 Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = ExAng\_New,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against Angina")+labs(x = "Angina", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) | Figure 4 Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = Ca,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against Ca")+labs(x = "Ca", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) |

|  |  |
| --- | --- |
| Figure 5 Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = Slope\_New,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against Slope")+labs(x = "Slope", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) | Figure 6 Chart, bar chart  Description automatically generatedggplot(data = clean, aes(x = Thal,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against Thal")+labs(x = "Thal", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) |
| Figure 7 Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = RestECG\_New,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against RestECG")+labs(x = "RestECG", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) | Figure 8 Chart, bar chart  Description automatically generated  ggplot(data = clean, aes(x = ChestPain,y = frequency(AHD),fill = factor(AHD)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: AHD Status against ChestPain")+labs(x = "ChestPain", y = "Frequency" , fill = "AHD")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "dark blue")) |

**Continuous Variables vs AHD status:**

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated  
Figure 9 Figure 10 Figure 11  
Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated Graphical user interface, text, application, email

Description automatically generated  
 Figure 12 Figure 13 sfsdfdfsdfdsfsdfsdfsdfsdfsdds

From our visualisations and explorations, we can hypothesize the following:

From Figure 1:   
The likelihood of males developing AHD is much higher compared to females. From figure 9, AHD seems to occur in people who are older. Hence, this observation seen in figure 1 may be a result of there being a greater proportion of older people amongst the males compared to females. However, on further exploration, it seems that even after taking into account both the variables of age and sex, more males at a younger age since to be developing AHD compared to females whose AHD statuses are observed mostly in older ages. This could illustrate that perhaps within the male demographic, there are certain habits (e.g. Dietary) that are contributing to a higher risk of contracting AHD (Harvard Medical School, 2016).

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated (Shows that spread of age is fairly similar in both male and females)  ggplot(data = clean) +geom\_boxplot(aes(x=Sex\_New, y=Age, fill=Sex\_New)) | Chart, box and whisker chart  Description automatically generated (Shows that males tend to develop AHD at a younger age compared to females)  ggplot(data = clean) +geom\_boxplot(aes(x=Sex\_New, y=Age, fill=AHD)) |
| Chart, histogram  Description automatically generated  ggplot(clean, aes(x=Age,y=frequency(AHD), fill=AHD)) +geom\_col() + facet\_grid(. ~clean$Sex\_New)+labs(x = "Gender", y = "Frequency") | Chart, bar chart, histogram  Description automatically generated  ggplot(clean, aes(x=Age,y=frequency(Age), fill=Age)) +geom\_col() + facet\_grid(. ~ clean$Sex\_New)+labs(x = "Gender", y = "Frequency") |

Figure 2:   
Fbs has almost no impact on AHD as percentage of people having AHD and Fbs compared to people having AHD and not having Fbs is both at around 50%. This is a contradiction to expert literature which states otherwise – people with fasting blood sugar above 90 puts them at a higher risk of experiencing heart diseases (Diabetes in Control, 2002)

Figure 3:   
People with exercised induced angina tend to also suffer from AHD compared with people who are not suffering from this type of chest pains. Exercise induced angina could be a symptom of AHD. From figure 1, it has been observed that regardless of age, males tend to have a higher likelihood of developing AHD compared to females. Below, it is shown that there is a greater proportion of males with angina compared to females, which may support the correlation between AHD & angina

Figure 4:   
The more major vessels coloured by fluoroscopy, the higher chances of that particular person suffering from AHD. Fluoroscopy works by helping Doctors find blockages in clogged heart arteries as contrast dye moves through them (Medline, 2019). The more major arteries detected through this procedure, the higher indication that these major arteries are clogged, which would of course translate higher risks of developing heart diseases.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated (~10% more males have > 0 arteries detected by fluoroscopy compared to females. Males who have 3 coloured major arteries doubles females)  ggplot(data = clean, aes(x = Sex\_New,y = frequency(Ca), fill = factor(Ca)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: Ca Status against Gender")+labs(x = "Gender", y = "Frequency" , fill = "Ca")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values =c("light blue", "cadet blue", "deep sky blue", "dark blue")) | Chart, bar chart  Description automatically generated (Shows that ~20% more males experience exercise induced angina compared to females)  ggplot(data = clean, aes(x = Sex\_New,y = frequency(ExAng\_New), fill = factor(ExAng\_New)))+geom\_bar(stat = "identity", position = "fill")+ggtitle("Stacked bar chart: Angina Status against Gender")+labs(x = "Gender", y = "Frequency" , fill = "Angina")+scale\_y\_continuous(labels = function(y)paste0(y\*100, "%"))+scale\_fill\_manual(values = c("light blue", "deeppink", "purple")) |

Figure 5:   
People whose slope of ST segment’s peak when exercising is down sloping or flat tend to have AHD compared to those whose slope is up sloping. A treadmill ECG stress test detects anomalies when there is a flat or down sloping ST segment observed (Rawat, 2019).

Figure 6:   
People with Thalassemia tend to develop AHD compared with people who are not suffering from Thalassemia. Additionally, males tend to develop Thalassemia more than females. Thalassemia is a blood inherited disease (Mayoclinic, 2019), which is an external uncontrollable genetic based factor rather than internal controllable behavioural factors such as unhealthy diets. Thalassemia seems to exhibit similar symptoms to AHD, such as anomalies in ECG readings, measured by Oldpeak.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated (Illustrates that males have a likelihood of inheriting Thalassemia)    ggplot(clean, aes(x=Age, y=frequency(Thal), fill=Thal)) +geom\_col() + facet\_grid(. ~ clean$Sex\_New)+labs(x = "Gender", y = "Frequency") | Chart, box and whisker chart  Description automatically generated (Illustrates that people with Thalassemia tend to have more anomalies with their ECG readings, illustrated by greater Oldpeak value measuring ST Depression)  ggplot(data = clean, aes(x = clean$Thal, y = clean$Oldpeak)) + geom\_boxplot(aes(x=Thal, y=Oldpeak, fill=Sex\_New)) + labs(x = "Thal", y = "Oldpeak", color="Sex") |

Figure 7:   
People with ST-T wave anomalies as well as hypertrophy have a higher likelihood of having AHD. ST-T wave anomalies are symptoms of heart problems (Friedberg, 1961) and it is no wonder that people having such anomalies in their readings are more likely of also having AHD. On the other hand, although hypertrophy is one of the symptoms of AHD, it is expected that it would not be such a great predictor given that muscular hypertrophy detected in readings could be a result of exercise and intense workouts (Healthline, 2019)

Figure 8:   
People experiencing asymptomatic chest pains tend to have AHD compared to other types of chest pains experienced. This category for chest pains could be symptoms of AHD as it has been identified as a symptom for Heart Attacks (Wong, 2020)

Figure 9:   
AHD seems to occur amongst older people. This could be attributed to the fact that as people age, their immune systems weakens (Dorshkind, 2013) which makes them more susceptible to diseases such as heart conditions.

Figure 10:   
People with AHD tend to have a higher blood pressure. High blood pressure has been identified as one of the symptoms leading to heart diseases (CDC, 2020) which could be possible explanation to the correlation between AHD and blood pressure

Figure 11:   
People with AHD tend to have higher levels of cholesterol. Apart from its links to heart disease (Carson, 2019), high levels of cholesterol has many links to the other factors that have been identified to be symptoms or indicators of AHD and other heart diseases, such as high blood pressure, ST-T segment anomalies and old age.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated  (Shows positive relationship between cholesterol and blood pressure)  ggplot(data = clean, aes(x = clean$Chol, y = clean$RestBP, colour=factor(clean$Sex\_New))) + facet\_grid(. ~ clean$Sex\_New) + geom\_point() + geom\_smooth(method='lm') +labs(x = "Cholesterol", y = "Blood Pressure", color="Sex") | Chart, scatter chart  Description automatically generated  (Shows positive relationship between cholesterol and ST-Depression (Oldpeak))  ggplot(data = clean, aes(x = clean$Chol, y = clean$Oldpeak, colour=factor(clean$Sex\_New))) + facet\_grid(. ~ clean$Sex\_New) + geom\_point() + geom\_smooth(method='lm') +labs(x = "Cholesterol", y = "Oldpeak", color="Sex") |

Figure 12:   
People with AHD tend to have lower heart rates. Having averagely lower heart rates is a symptom of heart conditions and heart attack, hence providing possible reasons to the observed graphs and plots seen in figure 12.

Figure 13:   
People with AHD tend to be experiencing ST-Depression on their ECG. ST-Depression refers to findings on an electrocardiogram wherein the ST-segment is abnormally low below the baseline. Again, such anomalies could also indicate the possibility of AHD as seen from the trend observed in figure 13.

**Conclusion & insights:**

Overall, apart from the Fbs factor, the other 12 variables seems to be possible predictors for AHD. Further supported by literature written by experts and reports done up by professional and medical authorities in the relevant fields, it is necessary to perform further analysis using other estimators such as odds ratio in order to narrow down the variables which have the greatest and most reliable prediction accuracy in determining AHD.

One interesting finding would be the higher likelihood of males developing AHD, or developing the factors with a strong relationship with AHD compared to females. As illustrated above, these factors could be certain internal or external characteristic of males such as genetics, diet and lifestyle which can be further investigated to explain this likelihood.

Apart from AHD, more males seem to experience Angina as well as inherit Thalassemia compared to females. Apart from the fact that these factors could contribute to AHD and that these factors have strong relationships with symptoms leading to AHD, it further supports the notion that males could have certain characteristics or genetics which would make them more susceptible to diseases compared to females.

It is also important to note that since the data set is quite small, observations and inferences made from exploratory and predictive analysis made not be representative or comprehensive of the global or even local population. As a result of the limitations imposed by our own data set, accuracy of the visualisation charts and predictions made subsequently may decrease.

1. Create a summary table that shows important information. Explain the findings.

Graphical user interface, text

Description automatically generated

Table

Description automatically generated  
Table

Description automatically generated  
Table

Description automatically generated

From our summary table, we observe the following:

* 1. Average age of males and females within the data set is relatively similar at the 55 – 60 age range. This adds consistency and reduces bias from disproportionate representation of males and females in the data set (Panzeri, 2015).
  2. There are more males than females suffering from AHD based on proportion, with 25/97 (25.8%) of females experiencing AHD compared to 114/206 (55.3%) of males experiencing AHD. This could be due to certain demographics and gender biased characteristics which could contribute to males developing AHD more compared to females (Harvard Medical School, 2016).
  3. People suffering from AHD tend to suffer asymptomatic chest pains compared to the other 3 types of chest pains, with more than 70% of females and males with AHD experiencing asymptomatic chest pains. This is consistent with most studies which states that males tend to experience chest pains more frequently and severely compared to their female counterparts (Bosner, 2009)
  4. People suffering from AHD tend to experience higher blood pressures compared to people who are not. Males with AHD on average have a resting blood pressure of 132 mmHg while males without AHD on average have a resting blood pressure of 130 mmHg. Females with AHD on average have a resting blood pressure 147 mmHg while females without AHD on average have a resting blood pressure of 129 mmHg. This difference in blood pressure is greater and more stark in females compared to males, hence using RestBP as a predictor for AHD may apply more for females compared to males. Additionally, in general, males and females have similar blood pressures.
  5. People suffering from AHD tend to have higher levels of cholesterol. Males with AHD on average have a cholesterol level of 246 while males without AHD on average have a cholesterol level of 232. Females with AHD on average have a cholesterol level of 276 while females without AHD on average have a cholesterol level of 257. Additionally, in general, females have higher cholesterol levels than males. This is a contradiction to most studies which indicate the opposite (Hazzard, 1985). One possible reason for this is that perhaps the data was not comprehensive and representative enough given its small size of 300 observations
  6. ST-T wave anomalies are rare occurrences whether a person has AHD or not, with ST-T wave anomalies being observed more in individuals diagnosed with AHD. Majority of the people with AHD tend to experience hypertrophy based ECG readings based on Estes Criteria.
  7. People suffering from AHD tend to have slower heart rates. Males with AHD on average have a heart rate of 138 bpm while males without AHD on average have a heart rate of 162 bpm. Females with AHD on average have a heart rate of 143 bpm while females without AHD on average have a heart rate of 154 bpm. Additionally, in general, males tend to have faster heart rates than females.
  8. People suffering from AHD tend to also be suffering from Angina, with an approximate 10% increase in Angina occurrences seen in people suffering from AHD compared to people who are not suffering from AHD. Additionally, in general, males tend to experience Angina more compared to females, with an approximate 5% more of males experiencing Angina compared to females.
  9. People suffering from AHD tend to have higher Oldpeak values. Males with AHD on average have Oldpeak values of 1.53 while males without AHD on average have Oldpeak values of 0.652. Females with AHD on average have Oldpeak values of 1.77 while females without AHD on average have Oldpeak values of 0.554. Oldpeak values are an indication of ST-depression anomalies found in ECG readings, which implies that people suffering from AHD will most likely have such anomalies which could be indicators or symptoms of AHD
  10. People suffering from AHD tend to have flat slopes on their readings during treadmill stress test, with more than 60% of such behaviours observed in both females and males; while people not suffering from AHD tend to have upward slopes on their readings with also more than 60% of such behaviours observed in both females and males
  11. People suffering from AHD have a greater proportion of > 0 coloured major arteries from fluoroscopy. The majority of people who are not suffering from AHD have 0 coloured major arteries from fluoroscopy. In general, males have less clogged arteries than females which is supported by the fact the females have higher cholesterol levels than males on average which has high causation effect to the clogging of arteries
  12. People suffering from AHD tend to also suffer from Thalassemia, with more than 50% of people suffering from AHD and also from reversible Thalassemia observed in both males and females. In general, males tend to suffer from Thalassemia more than females, with 40% of males suffering from Thalassemia compared to 30% of females suffering from Thalassemia.

1. Execute CART and another model of your choice taught in this module. Which model perform better? Explain.

**Logistic Regression:**

First, we split the data set into training set and test set. Training set would be used to train our logistic regression model while test set would be used to test and validate the accuracy of our train logistic regression model

Text, application

Description automatically generated

Using the glm() function, we build the logistic regression model with all of the variables and observe the results via summary() function.

Table

Description automatically generated

We then identify the variables which are significant at 5% and construct our final logistic regression model

A picture containing table

Description automatically generated

We then find out the odds ratio of the respective variables used to build the final logistic regression model and check the confidence interval of each odds ratio to ensure that it excludes 1 and is hence statistically significant

A picture containing text

Description automatically generated

All of the confidence intervals of the odds ratios of the variables used in the final logistic regression model are statistically significant. Using this model, we can go ahead and validates its accuracy via predicting the train set and test set. We will take the test set’s accuracy as the basis of comparison to our next model CART for a more accurate and reliable comparison

Graphical user interface, text, application

Description automatically generated

Further, we will ensure that there is no multicollinearity present in any of the variables used in the final logistic regression model via the vif() function. As all of the adjust GVIF values are below 2, the benchmark for detecting multicollinearity, multicollinearity is not present amongst the variables used.

A picture containing table

Description automatically generated

**CART:**

We first construct our CART model with all of the variables. To allow it to grow to its maximum length, we set the CP parameter to 0 and minsplit parameter to 2.

A picture containing graphical user interface

Description automatically generated

Chart, timeline, box and whisker chart

Description automatically generated

Text

Description automatically generated  
Text

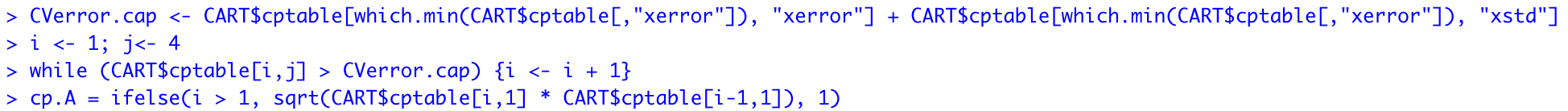
Description automatically generated

Graphical user interface, text, table

Description automatically generated  
Chart

Description automatically generated

After which, we will need to prune the maximum CART to its minimum at the tree’s weakest link. The weakest link is defined as the minimum value of complexity penalty that would trigger pruning. The derivation of this value is shown in the following:

  
Diagram

Description automatically generated  
Text

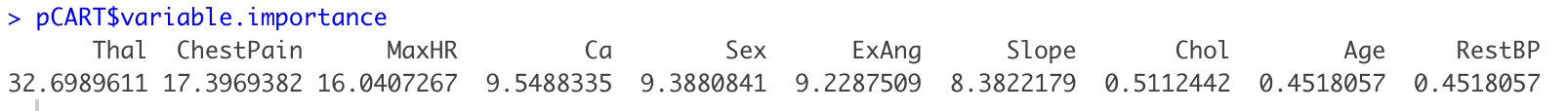
Description automatically generated

After pruning the tree, we use our pruned CART model to predict the AHD status on both the train set and the test set which was initially deduced when we created the logistic regression model. This is to ensure consistency when comparing logistic regression and CART in the subsequent analysis and evaluation.

Text

Description automatically generated

Using the variable.importance method, we can also view the variables deemed important by our pruned CART model which can be used additionally in evaluating factors which are important and accurate when predicting AHD status



Analysing the prediction tables on the test set by both CART and logistic regression, we get the following results:

A picture containing text

Description automatically generated A picture containing text

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | No | Yes |
| No | 41 | 8 |
| Yes | 18 | 24 |

CART:

Predicted, Actual

Accuracy: 71.4%  
Type I Error (False Positive): 16.3%  
Type II Error (False Negative): 42.9%

|  |  |  |
| --- | --- | --- |
|  | No | Yes |
| No | 34 | 15 |
| Yes | 15 | 27 |

Logistic Regression

Predicted, Actual

Accuracy: 67.0%  
Type I Error (False Positive): 30.6%  
Type II Error (False Negative): 35.7%

**Overall evaluation:**CART model seems to perform better than logistic regression. CART has a higher prediction accuracy of 71.4% compared to logistic regression at 67.0%. On average, this means that CART is able to predict whether a patient has or does not have AHD more accurately compared to logistic regression. CART’s overall error rate as well as type I error rate is also lower compared to logistic regression.

However, CART seems to perform poorer to logistic regression in terms of type II error. In this context, type II error would be more serious than type I errors, given that it would be more serious if the model predicts that a particular patient does not have AHD when he or she actually does, compared to the model predicting that a particular patient does have AHD when he or she actually does not. For type I errors, the repercussions would be wastage of resources which would be eventually be rectified once that particular patient has been correctly diagnosed while for type II errors, the repercussions would be life and death since the AHD would have been left untreated and will worsen. Hence, in this aspect, it is possible to consider that logistic regression could outperform CART when focussing on type II error rates.

1. Which of the two models would you recommend the Hospital to use? Explain.

I would recommend the Hospital to use CART nonetheless. Seen in question 3, CART has a better prediction accuracy compared to Logistic Regression. Though it is worse off in terms of type II error rates, there are many other advantages CART has over Logistic Regression in terms of model generation aspects.

CART is able to work with NA values while Logistic Regression cannot. In a data set of 300 observations, there are already 6 NA values encountered. Though we did clean it initially, CART saves us the hassle of cleaning NA values. Moreover, the cleaning of missing values often depends on first determining the behaviour of these missing values, whether they are missing completely at random, missing at random, or missing not at random, before using the best way to clean the data. Majority of the data cleaning process is determined by the analyst’s knowledge which may not always be accurate. Hence, CART saves the Hospital the trouble of data cleaning along with any other inaccuracies derived from it. Moreover, with bigger data sets that the Hospital definitely has given its huge database of patients, employees and medications, NA values are likely to occur in greater percentages (Singh, 2020).

Additionally, CART is more versatile than logistic regression, given that it can be used to predict both continuous and categorical variables where as logistic regression can only be used for predicting categorical variables and an additional classifier linear regression has to be implemented in order for the hospital to predict continuous variable. The versatility that comes with CART will definitely be advantageous to the Hospital given that they are not limited to predicting categorical variables such as AHD but can be able to predict other types of variables such as blood pressure and cholesterol which provides another angle of analysis which could prove to be useful and add to their research extending beyond just predicting AHD.

Moreover, CART uses 10 fold cross validation to train and tests the data, while Logistic Regression requires you to manually split your data set into training and testing data. This split determines the trade-off between overfitting your model and underfitting your model as well as model accuracy with training set and validation accuracy with test set or external data set. This may bring about further inaccuracies since analysts may not know the most accurate train test split ratio. For CART, it partitions the data set in 10 equal sizes, of which 1 is retained as validation data and 9 is used as training data. The cross validation process is then repeated another 9 times, with each of the 10 sub data sets used exactly once as validation data. The results from the 10 folds will then be averaged to give the final estimation. This makes CART superior to Logistic regression since CART trains and tests the model multiple times and the results produce would be a better indication on how it will perform on unseen data (Allibhai, 2018). Hence, with CART, it makes the process of train test splitting the data set redundant while raising prediction accuracies made

It is important to take note that our CART model was tested with the test data set derived initially from logistic regression only because since we are going to compare the accuracies of the 2 models, we would want to ensure consistency in doing so, which would of course require both models to be testing its accuracy against the same data set. In general, CART would be testing its accuracy against the original data set from which it was generated from.

1. Based on your chosen model above, explain the key findings to Hospital management.

Based on the CART model, we will be evaluating the variables deemed to be important and accurate predictors of AHD.

The first way to derive these variables is through the variable.importance or summary() method, to which it deduced the following variables as the most important factors in determining AHD (from most important to least important): Thal, ChestPain, MaxHR, Ca, Sex, ExAng, Slope, Chol, Age, and RestBP. Further filtering these values to correspond to the important variables identified from the summary() function, we obtain the following variables with measurements above 8 units:

1) Thal  
 2) ChestPain  
 3) MaxHR  
 4) Ca  
 5) Sex  
 6) ExAng  
 7) Slope

Another way to derive these variables is inspecting the variables used in the terminal nodes of the pruned CART model. Given the fact that the variables and the values used at the terminal nodes of the CART model produces the purest possible corresponding child nodes, this means that CART is able to use these values and confidently determine if that particular patient is suffering from AHD or not at the respective terminal nodes of the model. Thus, these are the following values extracted from the terminal node of the pruned CART model:

4) Ca=0,1 101 13 No (0.8712871 0.1287129) \*  
5) Ca=2,3 15 4 Yes (0.2666667 0.7333333) \*  
12) MaxHR>=143.5 21 6 No (0.7142857 0.2857143) \*  
13) MaxHR< 143.5 10 1 Yes (0.1000000 0.9000000) \*  
7) ChestPain=asymptomatic 65 7 Yes (0.1076923 0.8923077) \*

Hence from above, we derive the following:

|  |  |
| --- | --- |
| AHD = Yes | AHD = No |
| * Ca = 2, 3 * MaxHR < 143.5 * ChestPain = asymptomatic | * Ca = 0, 1 * MaxHR >= 143.5 |

Deriving the related variables from observing the terminal nodes is fairly consistent with the results obtain from the variable.importance method, given that Ca, MaxHR, ChestPain are within the top 5 most important variables identified by the CART. Furthermore, the variables identified by CART seems to also coincide with those identified by our logistic regression model, mainly Age, Sex, Chol, Ca and ChestPain based on odds ratio and confidence interval. Additionally, the variables selected also correspond to those identified to have a relationship with AHD status during exploration analysis and data visualisation via the various box plots and bar plots, further adding credibility and consistency to the results generated by CART.

From our generated CART model, it classifies patients with lower coloured major arteries as having no AHD will classifying them with higher coloured major arteries as having AHD. This is unsurprising as earlier proven in our data exploration and discussion of insights, major arteries coloured during the fluoroscopy procedure indicates clogging within these arteries and having more clogging at within major arteries would indicate anomalies and heart disease.

Apart from using the number of coloured major arteries detected during fluoroscopy, CART also uses maximum heart rate and asymptomatic chest pains as a metric to predict AHD which is viable, given that one of the symptoms of most heart diseases are averagely lower heart rates as well as this category of chest pains.

Additionally, medical conditions Thalassemia and Angina has also been selected as a predictor for AHD. This maybe because Thalassemia could weaken immunity systems making individuals more vulnerable to developing heart diseases (Mavrogeni, 2007) and that Angina could be a symptom of heart diseases itself (Government of Western Australia, 2019).

Lastly, given that ECG readings from treadmill stress tests often produce upward slopes and flat as well as downward slopes are considered anomalies (Rawat, 2019), behaviour of this slope recorded in ECG readings are indications of underlying health and heart issues and hence is consistent with the predictors selection of the “Slope” variable to determine AHD status.

However, it is important to take note that predictions generated from either CART or logistic regression are solely based on statistical figures which can be misleading (Lebeid, 2018). This problem is further amplified due to the nature of the data set that we are using. For example, small sample sizes such as the one used may project big numbers but have little statistical significance and representation due to its small size which may not be representative of the population (Ud-deen, 2020). The comparatively low accuracy rate and high type II error rate seen in CART and logistic regression is also another source of concern which may contribute to misleading findings. Hence, it is important for the Hospital to use any inference or predictions made by statistical models proposed such as in CART and logistic regression with precaution and with support from research and domain knowledge from the relevant experts to prevent their predictions from being counterproductive and working against them instead of for them (The Indian Express, 2019)

1. The cross validation error in CART is reported in the rpart package cp table. **If the outcome variable is continuous, this is fine. But if the outcome variable is categorical, an important information is missing.** What is the missing important information? Propose a way to obtain this information.

For continuous variable, CART uses mean to perform predictions, SSE (sum of square error) to determine best split and MSE (mean square error) to evaluate model performance whereas for categorical variables, CART uses majority to perform predictions, Gini Index to determine best split and misclassification error to evaluate model performance.

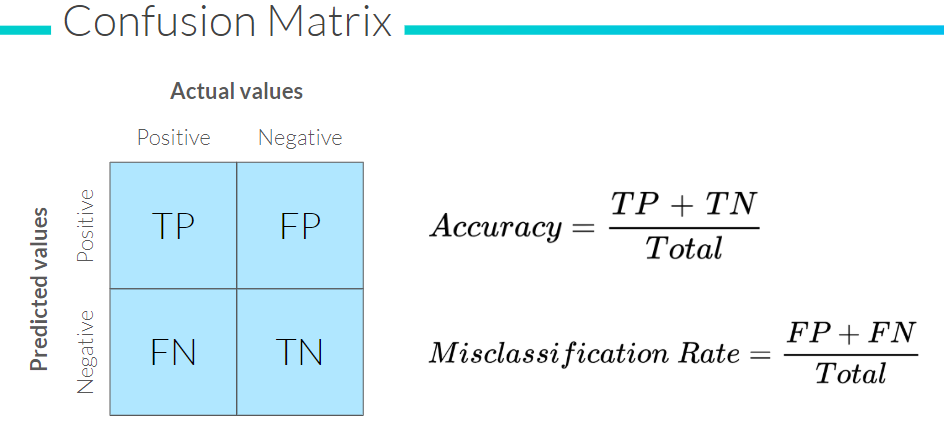
From the cp table created from the printcp() function, the following results are generated:

* 1. Root Node Error
  2. Complexity Penalty (CP)
  3. Node at which split occurs (nsplit)
  4. Trainset Error (rel Error)
  5. 10 Fold Cross Validation Error (xerror)
  6. Error Standard Deviation/1 Standard Error (Xstd)

Hence for continuous variables, based on the following formula, MSE and SSE can be derived (DataVedas, 2018):

Trainset MSE: Root Node Error \* rel error  
Trainset SSE: Trainset MSE \* n  
CV MSE: Root Node Error \* xerror  
CV SSE: CV MSE \* n

However based on the cp table, Gini Index as well as misclassification error of the model cannot be calculated. For misclassification error, we would need to compare predicted results against the actual data set and create a confusion matrix and derive the misclassification error rate from there (Testcraft, 2020):

  
*Credits: https://www.testcraft.io/artificial-intelligence-machine-learning-ebook/*

Hence, predicted results vs actual data in which we are predicting is missing from the cp table. To arrive at these values, we would need to filter out the column of the variable in which we are predicting from the actual data set. Using this, we would then create a 2 dimension table with the prediction results from CART via the table() function. Upon doing so and HENCE creating the confusion matrix, we would be able to derive the False Positive (FP) and False Negative (FN) values which would eventually help us calculate the misclassification rate for our CART model.

With reference to our earlier example in question 3, the misclassification error rate would be: = 28.6%

For Gini impurity of the model, we can first calculate the Gini indices for every node within the generated CART model and take the average of those Gini indices to find the Gini impurity of the overall CART model. Given that Gini Index is derived from the following formula: (Zhou, 2019), and that we are calculating the individual Gini indices for every node with only 2 possible outcomes (AHD = “Yes” and AHD = “No”), Gini impurity = p(AHD=“Yes”)[1 – p(AHD=“Yes”)] + p(AHD=“No”)[1 – p(AHD=“No”)] = 1 – – . The necessary p values are also missing from the cp table and we would need to manually hardcode the individual Gini impurity at each node with the help of the frame attribute from the generated CART model.

Graphical user interface, text, application

Description automatically generated

Hence, we arrive at an average Gini index of 0.343 for our generated pruned CART model from question 3.

Moving forward, another impurity metric that can be used by CART is entropy which is derived from: . Similar to what we did for Gini impurity, we are calculating entropy at every node before taking the mean to formulate the average entropy of the CART model, Entropy impurity = – + [–]. Again, these p values are also missing from the cp table and we would need to manually hardcode the individual entropy impurities at each node of the CART model with the help of the frame attribute again.

Graphical user interface, text, application

Description automatically generated

1. Read the research paper S.Chellammal, R. Sharmilax (2019). Comment on their approach. *Note: It is not required to know the techniques (MLP, SMO and NB) used in the research paper. Comment on their approach based on techniques and concepts learnt in any first course in Statistics module (e.g. AB1202 or equivalent) and this module.*

The research paper’s approach starts of by initially identifying the factors which are deemed to be important metrics for predicting AHD and creating the respective classifiers MLP, SMO and NB via correlation. After arranging the variables from highest correlation to AHD to lowest, there are then sequentially added from the highest correlation to the lowest to the classifiers, before each of the models generates its prediction on AHD status. Since there are a total of 13 variables, each model will be generated 13 times, adding the next highest correlation variable on the next iteration. After which, we identify the combination of variables which gives the highest prediction accuracy for each of the 3 classifiers respectively.

The second part of the approach is to fit variables into the 3 classifiers based on expert literature and evaluating each of the classifiers prediction accuracy for each combination of variables proposed by the respective referenced research paper and literature. Amongst this are a set of 10 variables which coincide with the combination of variables initially used to fit the classifiers during the first phase of the approach which eventually produced the best predictions for classifier NB & SMO.

The pros of this research paper is that it references external material and relies on expert literature on ascertaining the variables important to predicting AHD. It recognises the pitfalls of misleading statistics and predictions made by models as previously discussed and hence implements 5 different combinations of variables suggested by other research papers. They would then compare the accuracies of these generated predictions against the ones created based on correlation which would further enhance the reliability of the initial predictions made on correlation or identify its flaws and inaccuracies

Moreover, it also proposed 3 different classifiers and identifies the 3 respective combinations of variables which produces the best predictions for each of the classifier, which further increases the objectivity of its proposals and suggestions, making them more all-encompassing and definitive in eventually selecting the set of variables which have the best prediction accuracies in determining AHD.

The cons of this research paper is that it uses correlation as the basis of fitting variables into the classifiers, which creates issues and further raises the possibility of multicollinearity.

It is important to realise that correlation does not imply causation (Madhavan, 2019) and variables with a high correlation to AHD may not necessarily mean that these variables causes AHD and could hence be great predictors for it. Though causation implies correlation, but the reverse is not always true and the correlation observed may be a result of coincidence. To use correlation as a metric to hypothesize and select the variables to create our classifiers could be dangerous since it creates unrelated links between our factors and our prediction variable (Borwein, 2014).

Additionally, basing the accuracy of the predictors on correlation further creates issue when fitting the variables into the model. Theoretically, there are 13 x 13 different combinations of variables which could be fit into our model and in implementing the above approach, we are only testing 13 out of the many combination of variables which may not even be accurate in the first place given that correlation may not be the best metric to fit variables into our classifiers.

Moreover, adding variables based on its correlation to AHD could contribute to multicollinearity, where the independent factors used for predicting AHD are correlated to one another, which creates redundant information that could skew the results and accuracies of the classifier (Statistics How To, 2015)

Moving forward, instead of using correlation, there are many other ways of deriving the best combination of variables to use for their classifiers. As demonstrated earlier, we can identify important variables to predicting AHD via CART’s variable.importance method or analysing ita terminal nodes. Alternatively they could look into logistic regression and eliminate statistically insignificant variables at 5% level or determine important predictors via odds ratio and further validating this by ascertaining the confidence intervals of these odds ratios. They could also look into selecting variables based on the most accurate model with the lowest Akaike Information Criterion (AIC) which is a metric which determines how well a model fits into the data it is made from. The lower the AIC value, the better the model fits. As seen below, for logistic regression, the step() function is implemented to obtain these sets of information and acquire the logistic regression equation with the lowest AIC value and hence best fit model, at the end of which it determines that logistic regression model with Age, Sex, ChestPain, RestBP, Chol, MaxHR, Oldpeak, Ca, Thal

Table

Description automatically generated

Table

Description automatically generated  
A screenshot of text

Description automatically generated

**REFERENCES:**

Ai Multiple (2020, September 22). *Data Cleaning: What it is, why it matters, best practices & tools.* Retrieved 21 October 2020 from, <https://research.aimultiple.com/data-cleaning/>

Allibhai, E. (2018, October 3). *Hold Out vs Cross Validation in Machine Learning.* Retrieved 27 October 2020, from https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

Borwein, J. (2014, September 22). *Clearing up the confusion between correlation and causation.* Retrieved 28 October 2020, from <https://theconversation.com/clearing-up-confusion-between-correlation-and-causation-30761>

Bosner, S. (2009, December 14). *Gender differences in presentation and diagnosis of chest pain in primary care.* Retrieved 25 October 2020, from https://www.ncbi.nlm.nih.gov/pmc/  
articles/PMC2801475/

Carson, J. A. S. (2019, December 16). *Dietary Cholesterol and Cardiovascular Risk: A Science Advisory From the American Heart Association.* Retrieved 23 October 2020, from <https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000000743>

Center for Disease Control and Prevention (2020, May 19). *High Blood Pressure Symptoms and Causes.* Retrieved 13 October 2020, from https://www.cdc.gov/bloodpressure/about.htm

Cleveland Clinic (2020, July 31). *Cholesterol Numbers: What Do They Mean?* Retrieved 21 October 2020, from https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

Datavedas (2018, June 9). *Model Evaluation – Regression Models.* Retrieved 27 October 2020, from https://www.datavedas.com/model-evaluation-regression-models/

Diabetes in Control (2002, April 22). *Fasting Blood Sugar above 90 puts you at risk of heart disease.* Retrieved 1 November 2020, from <http://www.diabetesincontrol.com/fasting-blood-sugar-above-90-puts-you-at-risk-of-heart-disease/>

Dorshkind, K. (2013, March 1). *Causes, Consequences and reversal of immune system aging.*  Retrieved 19 October 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3582124

Friedberg, C. K. (1961, May 1). *“Non-specific” ST and T-wave Changes*. Retrieved 17 October 2020, from <https://www.ahajournals.org/doi/abs/10.1161/01.cir.23.5.655>

Gimenez, L. (2018, May 24). *6 Steps for Data Cleaning and why it matters.* Retrieved 23 October 2020, from https://www.geotab.com/blog/data-cleaning/

Government of Western Australia (2019, July 1). *Angina.* Retrieved 31 October 2020, from https://healthywa.wa.gov.au/Articles/A\_E/Angina

Harvard Medical School (2016, November). *Throughout life, heart attacks are twice as common in men than women.* Retrieved 10 October 2020, from https://www.health.harvard.  
edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women

Hazzard, W. R. (1985, January). *Atherogenesis: Why women live longer than men?* Retrieved 27 October 2020, from https://pubmed.ncbi.nlm.nih.gov/3965355

Healthline (2019, February 26). *Muscular Hypertrophy and Your Workout.* Retrieved 23 October 2020, from https://www.healthline.com/health/muscular-hypertrophy#definition

Khare, J. (2018, February 22). *Exploratory Data Analysis and its importance to your business.* Retrieved from https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/

# Learning By Marketing (2020). *The Gini Index.* Retrieved 13 October 2020, from http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/ Lebeid, M. (2018, August 8). *Misleading Statistics Examples – Discover the potential for misue of statistics & data in the digital age.* Retrieved 31 October 2020, from https://www. klipfolio.com/blog/how-to-spot-misleading-data Madhavan, A, (2019, September 20). *Correlation vs Causation: Understand the difference for your product.* Retrieved 1 November 2020, from https://blog.amplitude.com/causation-correlation Mavrogeni, S. (2007, February 20). *Cardiac aspects of ß-Thalassemia*. Retrieved 21 October 2020, from <https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-5/Cardiac-aspects-of-Thalassemia-Title-Cardiac-aspects-of-Thalassemia> Mayoclinic (2019, November 22). *Thalassemia.* Retrieved 5 November 2020, from https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995

MedlinePlus (2020, August 13). *Angina.* Retrieved 5 November 2020, from

https://medlineplus.gov/angina.html

MedlinePlus (2020, September 29). *Fluoroscopy.* Retrieved 2 November 2020, from https://medlineplus.gov/lab-tests/fluoroscopy/

Min Ma (2014, August 27). *Classification and Regression Trees with rpart and rpart.plot.* Retrieved 1 November 2020, from https://rstudio-pubs-static.s3.amazonaws.com/27179\_e64f0de316fc4f169d6ca300f18ee2aa.html

Molala, R. (2019, October 14). *Entropy, Information Gain, and Gini Index; the crux of a Decision Tree.* Retrieved 1 November 2020, from https://blog.clairvoyantsoft.com/entropy-information-gain-and-gini-index-the-crux-of-a-decision-tree-99d0cdc699f4

Panzeri, S. (2015, March 14). *Sampling Bias.* Retrieved 19 October 2020, from http://www.scholarpedia.org/article/Sampling\_bias

Patadiya, R. (2019, July 5). *Gini Index – CART Decision Algorithm in Machine Learning.* Retrieved 30 October 2020, from https://medium.com/@riyapatadiya/gini-index-cart-decision-algorithm-in-machine-learning-1a4ed5d6140d

Rajwanshi, K. (2018, May 9). *Data Cleaning: A guide to deal with NA values.* Retrieved 29 October 2020, from <https://www.linkedin.com/pulse/data-cleaning-guide-dealing-na-values-karan-rajwanshi/>

Rawat, S. (2019, August 10). *Heart Disease Prediction.* Retrieved 1 November 2020, from <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>

Singh, S. (2020, May 22). *Handling Missing Data.* Retrieved 1 November 2020, from https://towardsdatascience.com/handling-missing-data-12ac7c59eebc

Statistics How To (2015, September 15). *Multicollinearity: Definition, Causes, Examples.* Retrieved 17 October 2020, from https://www.statisticshowto.com/multicollinearity/

The Indian Express (2019, May 3). *Using AI to detect mental health disorders may be counterproductive.* Retrieved 24 October 2020, from https://indianexpress.com/article/  
lifestyle/health/using-ai-to-detect-mental-health-disorders-may-be-counterproductive-5708568/

Tyagi, N. (2020, March 24). *Understanding the Gini Index and Information Gain in Decision Trees.* Retrieved 27 October 2020, from https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8

Ud-deen, H. (2020, March 30). *Misleading Statistics & Data: How To Protect Yourself Against Bad Statistics.* Retrieved 19 October 2020, from https://www.klipfolio.com/blog/  
how-to-spot-misleading-data

Wong, C. K. (2002, January). *Recognizing “painless” heart attacks.* Retrieved 23 October 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1766972/

Wong, W. (2019, January 11). *Data Cleaning, Detection and Imputation of Missing Values in R Markdown — Part 2.* Retrieved 15 October 2020, from <https://towardsdatascience.com/data-cleaning-and-detecting-missing-values-from-australian-open-tennis-data-in-r-part-2-b4f6f32beffd>

Zhou, V. (2019, March 29). *A Simple Explanation of Gini Impurity.* Retrieved 29 October 2020, from https://victorzhou.com/blog/gini-impurity/