# A picture containing text Description automatically generated

BC2406: Analytics II

Computer Based Assignment:

Heart Disease Diagnostic with Analytics

Submitted by: Ng Chi Hui

Email Address: NGCH0118@e.ntu.edu.sg

Matriculation Number: U1922243C

### Nanyang Business School

# Introduction

A hospital would like to assess the usefulness of Analytics models in predicting heart disease and provided a set of data as recorded in AHD.csv.

A data dictionary is provided in Appendix A. AHD is the outcome variable to be predicted. A recent research paper on this dataset is included as a PDF file.

# Part A: Data Exploration and Preparation (30%)

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

Before data exploration can be carried out, data preparation and data cleaning have to be carried out first. The CSV file was first read into the R followed by factorization of the various variables as stipulated by the question. Following which several steps are taken to ensure the data is clean and structured enough to be analyzed which are stated as follows:

1. Renaming of factor variables into their representative categories for easier understanding and factorization of factors
2. Detection of NA/missing values and carrying out the appropriate actions to handle these values
3. Outlier Detection and appropriate actions to handle these outliers

**Step A: Renaming of Factor Variables and factorization of variables**

This step was carried out mainly to ensure easier understanding when plotting different graphs for visualization during data exploration. It was performed by changing the data type from numeric to strings. For example, initially Sex was coded as 0 and 1 which was later change to 0 for Female and 1 for Male.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Step B: Detection of NA/Missing Values and Handling**

**A picture containing chart

Description automatically generated**

There were 6 NAs in the entire dataset of which 4 belongs to Ca and 2 belongs to Thal. NAs values means that there exists a value, but this value is unknown for now. Both Ca and Thal are categorical variables which are easier to treat. The NA values consists of 4/303 and 2/303 of the entire dataset for Ca and Thal respectively. This is only 1.3% and 0.66% of the entire dataset respectively. These NA values are rather insignificant relative to the entire dataset. However, since the entire dataset is already rather small with only 303 rows, it will be better to replace the values with the value that appears most often in the dataset instead of deleting the entire row to prevent further reduction of data available. For Ca, both the median and mode are 0 thus the NA values are replaced with 0s. The mode of Thal is “normal” thus NAs in Thal columns are replaced with “normal”. Mode is used as a key metric for cleaning both Thal and Ca.

**Step C: Outliers Detection and Handling**

Having outliers in the dataset may affect the machine learning models and the analysis of various results in the dataset. Outliers in the dataset should be dealt with carefully and appropriately. When such erroneous data are found they should be removed (Grace, 2018). One of the ways to detect outliers is by using visualizations tools such as boxplots and also through using statistical methods like the IQR and median (Sharma, 2018). That being said, since this is a medical dataset thus each individual case or patient will have varying medical history and it will not be abnormal to have extreme anomalies. Thus, more care and considerations should be taken when identifying if values are anomalous and on whether they should be removed. The outlier though anomalous may be kept unless it is out of the norm even in contexts of abnormal cases presented in the medical world.

|  |  |
| --- | --- |
| **Boxplot of the Continuous Variables** | **Explanation and Handling Methods of Outliers** |
|  | No outliers were detected thus no data handling was required. |
|  | The summary function showed that the highest resting BP was 200 and there were 9 outliers with resting BP above 170.  To determine how these outliers should be dealt with various literature reviews were conducted. It was found that blood pressure above 180 or in the range indicated hypertensive crisis which requires urgent medical help (Sheps, 2019).  This value was recorded when the patient is resting, and it is unlikely for patients who are resting to have such high bp. If their blood pressure is in this range, it is likely that they already have a stroke. The readings were deemed erroneous and removed from the data set. |
|  | The boxplot for Serum Cholesterol also revealed the presence of outliers. Most of these outliers are in the 400 range.  On further research, the serum cholesterol levels are considered high when it is >240 mg/dl. (SingHealth, n.d.). However, cholesterol levels of more than 400 might be abnormal but it is still entirely possible in poorly managed patients (Bekerman, 2020). Also, higher cholesterol >240 mg/dl has been proven to double the risk of heart disease which is our predictor value. Thus, though uncommon, these values are not erroneous and thus should not be removed from the data. |
|  | Under MaxHR, one outlier value present at 71 bps.  While a MaxHR of 71 is extremely low, however, that is not entirely impossible. This might be due to insufficient exertion during the test or that the age of the person is relatively larger. The higher the age of the person the slower the max heart rate (Christou, D. D., & Seals, D. R. ,2008)  The MaxHR of 71 here belongs to a 67 year old Male which is quite old. Thus, this outlier is not removed. |
| **Chart, box and whisker chart  Description automatically generated** | No outliers removed as it is medically possible to have Oldpeak values of upwards of 4 (Neshat, 2010) |

At this step, data cleaning has been completed, and the data can be used for exploratory Analysis. The data exploratory analysis was mainly split into 2 phases, Univariate analysis and Bivariate Analysis. Below present some interesting findings during data exploration.

|  |  |
| --- | --- |
| Chart, radar chart  Description automatically generated | **Relationship between continuous variables and AHD, a categorical variable**  Since the problem statement is about AHD we need to focus on variables that affects AHD. The correlation between continuous and categorical variables cannot be mapped out easily such as using the cor( ) function thus, thus visualization was used  **AHD by Age**  In the age against AHD graph, we see that AHD is present in people who are generally older at least 55 years old.  By using a violin plot, we are able to see the distribution of the data. We observed there will be spike in the numbers who have AHD when the age is in the 60s range  **AHD by Chol**  Looking at the median Cholesterol of those with and without AHD, there is no significant difference. Perhaps, cholesterol levels is not highly correlated to risk to AHD. Over a small range of cholesterol values, between 225 and 225, the probability of AHD is higher.  **AHD by Max HR**  MaxHR of those without AHD is significantly higher than those with AHD. Younger people are more likely to have higher MaxHR and thus it is also likely that younger people are less likely to have AHD. (Christou, 2008)  **AHD by Oldpeak**  Since the outliers in Oldpeak are not removed due to reasons mentioned in front, the distribution is actually right skewed. From the violin box, there are significant number of people with low Oldpeak not having AHD. Likewise, it is likely that those with higher Oldpeak have higher probabilities of having AHD.  **RestBP**  The medians of RestBp is quite similar for those who has AHD and those without AHD. The distribution of datapoints is quite similar across those who have and not have AHD. There seem to be little correlation between AHD and RestBp. |
| Using a simple correlation matrix, we are able to quickly see the relationship between continuous variables Chart  Description automatically generated   1. MaxHR has a moderately strong and negative correlation with Age and Oldpeak 2. Chol has relatively weak correlation with the rest of the continuous variables 3. RestBP has relatively weak correlation with the rest of the continuous variables   Note: The description of strong, weak and moderately strong is relative to the correlation numbers in the matric where moderately strong is use to describe correlation of plus minus 0.3 to plus minus 0.4.  With that in mind, we can choose to focus our analysis on MaxHR, Oldpeak and Age. | |
| Chart, scatter chart  Description automatically generated | As age increases, MaxHR decreases. Patients with lower MaxHR are more likely to have AHD. If a patient has AHD, MaxHR will decrease more slowly with age compared to those without AHD. |
|  | Upon further analysis, it appears that those that have RestBp > 150 are more likely to have AHD |
| AHD against Categorical Variables | |
| Chart  Description automatically generated | The dataset consists of a higher proportion of Males than Females. By using a side by side bar plot, we are able to know the relative proportions of those with AHD within each sexes.  It’s observed that males are more likely to suffer from heart disease, AHD compared to their females counterparts. Thus sex might be a plausible predictor for AHD. |
| Chart, bar chart  Description automatically generated | We are able to observe that the most common occurrence of Chest Pain is asymptomatic followed by non-anginal, atypical and then typical.  Asymptomatic pain seems to be a major cause of AHD. Whereas in other pain types like non-typical pain, atypical pain and typical pain, there are significantly lesser people with AHD. It is very likely that asymptotic chest pain will be a significant predictor of AHD later in the models. |
| Chart, bar chart  Description automatically generated | The distribution of data of patients with FBS <120mg/dl is significantly higher than those with FBS of > 120 mg/dl.  We can broadly classify those with FBS <120 mg/dl not having diabetes and FBS > 120 mg/dl having diabetes.  There is no significant relationship between FBS and the presence of AHD. Thus, preliminarily, it will be unlikely that FBS will be a predictor for AHD. |
| A picture containing graphical user interface, chart  Description automatically generated | Patients with abnormal RestECG - ST-T Wave Abnormality and Left Ventricular Hypertrophy are more likely to get AHD. Whereas, those with normal RestEcg are unlikely to have AHD.  The dataset seems to contain an extremely small proportion of ST-T Wave Abnormality cases and the unbalanced data might be concerning. |
| Chart, bar chart  Description automatically generated | The dataset consists of more cases of no exercise induced angina  Exercise Induced Angina seems to be correlated those having AHD. More patients with ExAng have AHD. While those without ExAng are less likely to have AHD. |
|  | People with upsloping Peak Exercise ST Segment are less likely to have AHD. While those with flat and downsloping ST segments are more likely to have AHD |
| Chart  Description automatically generated | When fluoroscopy was performed if the number of Ca being >= 1, the proportion of those with AHD was significantly higher. Whereas those without Ca  (Ca = 0) are more likely not to have AHD |
| Chart, bar chart  Description automatically generated | It can be observed that those without thalassemia meaning normal cases have a higher proportion of cases without AHD.  In both fixed and reversable Thalassemia, there were more people with AHD. |

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

|  |  |  |
| --- | --- | --- |
| Summary Table | | |
| **Variables** | **Diagram** | **Description & Key Statistics** |
| Age |  | Age is normally distributed. Most datapoints are in the age between 50s to 60s.  Mean Age: 54.29252  Min Age: 29  Max Age: 77  Median Age: 55  There is only 1 data point in age group 20-30 and very few young adults.  **Older people are more likely to have AHD** |
| Sex |  | Dataset consists of significantly more males than females  **Count:**  Males: 202  Females: 92  **Males are more likely to have AHD compared to females** |
| Chest Pain | Chart, bar chart  Description automatically generated | **Count:**  Asymptomatic: 139  Non Anginal Pain: 84  Atypical Angina: 49  Typical Angina: 22  **Patients with asymptomatic pain more like to have AHD whereas those with the other 3 types of pains are less likely to have AHD** |
| RestBp |  | RestBp is normally distributed in the dataset.  RestBp > 150 are more likely to have AHD  Mean: 130.1633  Median: 130  Min: 94  Max: 170  **Those with RestBP > 150 are more likely to have AHD** |
| Chol |  | Mean: 245.9354  Median: 240  Min: 126  Max: 564  **Those with higher cholesterol are more likely to have AHD though over a very small range of values** |
| Fbs |  | **Count**  FBS < 120 mg/dl (Normal): 253  FBS > 120 mg/dl (Normal): 41  **No significant relationship between FBS and the presence of AHD.** |
| RestECG |  | Significantly more datapoints of those with normal and left ventricular hypertrophy.  **Count:**  Normal: 147  ST-T-Wave Abnormality: 3  Left ventricular Hypertrophy: 144  **Patients with abnormal RestECG, those with ST-T Wave Abnormality and Left Ventricular Hypertrophy are more likely to get AHD.** |
| MaxHR |  | Mean: 149.5476  Median: 153  Min: 71  Max: 202  **MaxHR of those without AHD is significantly higher than those with AHD. Age and Oldpeak has a strong and negative correlation with MaxHR.** |
| ExAng |  | **Count:**  No Exercise Induced Angina: 201  Exercise Induced Angina Present : 93  **Exercise Induced Angina seems to be correlated those having AHD. People with Exercise Induced Angina are more likely to have AHD** |
| Oldpeak |  | Skewness: 1.271397  Distribution is right skewed and very high proportion of 0 relative to the rest of the distribution.  Mean: 1.021429  Median: 0.8  Min: 0  Max: 6.2  **People with higher Oldpeak values are more likely to have AHD** |
| Slope |  | Distribution contains of significantly higher number of points for upsloping and flat slope.  **Count**  Upsloping: 139  Flat: 136  Downsloping : 19  **People with upsloping Peak Exercise ST Segment are less likely to have AHD. While those with flat and downsloping ST segments are more likely to have AHD** |
| Ca |  | **Count**  0 : 174  1 : 64  2 : 36  3 : 20  **When fluoroscopy was performed i.e. the number of Ca being >= 1, the proportion of those with AHD was significantly higher. Whereas those without Ca performed are more likely not to have AHD** |
| Thal |  | **Count**:  Fixed: 18  Normal: 165  Reversable 111  **Normal cases have a higher proportion of cases without AHD. Whereas, those with fixed and reversable Thalassemia, are more likely to have AHD** |
| AHD |  | AHD is the target variable  **Count**  No: 161  Yes: 133 |

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

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

The two models that were chosen are the CART model and the logistic regression model. Both model used the same train set to train and same test set to test the performance of each model. This is because later on, we will like to compare between the models and this ensures fairness and less biasness being introduced. Also, to ensure standardization and reproducibility seed was set to 2 for the train test split as well as for the 10 fold cross validation where we want to ensure randomness and all other random assignments. The ratios for the split of dataset is 70% for training and 30% for test.

The train set consists of a total of 206 samples of which 113 samples without AHD and 93 with AHD. The train set split is approximately 55: 45 without AHD and with AHD respectively thus it is not necessary to balance the dataset with tools like SMOTE as it is already quite balanced.

**CART Model:**

**Step 1: Growing the tree to the Maximal**

The rpart function was used to perform CART analysis



The method was set to “class” as we are trying to predict a categorical variable, AHD here. The default cp was changed from 0.01 to 0. This is to allow the tree to grow to the maximum to be pruned later. When cp = 0 no penalty will be imposed on growing the tree allowing it to grow to its maximum.

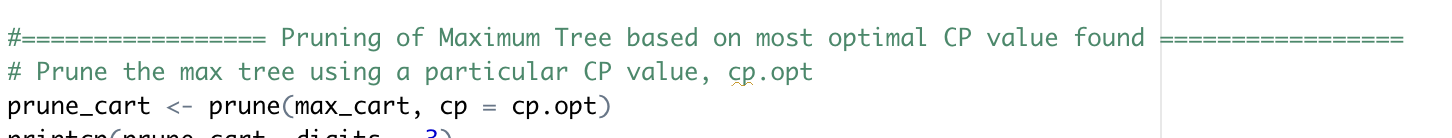
|  |  |
| --- | --- |
|  | This is a screen capture of the maximal tree obtained however this tree should not be used as it will almost always result in overfitting. It also does not provide any useful information that aid in the user’s understanding |

**Step 2: Pruning the tree**

**Chart

Description automatically generated**

To prune the tree, we use the 1 SE rule , the best cp value will be calculated by taking minimum CVerror + 1SE. The CVerror + 1SE can be seen by the dotted line. In this case, the optimal complexity parameter used will be 0.0395079.

****

**Step 3: Printing of the optimal tree for analysis**

**Diagram

Description automatically generated**

**Analysis of Variable Importance**

**Chart

Description automatically generated**

**Accuracy of the CART Model:**

**Confusion Matrix for CART Training Data Set**

|  |  |  |
| --- | --- | --- |
|  | CART Train Set Prediction | |
| Train Set Actual | No | Yes |
| No | 96 | 17 |
| Yes | 14 | 79 |

Overall Accuracy of Train Set: 0.8495146

**Confusion Matrix for CART Test Data Set**

|  |  |  |
| --- | --- | --- |
|  | CART Test Set Prediction | |
| Test Set Actual | No | Yes |
| No | 42 (TN) | 6 (FP) |
| Yes | 9 (FN) | 31 (TP) |

Overall Classification Accuracy on Test Set: 0.8295455

1. Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN = (6 + 9)/(42 +6 + 9 + 31) = 0.170454545 = 17%
2. Precision (**true** positives / **predicted** positives) = TP / TP + FP = 31/ (31 + 6) = 0.837837838 = 83.8%
3. Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP = 42/(42 +6)= 0.875 = 87.5%
4. TPR = TP / TP + FN = 31/(31+9) = 0.775 = 77.5%
5. TNR = TN/FP + TN = 42/ (6+42) = 87.5%
6. FPR = FP/ (FP + TN) = 6/ (6+42) = 12.5%
7. FNR = FN/(TP+FN) = 9/ (31+9) = 22.5%

**Logistic Regression**

The Logistic Regression helps us model after a binary variable, in this case it will be AHD. In this study, backwards stepwise selection method was being performed. (Choueiry, n.d.) We started by considering the model that contains all the variables at each step of the model, we choose the most insignificant variable to remove until the final equations contains only significant values with threshold p-value determined by the user.

To simply the codes, I have chosen to use the function step with the direction backward to perform the steps. It produces relatively similar to doing it manually. The manual implementation can be found it the R script.



Table

Description automatically generated

In this study, although not all variables were statistically significant at 5% alpha, however, it was decided that these variables shall remain. For example, variables like MaxHR and Oldpeak though they show low statistical significance as we do not have enough domain knowledge to remove them.

Following which we can check for multi-collinearity by using the VIF (Variance Inflation Factors). Looking at the rightmost column since most variables have values that are less than 2, we will not be removing any values

A picture containing text

Description automatically generated

**Confusion Matrix for Logistic Regression Training Data Set**

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression Train Set Prediction | |
| Trainset Actual | No | Yes |
| No | 103 | 10 |
| Yes | 20 | 73 |

Overall Classification Accuracy on Train Set: 0.8543689

**Confusion Matrix for Logistic Regression Test Data Set**

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression Test Set Prediction | |
| Trainset Actual | No | Yes |
| No | 46 (TN) | 2 (FP) |
| Yes | 12 (FN) | 28 (TP) |

Overall Classification Accuracy on Train Set: 0.8409091

1. Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN = (2 + 12)/(46 +2 + 12 + 28) = 15.9
2. Precision (**true** positives / **predicted** positives) = TP / TP + FP = 28/ (28 + 2) = 93.3%
3. Specificity (**true** negatives / all **actual** negatives) = TN / TN + FP = 46/(46 +2) = 95.8%
4. TPR = TP / TP + FN = 28/(28+12) = 70%
5. TNR = TN/FP + TN = 46/ (2+46) = 95.8%
6. FPR = FP/ (FP + TN) = 2/ (2+46) = 4.2%
7. FNR = FN/(TP+FN) = 12/ (28+12) = 30%

**Summary of Comparison of Accuracy Measures across both Models**

|  |  |  |
| --- | --- | --- |
|  | **CART Model** | **Logistic Regression** |
| **True Positive Rate (TPR)** | **77.5%** | 70% |
| **Type 1 Error (FPR)** | 12.5% | **4.2%** |
| **Specificity** | 87.5% | **95.8%** |
| **Type 2 Error (FNR)** | **22.5%** | 30% |
| **TNR** | 87.5% | **95.8%** |
| **Precision** | 83.8% | **93.3%** |
| **Overall Model Accuracy** | 83% | **84%** |
| **Misclassifications Error** | 17% | **15.9%** |

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

I will recommend the CART model to the hospital. Although base on the models, logistic regression boost a slighter higher overall model accuracy. However, there are several alarming statistics such that CART is preferred.

First, CART has a higher TPR of 77.5% compared to that of Logistic Regression 70%, it is 7.5% higher which is quite a large margin. The TPR refers to when the patient has AHD and is being correctly identified. CART also boost a lower FNR .With a lower FNR, this means that when the patient has AHD there will be lesser chance of them being incorrectly identified as not having the disease. With AHD, often going unnoticed until the condition of the patient significantly worsen such as only when the person have a heart attack or when one starts experience chest pains between 70 to 90% of the blood vessel is already blocked (Publishing, 2020), there are severe consequence of being unable to detect AHD accurately. Since hospitals are often dealing with life and death decisions, accuracy is an important factor for consideration. The model for this kind of prediction problem should be that there needs to be a high TPR so that patients that do not show signs of heart blockage can be detected early and treated. It also requires a low FNR, so that patients who have AHD will not be wrongly marked as not having AHD.

Also, although CART has a slightly higher misclassification error, it is mainly due to the significantly higher false positive rate. In the context of predicting heart disease, it’s better to have higher FP rate that a high FN rate. A high FP only means a false alarm for the patient whereas a high FN will likely cause a life as patient with AHD are incorrectly classified as not having AHD.

The second reason that cart is preferred is due to the assumption that medical staff are not statistically trained, CART produces a decision tree which is relatively much simpler for the layman to understand. To understand logistic regression one needs to understand statistics such as odds ratio and the medical staff might not be able to interpret the model correctly. (Bock, 2020). With CART, medical staff just needs to read from top down to make an accurate decision or prediction of whether the patient is likely to have AHD.

Lastly, CART is able to automatically handle missing values thus this require less work from the hospitals to spend large amount of time to clean the dataset that they have collected.

In conclusion, CART should be the model that is recommended to the hospitals due to higher TPR and FNR though its overall model prediction is lower than logistic regression as the difference is trivially small. Also, with the ease of interpretation and the lack of need to do tedious data cleaning CART should be recommended to hospitals.

That being said, despite recommending CART, more research should be done before this can be implemented. This is because there are several limitations in my study. First, we are limited by the dataset. The current dataset used by this study is too small for proper analysis and machine learning. It’s was noted that by choosing a different ratio for the train test split will skew my model prediction significantly, in fact, by change the train test split from 70:30 to 80:20 results in better prediction results for logistic regression. The current statistics presented in here are remotely similar between CART and Logistic Regression and thus are not accurate representation of the entire population/patients group. Thus, with a larger dataset, the results might result in completely different.

There were also several key assumptions being made, primarily on cleaning the outliers. To better handle the data, perhaps it will be better to rely on expert opinion. Second, the threshold values chosen here was based of common threshold values used. Perhaps, industry standards threshold values will give us better accuracy on the data or even with the use of methods like ROC curve to measure the performance of classification model and to set an ideal threshold can improve the results of the dataset.

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

Chart

Description automatically generated

Thalessemia is the most significant variable in predicting AHD which is shown in the bar plot as well as the root node of the pruned tree followed by Old Peak and Chest pain which is shown in level 2 of the tree assuming root node is level 1 so on and so forth.

Diagram

Description automatically generated

The tree provides and intuitive way for hospital staff to filter through patients who have AHD. Reading the tree from top to bottom and by taking the patient history the hospital could easily predict who will be likely for have AHD.

1. If the patient has thalassemia of normal type, Chest pain of either non anginal pain or atypical pain, the patient has 9% chance of having AHD
2. If the patient has thalassemia of normal type but does not have chest paint types of either non anginal pain or atypical pain (meaning chest pain of either typical and asymptomatic) , and if the number of major vessels coloured fluoroscopy is equals to zero, he or she has 22% chance of AHD
3. If the patient has thalassemia of normal type but does not have chest pain of either non anginal pain or atypical pain (meaning chest pain of either typical and asymptomatic), and if the number of major vessels coloured fluoroscopy is not equals to zero(meaning Ca of either 1/2/3), he or she has 71% chance of AHD
4. If the patient does not have thalassemia of normal type (meaning thalassemia of either fixed defect or reversable defect) but has oldpeak of <0.7 and Chol of <241, he or she will have 15% chance of getting AHD.
5. If the patient does not have thalassemia of normal type(meaning thalassemia of either fixed defect or reversable defect) but has oldpeak of <0.7 and Chol of >241, he or she will have 77% chance of getting AHD.
6. If the patient does not have thalassemia of normal type(meaning thalassemia of either fixed defect or reversable defect) and has oldpeak of > 0.7 they have a 87% chance of AHD

Note that the tree only provides a good basis for prediction of AHD base on the various factors mentioned. However, this should not be substitute for actual testing. It was also noted that there was a flaw with this CART model in which on part requires results from fluoroscopy in medical terms known as angiogram fluoroscopy, however if fluoroscopy is to be carried out there will be no need for this prediction as the doctors will be able to use the results to directly detect if the patient suffers from heart disease

In conclusion, CART is the recommended model in this study. However, in order to reach a more definitive answer as to which model is the best. There needs to improvement in the data accuracy and data size.

Regarding the data accuracy, as much as possible the data collection team needs to ensure values are entered correctly and there’s no outliers. Regarding the outliers, in this dataset there was an extremely low MaxHr value though not impossible, medical professionals collecting the data need to ensure that patient is exerting their heart to the maximum as it is likely that the patient did not exert enough.

Secondly, regarding data size, it was earlier mention that this model was built with limited data and this could limit the results achieve in terms of accuracy of models and representation of the population. For instance, in the age category the youngest person is 29 years old thus this model may not be applicable to predict AHD in babies or people younger that. The models implemented are likely not accurate if we were to predict AHD for younger persons in general. Thus more datapoints should be collected to allow a more accurate model to be build, trained and test so that more conclusive results trends and patterns can be observed.

Thirdly, this CART model is not that useful as it will require some form testing like Thalassemia if the patient is not known to have it. It also suggested that fluoroscopy should be done which is not ideal, if fluoroscopy has to be done there’s will be no need for this prediction. Perhaps, more qualitative variables can be used for prediction instead so just by taking the patient medical history and using the prediction models, doctors will already have a certain percentage of accuracy that these patients have heart disease. Some example will be if smoking, family history of heart disease, weight, alcohol use. These are already known factors for causes of heart disease and when a prediction model is created with these variables it will be more practical for use in the hospitals.

That being said, perhaps a proper study on the causes of heart disease should be done first. In the current context, we can only find out correlation of the variables however, having correlation does not always equals to causality. By finding out the causes of heart disease first perhaps we are able to build a more practical and useful model.

**Part C: Advanced Concepts (30%)**

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.

**What is the missing information?**

Text

Description automatically generated

As mentioned above in the CART model section, when rpart is run, 10 fold cross validation is being performed automatically. This means that the data is being trained on 9 of the 10 subsets available and the tested on the remaining one unseen test dataset. This process is continually repeated until 10 train and test set errors are being produced.

When we are trying to predict a continuous, numeric outcome variable the cross validation error or rather the xerror labelled in the cptable is an estimate of the error that will be obtained if the model is being applied to unseen data. The xerror here represents the expected Mean Squared Error, MSE that occurs when we are using the model and unseen data. This mean square error can be calculated because for numeric output as we have the integer values to calculate MSE or RSME. Note that MSE here refers to average squared difference between the estimated values and the actual value.

However, in categorical variables, we cannot use MSE or RMSE or rather it is not ideal because categorical variables are not “integer values” (usually they only have a selected range i.e 1, 2, 3, 4) thus we cannot quantify the differences numerically. This is especially true for multinomial categorical variables, we cannot simply use the MSE of the model as the differences between each level are not equal. To simplify things, let me give you an example.

For instance, the response variable, y we are trying to predict here is a multinomial categorical variable with categories 1-5 stars Thus using the model, we have come up with some predicted y values, y^.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Y^ (Predicted Y) | Actual Y values | Difference | Squared Errors |
| 1\* | 5\* | 4\* | 16\* |
| 5\* | 1\* | 4\* | 16\* |

This cause the model to overpredict and underpredict at the same time reducing accuracy. Thus, when Y is categorical variable, this is a classification problem instead of regression problem. In categorical variables, we have finite and countable class labels, which does not corresponds to numbers like in continuous Y. Therefore we cannot use MSE/RSME because it will be it will be difficult to find and quantify difference between label ‘a’ and label ‘b’. We only know that the expected and observed values are different as stated in the example. Instead other measures such as accuracy, geometric mean , precision, recall and ROC etc should be used. If we choose to use performance matrix like accuracy, TPR, TNR, FPR, and FNR this cannot tell us about the degree of error.

In summary, if the outcome variable is categorical, the missing important information is the MSE

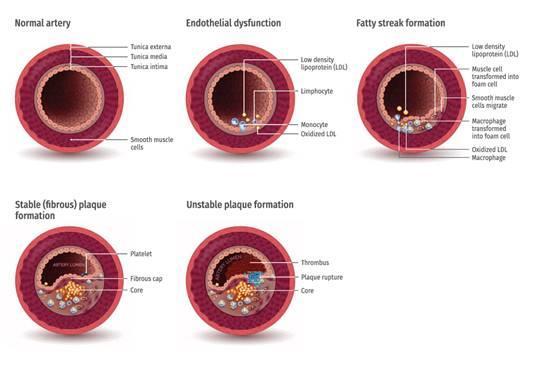
However, here we are measure AHD, which is a binary categorical variables it will be not that bad to use MSE because the differences in out prediction will always be the same i.e. 0 or 1. There will not be overprediction or underprediction that results in the same MSE as mentioned in the example above. Though, this doesn’t mean that MSE is a good measure for this dataset because for MSE to work there is an underlying assumption that the dataset distribution is a normal distribution (Khan, 2019)

Moreover, AHD, the narrowing of blood vessels is often cause by atherosclerosis (www.nhlbi.nih.gov, n.d.) and the disease progression is more often than not, not a binary problem.

**Propose a way to obtain this information**

Since heart disease/AHD comes in many stages, we can changing the response variable AHD to a continuous variable such that it measures the extents of the blood vessel narrowing. Note that multi-nominal categorical variables should also not be used due to the reasons mentioned above. Therefore, by changing the variable to a numeric, continuous variable we will be able to predict the exact percentage of the blood vessel narrowing allowing MSE to be obtained. This will also allow us to quantify and describe how serious a patient condition is and is more definitive measure for doctors.

Thus, by changing AHD to a continuous variable, the model can predict the exact percentages of the narrowing of the blood vessel.



(Oh, n.d.)

Since there are the 5 main stages of atherosclerosis, based on the exact percentage derive from the model doctors can then categorize them into the stage of the disease the patient is in.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Normal | Endothelial Dysfunction | Fatty streak Formation | Stable plaque formation | Unstable plaque formation |
| 0-20% | 20-40% | 40-60% | 60-80% | 80-100% |

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.***

Let’s first understand the approaches that was being used in the paper. The proposed approach were:

* 1. Ranking the attribute according to the correlation measure
  2. Performing classification using 3 different classifiers and comparison of accuracy across different classifier models
  3. Recommendation of relevant features for the chosen classifiers based on accuracy.

The research paper focused on choosing relevant attributes or independent variables among the initial 13 attributes namely age, sex, cp, tresbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca and cal and using these to predict “num” the dependent variables.

**Step 1:**

**Table

Description automatically generated**

The research paper determined the correlation of all independent variables against dependent variables, num. However, it is unclear which correlation scale was being used or how the correlation was being determined and thus the impact on the correlation calculations cannot be evaluated. Base on the correlation measure that we have learnt till now, to my knowledge Pearson Correlation can only be used to measure the correlation of two variables that are both in continuous scale. Since num is a categorical variable, Pearson correlation might not apply here. Correlation here is measure between categorical and categorical variable and between categorical and continuous variable. Between categorical and categorical variables, perhaps a chi square test was used and between categorical and continuous perhaps annova was used. There are numerous ways of measure correlation here, but it was unclear how it’s being achieved here.

It has been also found that the common method of evaluating correlation between continuous and categorical variable more often than not lead to the discretization of the continuous variables that can lead to information loss and after the models predicted later on. (Chrobak, 2015)

Also, we can’t be sure if correlation is the best metric to choose the variables to use in predicting heart disease. This is because correlation does not equal to causation. Although, there may be a high correlation value, but this might not actually cause heart disease in real life. There needs be more consideration in place when choosing the variables to be placed into the models.

**Step 2:**

In this step, variables with the highest correlation values are introduced one by one into each one of the classifier NB, NLP and SMO in order from the highest to the lowest. Classifier accuracy for each of the model was measured at each step a new variable was added in and compared against each other. This ordered selection of the variables fails to consider that other possible combinations of variables that might give higher prediction accuracy of considers. It does not consider the full model which included all the variables.

This method of selection is similar to a forward stepwise selection (Choueiry, n.d.) often used in regression analysis where they begin the model with no variables, the start by adding the most significant variables into the models. In this case, significance is determined by their correlation values. The process only stops when all 13 variables have been added into the model and tested for accuracy. Thus, by using forward stepwise selection, the best combination of variables might not be achieved.

**Step 3:**

In determining the best models, only one key metric was being looked at which was the classifier accuracy. The formula for model accuracy is (TP + TN)/ (TP+TN+FP +FN), thus the high accuracy may be caused by high TN values. High TN values is not ideal nor important in predicting heart disease as the key metrics we should be looking at should be how well we can predict heart disease and not how well we can predict heart disease not occurring.

Thus, in determining the best model, we need to look beyond overall model accuracy only instead things like precision (TP / (TP + FP)) and recall (TP / (TP + FN)) (developers.google.com, 2020) should be taken into consideration, such can be visualise using things like confusion matrix, ROC and AUC. We also need to ensure the dataset is balanced as an imbalanced dataset can negatively skew accuracy measures.

To re-emphasis, in predicting heart disease we need to look at achieving a high TPR to ensure patient with “nums” are predicted correctly and a low false negative rate so that we will not let patients with heart disease assume they do not have it. We also need to consider if the variables used in prediction are practical for real life use, for example, if the prediction calls for fluoroscopy there is essentially no need for the prediction model because a fluoroscopy can accurately tell us if the patient has heart disease or not. More thought needs to be put into choosing the variables such as using the common causes of heart disease like smoking, family history etc might make a better model that has more practical use. Note that causes of heart diseases should also be used here or be an addition consideration point because correlation here doesn’t necessarily causes heart disease.

# Bibliography

Sharma, N. (May, 2018). *https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba*. Retrieved from towardsdatascience: https://towardsdatascience.com

Grace, K. (2018). *https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/*. Retrieved from theanalysisfactor: www.theanalysisfactor.com

Sheps, S. G. (26 Jan , 2019). *https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/hypertensive-crisis/faq-20058491*. Retrieved from mayoclinic: www.mayoclinic.org

https://www.singhealth.com.sg/patient-care/conditions-treatments/cholesterol-management/other-information. (n.d.). *singhealth*. Retrieved from singhealth: https://www.singhealth.com.sg

SingHealth. (n.d.). Retrieved from https://www.singhealth.com.sg/patient-care/conditions-treatments/cholesterol-management/other-information

Bekerman, J. (24 Aug, 2020). *wedmd*. Retrieved from wedmd: https://www.webmd.com/cholesterol-management/understanding-your-cholesterol-report

Seals, D. D. (15 May , 2008 ). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2494835/

Christou, D. D. (2008). Retrieved from https://www.researchgate.net/publication/309210947\_Heart\_Disease\_prediction\_using\_Machine\_learning\_and\_Data\_Mining\_Technique: researchgate

Donges, N. (23 04, 2018). *machinelearning-blog.* Retrieved from https://machinelearning-blog.com: https://machinelearning-blog.com/2018/04/23/logistic-regression-101/#:~:text=A%20disadvantage%20of%20it%20is,it's%20decision%20surface%20is%20linear.&text=Since%20its%20outcome%20is%20discrete,for%20its%20vulnerability%20to%20overfitting.

*my.clevelandclinic.org*. (n.d.). Retrieved from https://my.clevelandclinic.org/: https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease

Publishing, H. H. (June, 2020). *health.harvard.edu*. Retrieved from health.harvard.edu: https://www.health.harvard.edu/heart-health/a-closer-look-at-your-coronary-arteries

Bock, T. (2020). *www.displayr.com*. Retrieved from www.displayr.com: https://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression/

Choueiry, G. (n.d.). *quantifyinghealth.com*. Retrieved from quantifyinghealth.com: https://quantifyinghealth.com/stepwise-selection/

*developers.google.com*. (10 2, 2020). Retrieved from developers.google.com: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Khan, R. (15 November, 2019). *Why Using Mean Squared Error(MSE) Cost Function for Binary Classification is a Bad Idea?* Retrieved from Towards Data Science: https://towardsdatascience.com/why-using-mean-squared-error-mse-cost-function-for-binary-classification-is-a-bad-idea-933089e90df7#:~:text=There%20are%20two%20reasons%20why,we%20assume%20a%20Gaussian%20prior.

Khan, R. (15 Nov, 2019). Retrieved from https://towardsdatascience.com/why-using-mean-squared-error-mse-cost-function-for-binary-classification-is-a-bad-idea-933089e90df7

*www.nhlbi.nih.gov*. (n.d.). Retrieved from www.nhlbi.nih.gov: https://www.nhlbi.nih.gov/health-topics/atherosclerosis#:~:text=Atherosclerosis%20is%20a%20disease%20in,other%20parts%20of%20your%20body.

Oh, S. L. (n.d.). *www.researchgate.net*. Retrieved from www.researchgate.net: https://www.researchgate.net/figure/The-pathology-of-coronary-artery-disease-CAD\_fig1\_317336647

Neshat, M. (March, 2010). *www.researchgate.net*. Retrieved from www.researchgate.net: https://www.researchgate.net/figure/CLASSIFICATION-OF-MAXIMUM-HEART-RATE\_tbl1\_44260568

Chrobak, M. (17 July , 2015). *www.journals.elsevier.com*. Retrieved from www.journals.elsevier.com: https://www.journals.elsevier.com/information-processing-letters