HEART DISEASE PREDICTION

Introduction

Heart disease, also known as cardiovascular disease, is a disorder characterized by clogged blood vessels that can result in a heart attack.[1] According to the World Health Organization, heart disease has been the greatest killer among diseases worldwide over the past 20 years.[2]

Every year, heart diseases have caused an estimated 17.9 million mortalities globally. It now accounts for 16% of total mortalities from all causes. Not just that, it became the main cause of death among Malaysians in 2019, with 173,746 deaths reported by the New Straits Times.[3]

Risk factors of heart disease fall into three main categories[4] :

1. Unchangeable risk factors - age, gender, heredity;
2. Modifiable risk factors - smoking, blood cholesterol, blood pressure, obesity, diabetes; and
3. Contributing risk factors - stress, alcohol, diet and nutrition.

Data Dictionary

The dataset used in this report is the **Statlog (Heart)** dataset from UCI Machine Learning Repository.[5] It consists of **270** cases. The following table describes the **14** variables and measurements for each of the variables from the dataset.

No.	Names	Description	Measurement
1	age	Age	in years
2	sex	Sex	0 = female, 1 = male
3	cp	Chest pain type	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic
4	restbp	Resting blood pressure	in mm Hg
5	chol	Serum cholesterol	in mg/dl
6	fbs	Fasting blood sugar > 120 mg/dl	0 = false, 1 = true
7	ecg	Resting electrocardiographic results	0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy
8	hrate	Max heart rate achieved	integer
9	exang	Exercise induced angina	0 = no, 1 = yes
10	stdep	ST depression induced by exercise relative to rest	integer or float
11	slope	The slope of the peak exercise ST segment	1 = upsloping, 2 = flat, 3 = downsloping
12	fluor	Number of major vessels colored by fluoroscopy	0 - 3
13	thal	Thallium stress test results	3 = normal, 6 = fixed, 7 = reversible
14	target	Diagnosis of heart disease	1 = absent, 2 = present

The variables include risk factors (age, sex, blood pressure, cholesterol, blood sugar), symptoms (chest pain, exercise-induced angina), test results (resting electrocardiography, max heart rate, ST depression, ST segment, fluoroscopy, thallium stress test) and finally the target variable (diagnosis of heart disease).

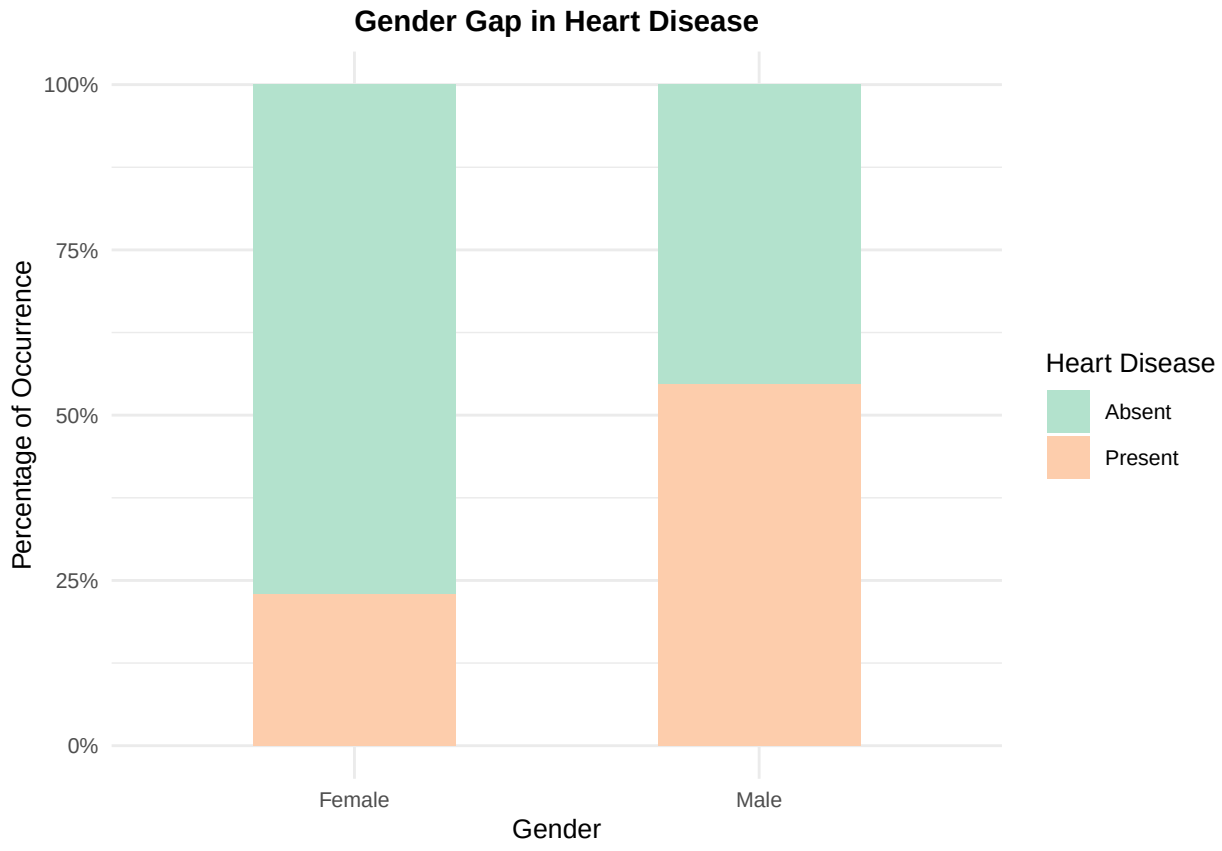
As heart disease is a major public health concern, it is essential to be knowledgeable about the symptoms and factors of heart disease. By analyzing the healthcare data, we can gain profound insights into the disease. Predictive modeling can also be used to predict the diagnosis, which may assist in the early detection and prevention of heart disease.

Exploratory Data Analysis

Exploratory Data Analysis is a practical approach for making sense out of data.[6] Throughout this section, different methods are applied to investigate and reveal the relationships between variables.[7]

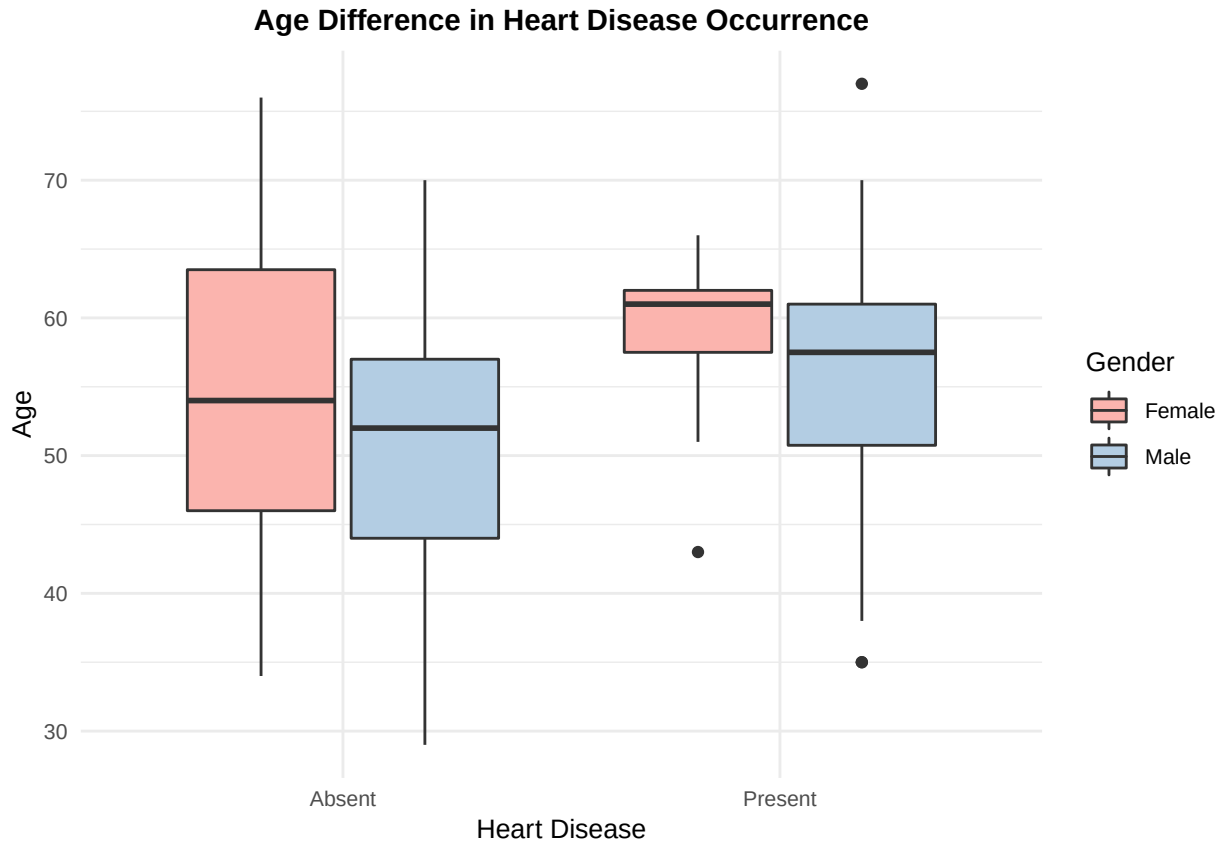
Question 1

Which gender has a higher risk of heart disease?



The bar chart above illustrates the percentage of heart disease occurrence in different genders. Among female respondents, less than a quarter of them are diagnosed with heart disease. Contrarily, heart disease is present in more than half of the male respondents. Hence, it can be said that male has a higher risk than female.

Due to men's higher absolute risk relative to women, cardiovascular disease has been seen as a male disease for a long time. There are a variety of reasons that can explain the gender difference in heart disease. For instance, psychosocial and behavioral factors such as excessive alcohol consumption and smoking, are in favor of women.[8]

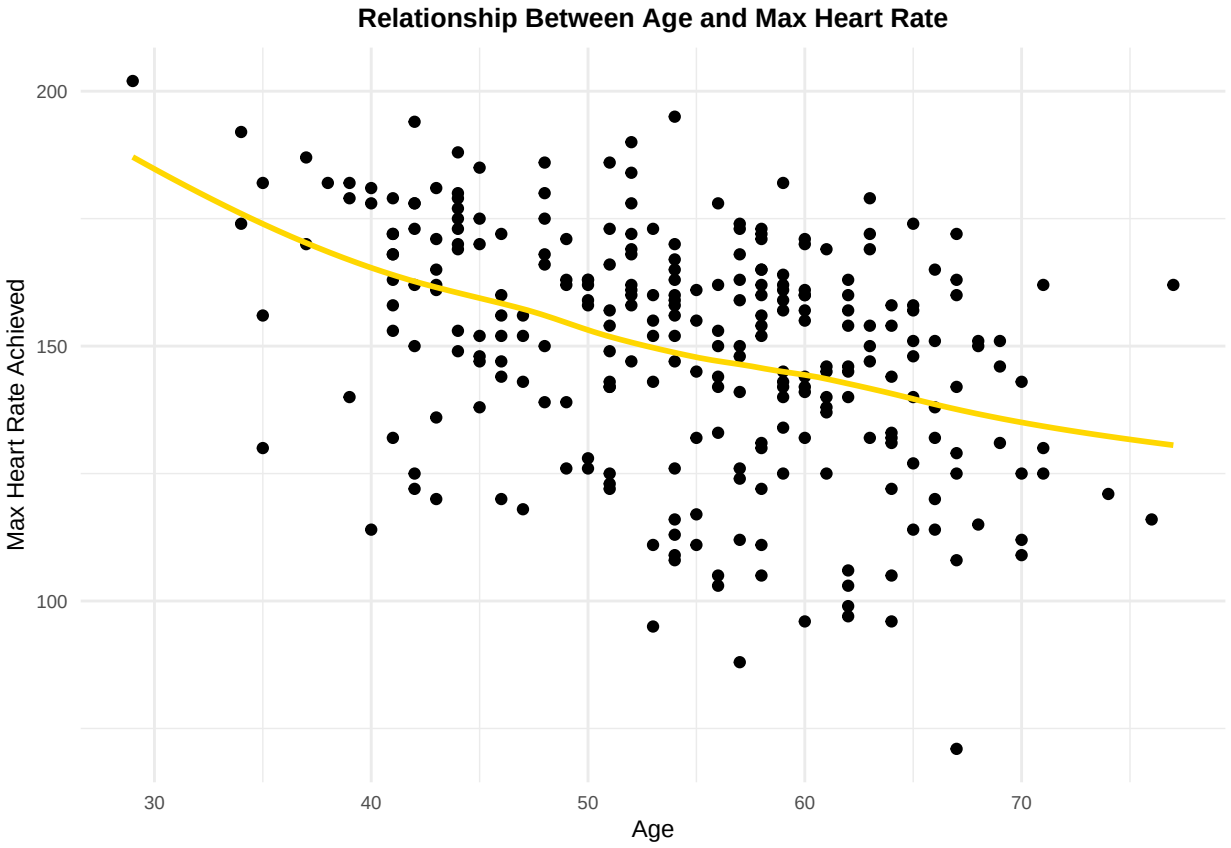


As shown in the boxplot, there is an age gap between genders in terms of heart disease diagnosis. The interquartile range for men with heart disease is much larger than their female counterparts. The boxplot for female with heart disease is also heavily skewed to the left in comparison with male's, indicating that female get heart disease later in life. A study in 2010 has revealed that cardiovascular disease develops 7 to 10 years later in women than in men. In spite of that, it remains the leading cause of mortality in women.[9]

Heart attack can be more fatal in women. Due to their smaller body size and blood vessels, blockages of the same size can be more serious. The narrower vessel channel also makes surgery and angioplasty procedures more complicated.[1] Hence, the risk should not be underestimated due to the false sense of security that females are protected against cardiovascular disease.[9]

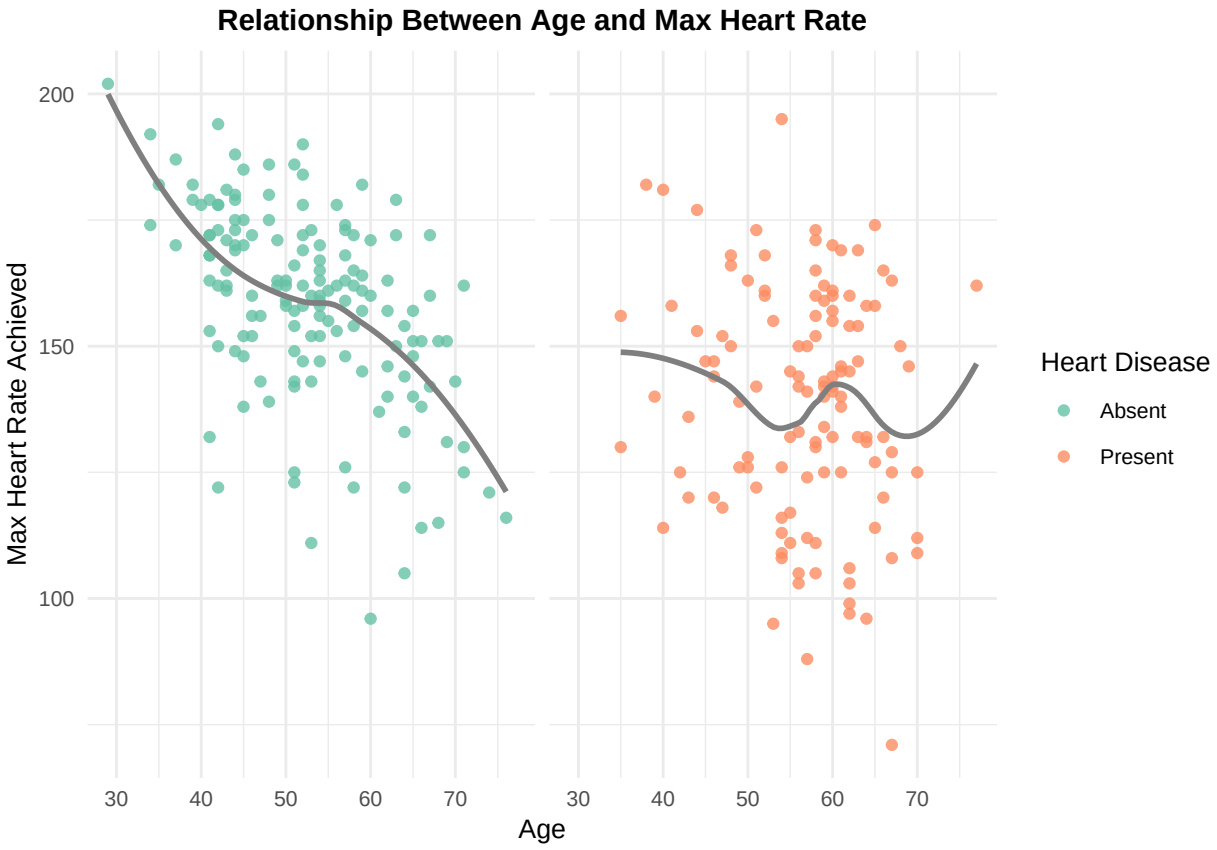
Question 2

What is the relationship between heart rate and age?



To identify the relationship, a scatterplot was obtained by plotting the age of respondents against their maximum heart rate achieved. From the plot above, it shows that age and maximum heart rate are inversely related, in which the older the age, the lower the maximum heart rate.

Many studies have shown that the maximum heart rate declines with age.[10] The major reason behind the age-related decrease is due to a drop in the intrinsic heart rate. This implies that the heart's own pacemaker (a cluster of cells responsible for initiating each heartbeat) has lost its functional efficiency.[11]



The graph prior to this shows that the maximum heart rate decreases steadily with age. However, after splitting the graph based on the diagnosis of heart disease, the previous statement no longer holds true for those who suffered from heart disease.

The line for those with heart disease has a more irregular pattern. It also shows that those who suffered from heart disease have a lower maximum heart rate at an earlier age, reflecting poor heart functionality. Hence, an abnormally low maximum heart rate may be used as a risk marker for younger people to detect heart disease.

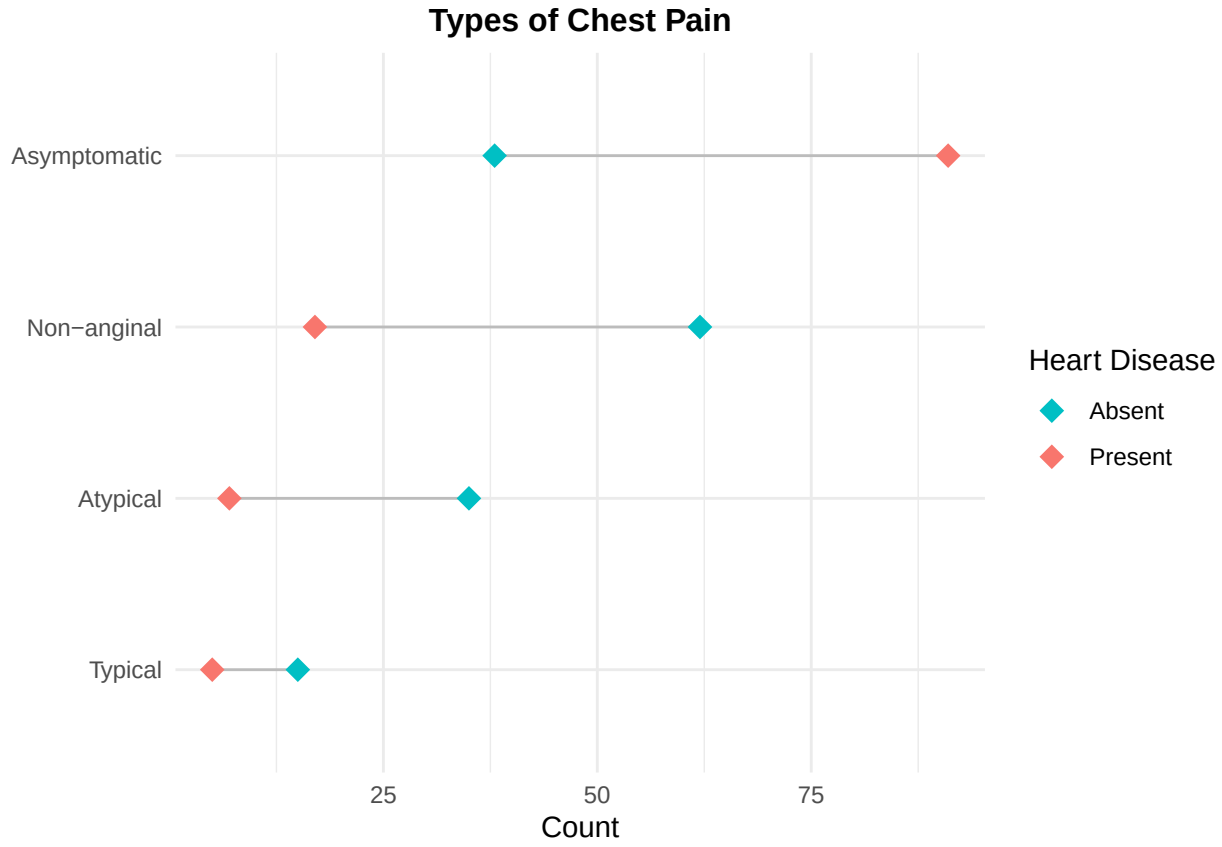
Question 3

Which type of chest pain is the most prevalent in individuals with heart disease?

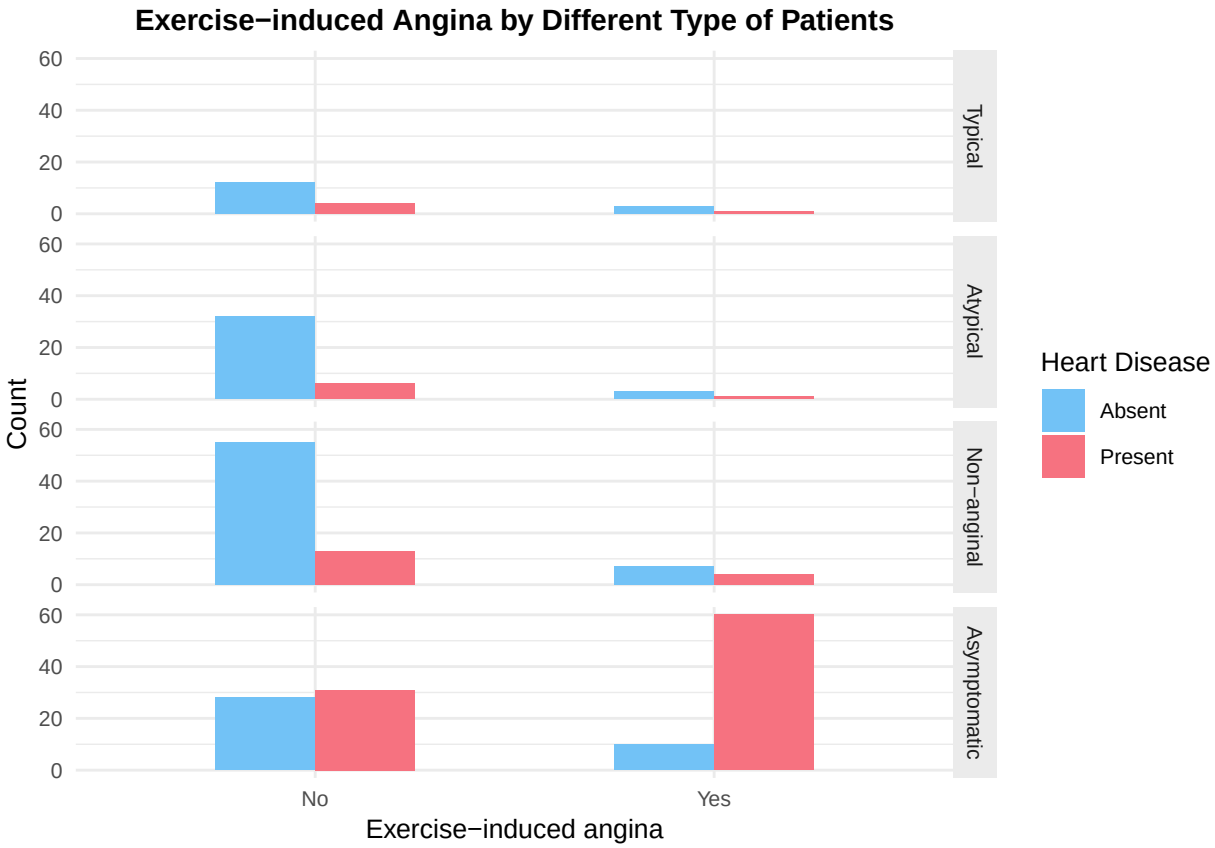
The type of chest pain contains 4 different levels:[12] [13]

1. Typical angina - caused by physical exertion or emotional stress
2. Atypical angina - does not cause any typical chest pain but will radiate to other parts of the body
3. Non-anginal pain - resembles heart discomfort in people who do not have heart disease
4. Asymptomatic - showing no symptoms or minimal symptoms

Angina is described as a kind of discomfort caused when the heart doesn't get enough oxygen-rich blood.



From the plot, it is evident that the 'Asymptomatic' type is the most prevalent among the respondents with heart disease. This means that most of them actually do not show any symptoms of heart disease. A person who appears to be healthy may be experiencing painless damage to the heart. The first sign of silent heart disease can be fatal: a heart attack or even death.[1]



From the bar chart above, exercise-induced angina is found to have a higher occurrence in respondents suffering from heart disease, specifically with those who are otherwise asymptomatic. This means that even if they do not suffer from any chest pain during day-to-day activities, they might still experience some discomfort during physical exercise.

In fact, it was proven that exercise-induced angina is a common complaint among cardiac patients, which can be triggered further when exposed to a cold environment.[14] Thus, this symptom can be served as an indicator of potential heart disease.

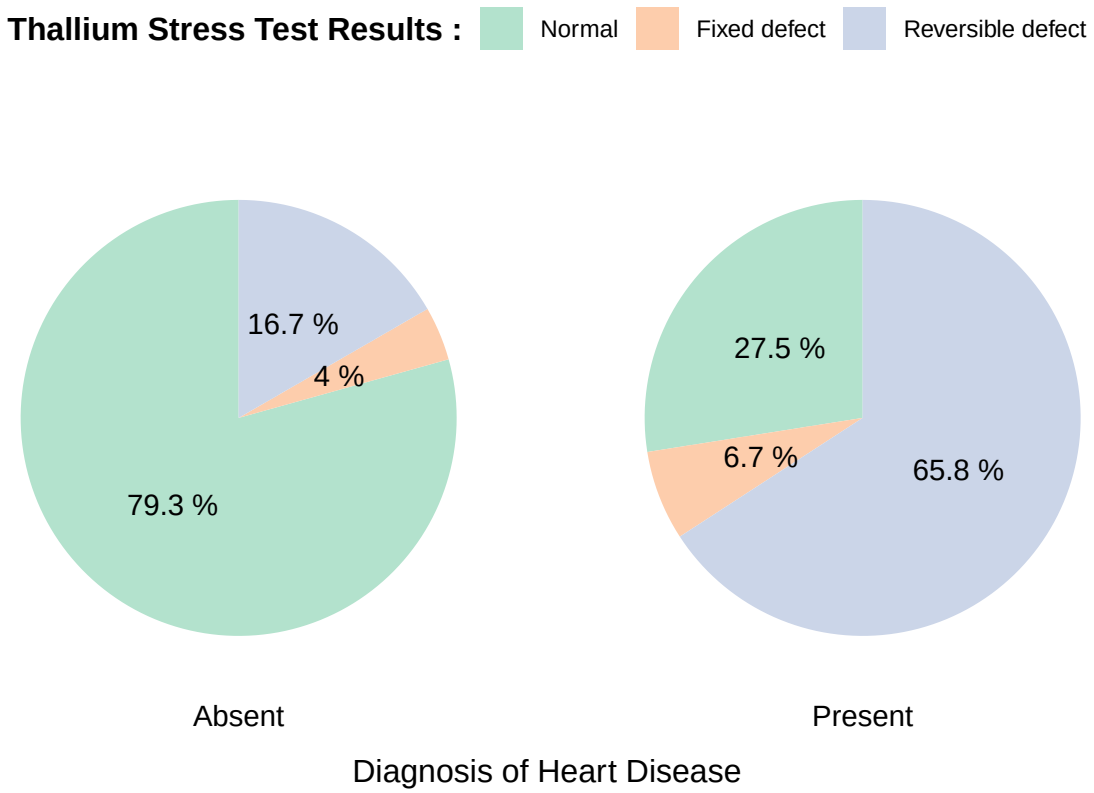
Question 4

Can thallium stress tests detect heart disease?

A thallium stress test is a nuclear imaging test that evaluates blood flows in the heart. The heart is photographed while doing exercise and at rest.[15] It can provide diagnostic information about patients with known or suspected cardiovascular disease.[16]

	Normal	Fixed	Reversible	Total
Absent	119	6	25	150
Present	33	8	79	120
Total	152	14	104	270

Based on the dataset, the test results are categorized as normal, fixed defect and reversible defect. Fixed defects refer to abnormalities present on scan both during cardiac stress and at rest. Reversible defects are those that are present on the post-stress scan but show improvement in the resting scan.[16]



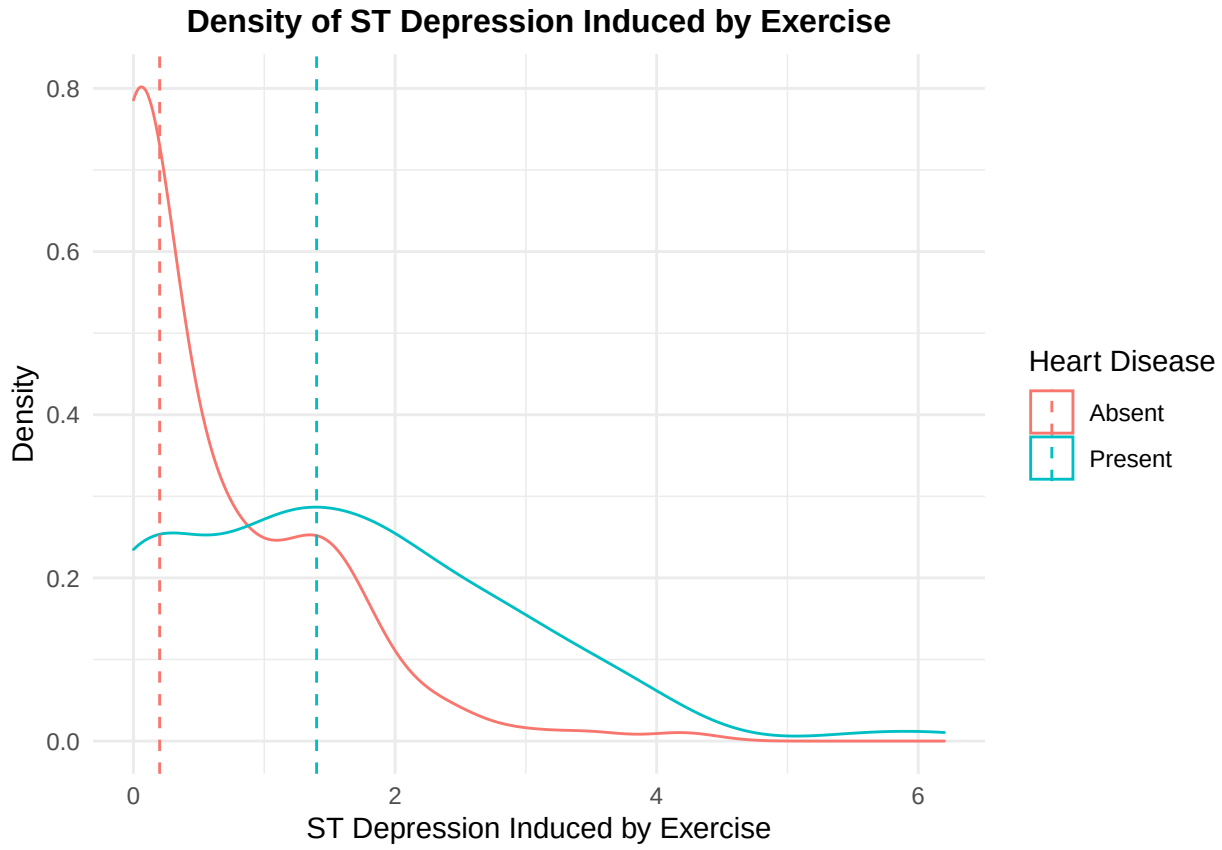
Based on the pie chart, almost 80% of respondents without heart disease have normal test results. On the other hand, 66% of respondents with heart disease have reversible defects. Thus, it demonstrates that thallium stress test can be quite effective in detecting heart disease and the results may contribute to the prediction of diagnosis.

Nonetheless, normal results still account for more than a quarter of the respondents with heart disease. Hence, the risk should not be underestimated even with normal stress test results.

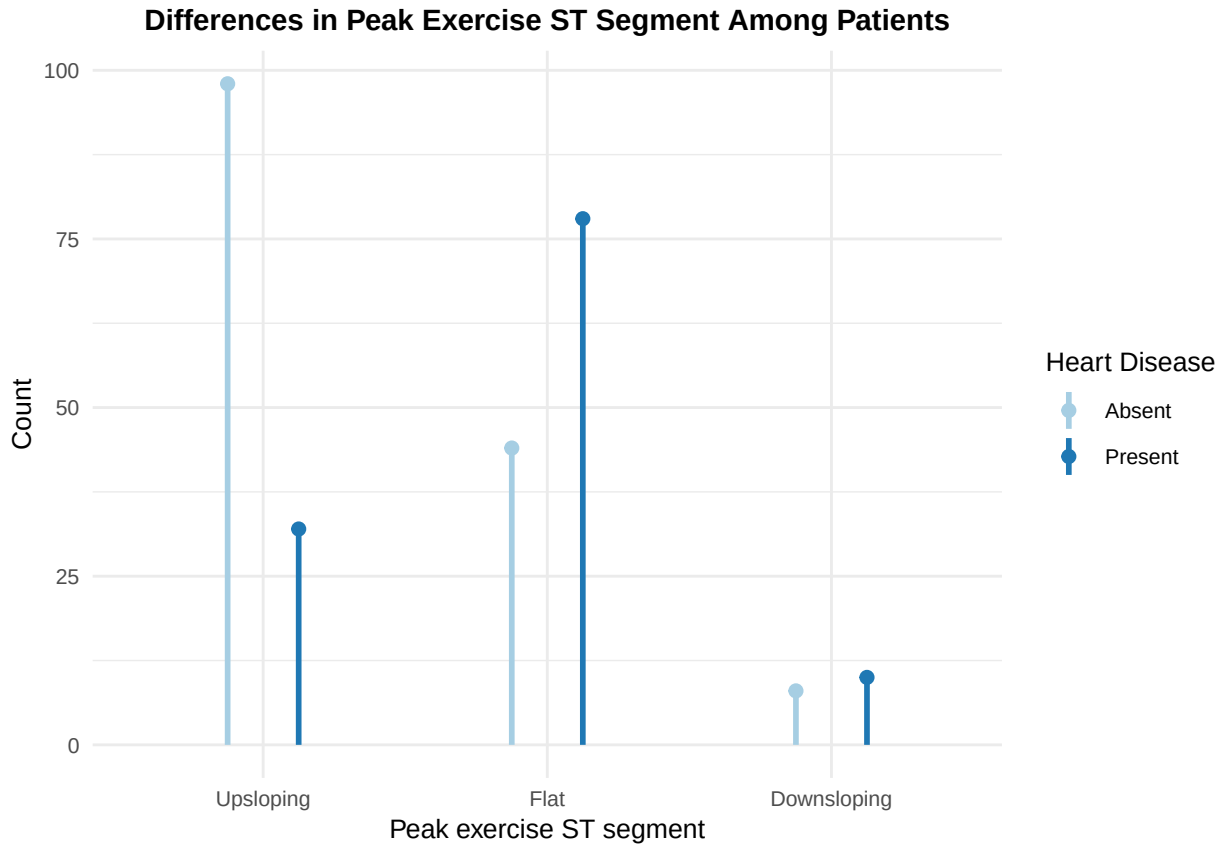
Question 5

How does the ST segment differ between individuals with and without heart disease?

Electrocardiogram (ECG) measures electrical activities of the heart. The ST segment reflects the period of zero potential between ventricular depolarization and repolarization. The ST segment can be displaced upwards (elevation) or downwards (depression).[17]



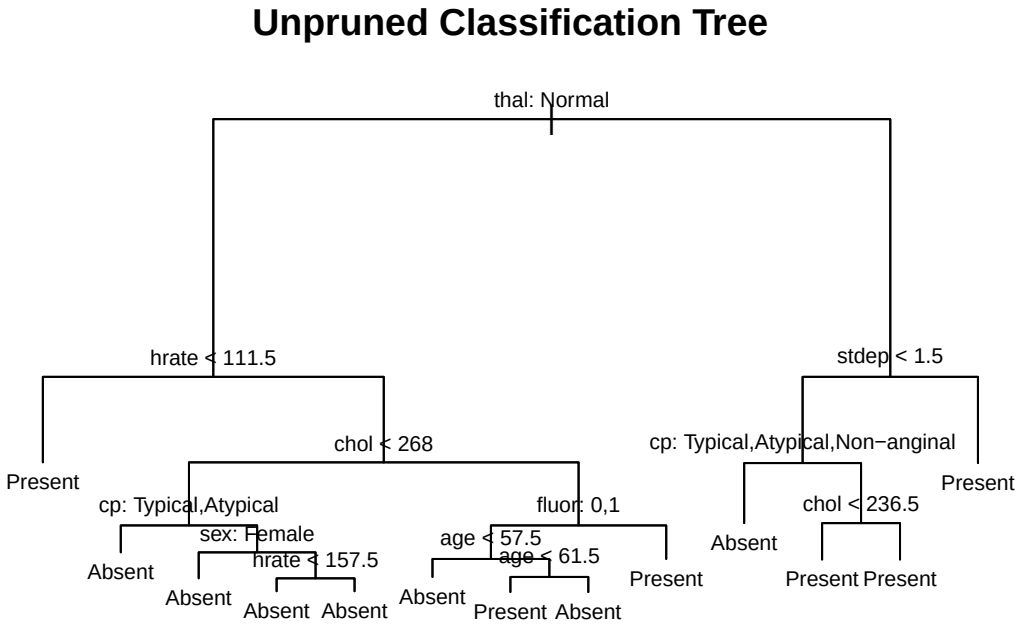
ST depression refers to an ECG result in which the trace in the ST segment is unusually low.[18] The acceptable level of ST segment depression is less than 0.5mm, while reading that is more than 0.5mm is considered pathological.[19] The graph above supports that higher ST depression is related to the presence of heart disease. It can be seen that the dashed line that represents the median is higher for patients with heart disease than non-diseased patients.



Besides the reading of ST depression, the characteristic of ST segment is also important. As shown above, respondents who are not diagnosed with heart disease have the highest rate of showcasing an upsloping ST segment during exercise, followed by a flat one. On the flip side, those with heart disease are more likely to exhibit a flat ST segment. Previous studies have proved that the normal ST segment during exercise has an upward curve. Flat or downsloping ST segments might indicate possible abnormalities.[20]

Predictive Modeling

A decision tree is a predictive model which can be used to represent both classifiers and regression models. In this study, classification trees are used to classify the diagnosis of heart disease into a predefined set of classes (absent/present) based on their attribute values (e.g., age, gender, chest pain type, etc).[21]



By fitting the classification tree using all of the predictors, the tree turned out to be quite complicated (i.e., has too many nodes), reducing the usefulness of its straightforward graphical representation. Branches that are not contributing to the generalization accuracy can be trimmed to simplify the interpretation of the hierarchical structure and improve its comprehensibility. This process is called pruning.[21]

Summary of Unpruned Classification Tree

```
##
## Classification tree:
## tree(formula = target ~ ., data = heart, subset = training)
## Variables actually used in tree construction:
## [1] "thal"  "hrate" "chol"  "cp"    "sex"   "fluor" "age"   "stdep"
## Number of terminal nodes:  13
## Residual mean deviance:  0.5484 = 66.9 / 122
## Misclassification error rate: 0.1185 = 16 / 135
```

The summary above shows that the tree has 13 terminal nodes and a misclassification error rate of 11.85%. 8 out of 13 predictors have been used in the tree construction.

Accuracy of Unpruned Classification Tree

```
##          target.test
## tree.pred Absent Present
##   Absent      67      17
##   Present     17      34
```

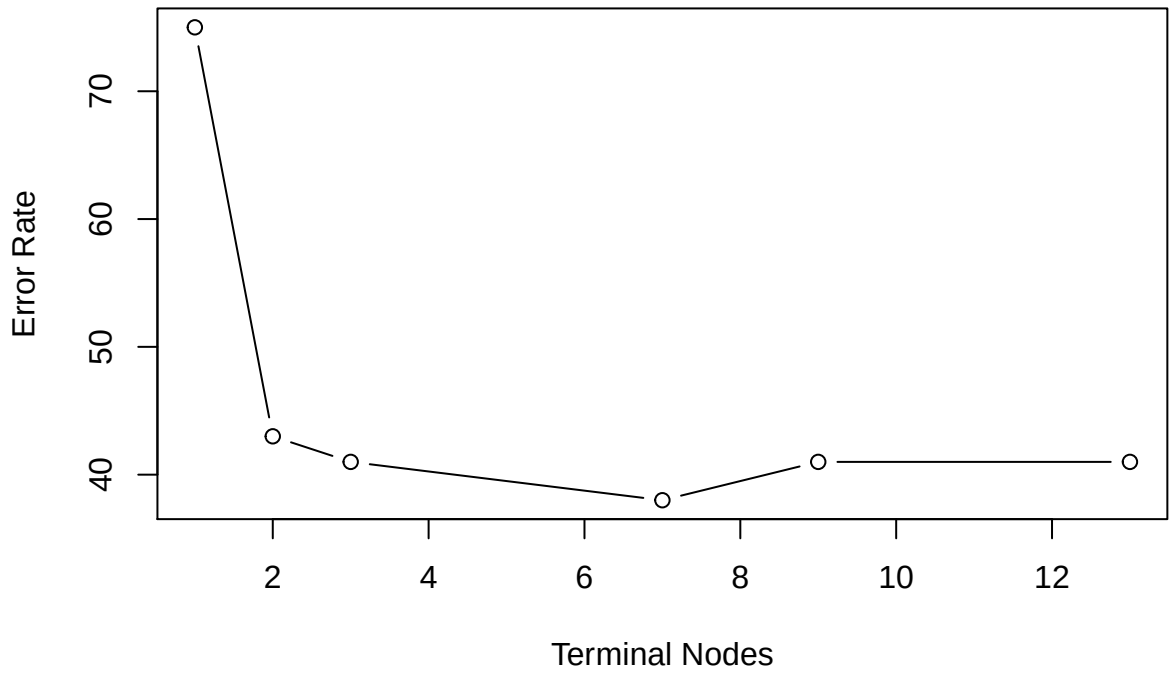
```
## [1] 0.7481481
```

By evaluating the tree model using the test set, the unpruned tree has an accuracy rate of 74.81%.

Pruning

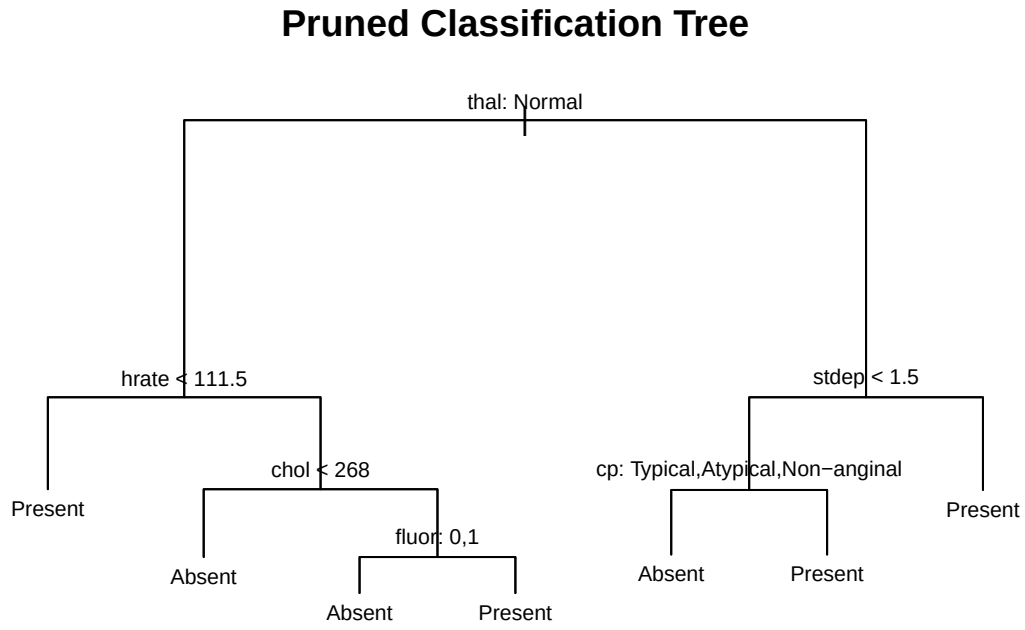
Cross-validation and cost complexity pruning is utilized to determine the optimal level of tree complexity. The misclassification error rate is used as the cost function to guide the pruning process.[21]

```
## $size
## [1] 13  9  7  3  2  1
##
## $dev
## [1] 41 41 38 41 43 75
##
## $k
## [1] -Inf    0    1    2    5   35
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```



The cross-validation of terminal nodes amount (size) and the corresponding error rate (dev) shows that a tree with 7 terminal nodes will result in the lowest misclassification error rate.

The parameter is set to obtain the seven-node tree.



The above shows that thal (thallium stress test) is the root node, which is the best attribute to predict outcomes. It then splits into the decision nodes according to the rules. If it obeys the rule, it will go down to the left child node and to the right if it doesn't (e.g. if the test result is normal, it will go to the node with heart rate < 111.5; if the test result is not normal, it will go to the node with ST depression < 1.5). This will carry on until it reaches one of the seven terminal nodes.

Summary of Pruned Classification Tree

```
##
## Classification tree:
## snip.tree(tree = tree.heart, nodes = c(10L, 13L, 22L))
## Variables actually used in tree construction:
## [1] "thal"  "hrate" "chol"  "fluor" "stdep" "cp"
## Number of terminal nodes:  7
## Residual mean deviance:  0.6988 = 89.45 / 128
## Misclassification error rate: 0.1333 = 18 / 135
```

The summary shows that the tree now has 7 terminal nodes and a misclassification error rate of 13.33%. Only 6 of the predictors have been used to construct the pruned tree.

Details of Pruned Classification Tree

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 135 187.100 Present ( 0.48889 0.51111 )
##    2) thal: Normal 75  86.990 Absent ( 0.73333 0.26667 )
##      4) hrate < 111.5 5   0.000 Present ( 0.00000 1.00000 ) *
##      5) hrate > 111.5 70  72.740 Absent ( 0.78571 0.21429 )
##       10) chol < 268 44  26.810 Absent ( 0.90909 0.09091 ) *
##       11) chol > 268 26  35.430 Absent ( 0.57692 0.42308 )
##         22) fluor: 0,1 20  24.430 Absent ( 0.70000 0.30000 ) *
##         23) fluor: 2,3 6   5.407 Present ( 0.16667 0.83333 ) *
##    3) thal: Fixed defect,Reversible defect 60  57.170 Present ( 0.18333 0.81667 )
##      6) stdep < 1.5 34  42.810 Present ( 0.32353 0.67647 )
##       12) cp: Typical,Atypical,Non-anginal 12  15.280 Absent ( 0.66667 0.33333 ) *
##       13) cp: Asymptomatic 22  17.530 Present ( 0.13636 0.86364 ) *
##      7) stdep > 1.5 26   0.000 Present ( 0.00000 1.00000 ) *
```

Accuracy of Pruned Classification Tree

```
##           target.test
## prune.pred Absent Present
##    Absent      69      17
##    Present     15      34
```

```
## [1] 0.762963
```

The pruned tree has a prediction accuracy rate of 76.29%, which is a slight improvement from the unpruned tree.

Comparison of Pruned and Unpruned Classification Trees

	Unpruned	Pruned
No. of Terminal Nodes	13	7
Predictors Used	8	6
Accuracy Rate	74.81%	76.29%
Misclassification Rate	11.85%	13.33%

Comparing both trees, the pruned tree has only 7 terminal nodes, which makes the classification process more simplified than the unpruned tree with 13 terminal nodes. The pruned tree only used 6 out of the 13 predictors. In contrast, 8 predictors were fit into the unpruned tree.

The diagnostic accuracy has increased around 1.5% after pruning. However, the misclassification rate had also gone up at the same rate. Pruning a tree will increase the misclassification error rate of training data, but should decrease the error rate on independent test data. This is because in high uncertainty domains, all but the first two or three levels of the tree were being removed.[22]

Users should choose the appropriate tree based on their criteria.

Conclusion

In conclusion, men have higher odds of developing heart disease than women. Nevertheless, women should not take this lightly as heart disease remains the biggest cause of death among women. We also found out that younger people with heart disease have abnormally low maximum heart rates, whereas the heart rate of those without heart disease decreases steadily with age.

Despite technological advancements and increased health consciousness, heart diseases still claim millions of lives every year. Excruciatingly, this catastrophic event can occur without any warning. The analysis shows that most patients with heart disease do not suffer from chest pain. However, exercise-induced angina is found to be more common in cardiac patients. A thallium stress test seems to be effective in detecting heart disease. In spite of that, a normal test result does not equate to risk-free as it still accounts for more than a quarter of those suffering from heart disease.

There are also other means of measure that can be employed to detect heart disease. For example, the ST segment in electrocardiogram (ECG) can be used to monitor the risk. A high number of ST depression or a flat ST segment during exercise might indicate the presence of heart disease. The susceptibility to heart disease can also be estimated with the help of predictive modeling. In this study, decision tree was used to predict the diagnosis of heart disease.

Many factors play a part in our heart health, but not all of them can be altered. Preventive measures and early diagnosis need to be emphasized, especially towards high-risk subjects. Genetics can't be altered, but diets and everyday habits sure do. Regardless of one's genetic profile, lifestyle factors may reduce the risk of heart disease.

References

1. Cohn DP, Cohn DJ. *Fighting the Silent Killer: How Men and Women Can Prevent and Cope with Heart Disease Today*. 0th ed. A K Peters/CRC Press; 1993. doi:10.1201/9781439864753

2. WHO reveals leading causes of death and disability worldwide: 2000-2019. December 2020. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.

3. New Straits Times. Heart disease top killer of Malaysians in 2019. *NST Online*. November 2020. https://www.nst.com.my/news/nation/2020/11/644515/heart-disease-top-killer-malaysians-2019. Accessed May 30, 2021.

4. American Heart Association. Understand your risks to prevent a heart attack. June 2016. https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack. Accessed May 30, 2021.

5. Dua D, Graff C. UCI machine learning repository. 2017. http://archive.ics.uci.edu/ml.

6. Myatt GJ. *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. Hoboken, N.J: Wiley-Interscience; 2007.

7. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org.

8. Weidner G. Why do men get more heart disease than women? An international perspective. *Journal of American College Health*. 2000;48(6):291-294. doi:10.1080/07448480009596270

9. Maas AHEM, Appelman YEA. Gender differences in coronary heart disease. *Netherlands Heart Journal*. 2010;18(12):598-603. doi:10.1007/s12471-010-0841-y

10. Kostis JB, Moreyra AE, Amendo MT, Di Pietro J, Cosgrove N, Kuo PT. The effect of age on heart rate in subjects free of heart disease. Studies by ambulatory electrocardiography and maximal exercise stress test. *Circulation*. 1982;65(1):141-145. doi:10.1161/01.CIR.65.1.141

11. Dworkin G. Changes in max heart rate with aging. *Vein Specialists of Tampa*. February 2017. https://www.tampaveinspecialists.com/changes-in-max-heart-rate-with-aging/. Accessed May 30, 2021.

12. American Heart Association. Angina(Chest pain). *Angina (Chest Pain)*. July 2015. https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain. Accessed May 30, 2021.

13. Wong C-K. Recognising ”painless” heart attacks. *Heart*. 2002;87(1):3-5. doi:10.1136/heart.87.1.3

14. Brown CF, Oldridge NB. Exercise-induced angina in the cold: *Medicine & Science in Sports & Exercise*. 1985;17(5):607???612. doi:10.1249/00005768-198510000-00015

15. Nelson J. Thallium stress test: Purpose, procedure, and risks. *Healthline*. June 2012. https://www.healthline.com/health/thallium-stress-test.

16. Tighe DA, Gentile BA, Chung EK, eds. *Pocket Guide to Stress Testing*. Second edition. Hoboken, NJ: Wiley; 2020.

17. Ashley EA, Niebauer J. *Cardiology Explained.* London: Remedica Medical Education; Publishing; 2004.

18. Okin PM, Devereux RB, Kors JA, et al. Computerized st depression analysis improves prediction of all-cause and cardiovascular mortality: The strong heart study. *Annals of Noninvasive Electrocardiology*. 2001;6(2):107-116. doi:10.1111/j.1542-474X.2001.tb00094.x

19. Rawshani A. The ST segment: Physiology, normal appearance, ST depression & ST elevation. *ECG & Echo Learning*. August 2018. https://ecgwaves.com/st-segment-normal-abnormal-depression-elevation-causes/. Accessed May 30, 2021.

20. Hill J. ABC of clinical electrocardiography: Exercise tolerance testing. *BMJ*. 2002;324(7345):1084-1087. doi:10.1136/bmj.324.7345.1084

21. Rokach L, Maimon O. *Data Mining with Decision Trees: Theory and Applications*. Second edition. Hackensack, New Jersey: World Scientific; 2015.

22. Mingers J. An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning*. 1989;4(2):227-243. doi:10.1023/A:1022604100933