Building Decision Tree Predictive Models for Heart Disease

Student Name: Wangjun SHEN

Student ID: 110248810

School: The University of South Australia

Content Table

Introduction

The objective of this assignment is to utilize decision tree models to predict heart disease, using visually engaging displays such as tables and graphics to enhance presentation. Machine learning algorithms have gained popularity in detecting heart disease due to their potential for early diagnosis and improved accuracy. This assignment will examine the dataset and extract relevant features for analysis, taking feedback into account to optimize predictive power. Through the use of descriptive statistics and visual aids such as value distributions, histograms, bar plots, and box plots, we will gain insights into the dataset and investigate correlations between important variables. Decision tree models will be constructed using R, with multiple models fitted and their performance compared using metrics such as accuracy, precision, recall, F-score, and AUC. The aim of this assignment is to develop a comprehensive analysis of heart disease detection using decision trees that addresses previous feedback on result presentation and adheres to the assignment requirements.

Related Work

Heart disease poses a significant public health concern due to its high prevalence and associated morbidity and mortality rates. According to the World Health Organization (WHO), cardiovascular diseases, including heart disease, are the leading cause of death globally (WHO, 2020). For instance, a study by Mozaffarian et al. (2015) estimated that approximately 17.9 million deaths occur each year due to cardiovascular diseases. This alarming statistic underscores the urgent need for effective methods of heart disease detection and management. The growing interest in utilizing machine learning algorithms for heart disease detection stems from their potential to enhance diagnostic accuracy, risk prediction, and personalized treatment strategies. By harnessing the power of machine learning, healthcare practitioners can benefit from more accurate and efficient tools for early detection and intervention, leading to improved patient outcomes and a reduction in the burden of heart disease on individuals and healthcare systems.

Machine learning algorithms have shown promise in detecting various diseases. Previous research has highlighted key themes and findings in this field. For instance, Géron (2019) discusses the use of machine learning for predicting cardiovascular disease risk, emphasizing the importance of feature engineering and model interpretability. In another study, Attia et al. (2019) utilized deep learning to analyze ECG data and accurately detect atrial fibrillation, a common heart rhythm disorder. Their work demonstrates the potential of machine learning in improving early detection and diagnosis. Furthermore, Dey et al. (2018) present a comprehensive review of machine learning techniques for diagnosing heart disease and predicting risk, encompassing various models and datasets. Their study provides valuable insights into the strengths and limitations of different approaches.

Data Exploration

The same feature set that was extracted in Assignment 1 will be used to predict heart disease. Based on feedback received, the background section was the main area of focus, and no specific issues were raised with the selected feature set. Therefore, the previously chosen feature set will be maintained as it includes relevant variables associated with heart disease. These features encompass demographic information such as age, sex, and chest pain type, as well as physiological indicators like blood pressure, cholesterol levels, and maximum heart rate.

Variable	Description	Data Type
id	A unique ID that identifies a participant in the study	Integer
age	Age in years	Integer
sex	Male and Female were recorded	Character
cp	Chest Pain type: typical angina; atypical angina; non-anginal pain; and asymptomatic	Character
thalach	Maximum heart rate achieved	Integer
exang	Exercise induced angina (True/False)	Logical
oldpeak	ST depression induced by exercise relative to rest	Numeric
major_vessels	Number of major vessels (0-3) colored by flourosopy	Integer
restwm	Rest wall motion abnormality: none; mild or moderate; moderate or severe; akinesis or dyskmem	Character
target	Heart disease diagnosed (disease/no disease)	Character

Table 1: Features Information for the Selected Features

When examining a dataset, it's crucial to verify for any absent values. This is because absent values can influence the precision and credibility of the analysis, resulting in prejudiced outcomes and less efficient models. Spotting and resolving absent values can enhance the quality of the analysis.

Feature	Missing Count
id	0
age	0
sex	0
cp	0
thalach	0
exang	0
oldpeak	0
major_vessels	0
restwm	7
target	0

Table 2: Count Result for Missing Value of Each Feature

The majority of features in the dataset are complete, except for the restwm feature, which has 7 missing values. We will remove these values directly. The dataset is sufficiently large, so their removal will have no impact.

To begin, an examination will be made of how the categories are distributed:

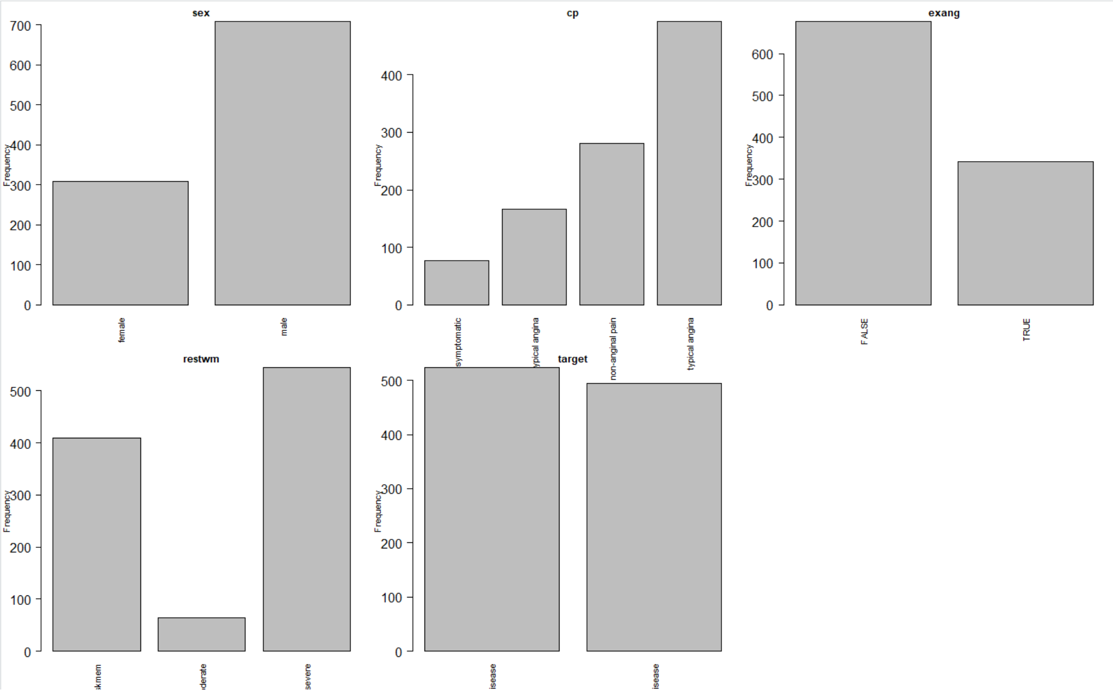

Figure 1: Distribution for Categories Features

The sex variable has 309 samples for females and 709 for males. The cp variable has four categories, with typical angina being the most common at 493 occurrences. The exang variable has two categories, with 677 instances of FALSE and 341 instances of TRUE. The restwm variable has three categories, with moderate or severe being the most frequent at 544 occurrences. The target variable has two categories, with 523 cases of disease and 495 cases of no disease.

Then explore Numeric Features:

	Min	Q1	Median	Mean	Q3	Max	Missing	Unique	Skewness	Kurtosis
Min	29.00	48.00	56.00	54.45	61.00	77.00	0	41	0.2518735	0.54736109
Q1	71.0	132.0	152.0	149.2	166.0	202.0	0	91	0.5158395	0.08791252
Median	0.000	0.000	0.800	1.075	1.800	6.200	0	40	1.2011371	1.26939911
Mean	0.0000	0.0000	0.0000	0.7593	1.0000	4.0000	0	5	1.2477696	0.65938246

Table 3: Summarize for Numeric Features

The above table presents descriptive statistics for four features that are related to heart disease. These features include the minimum heart rate, the first quartile heart rate, the median heart rate, and the mean heart rate. The minimum heart rate ranges from 29 to 77, with an average of 54.45. The first quartile heart rate ranges from 48 to 166, with an average of 149.2. The median heart rate ranges from 56 to 202, with an average of 149.2. The mean heart rate ranges from 0 to 4, with an average of 0.7593. The skewness and kurtosis values suggest that the distributions of these features are approximately normal, except for the mean heart rate which has a positive skew and a higher kurtosis value. These features can provide valuable insights into the relationship between heart rate and heart disease, and can be used to build decision tree models for predicting heart disease.

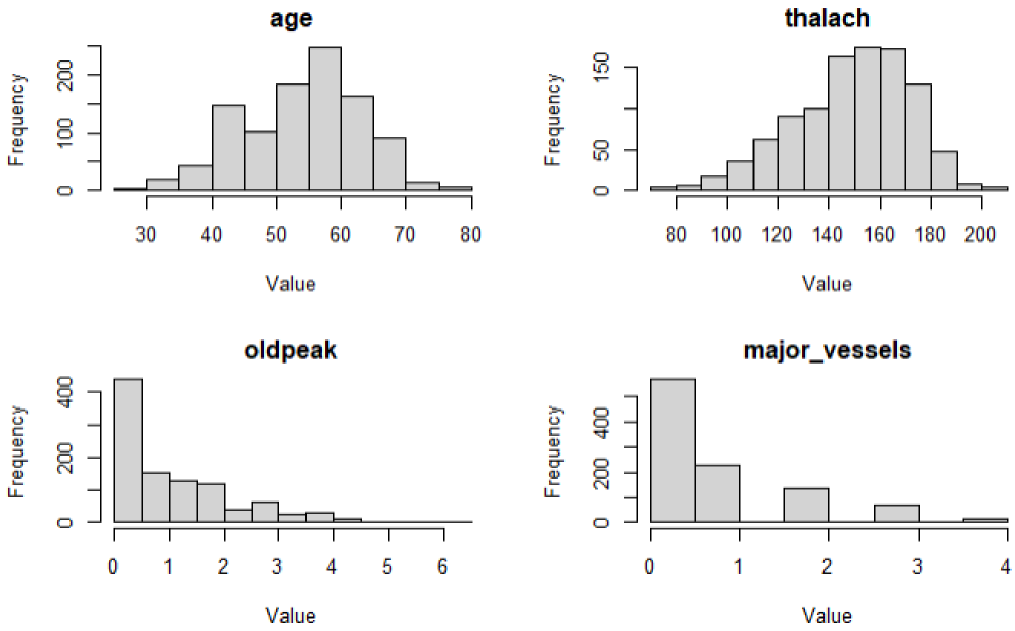

Figure 2: Distribution for Numeric Features

The above figure shows the visualization results of the distribution of numeric features, which is convenient for intuitive observation.

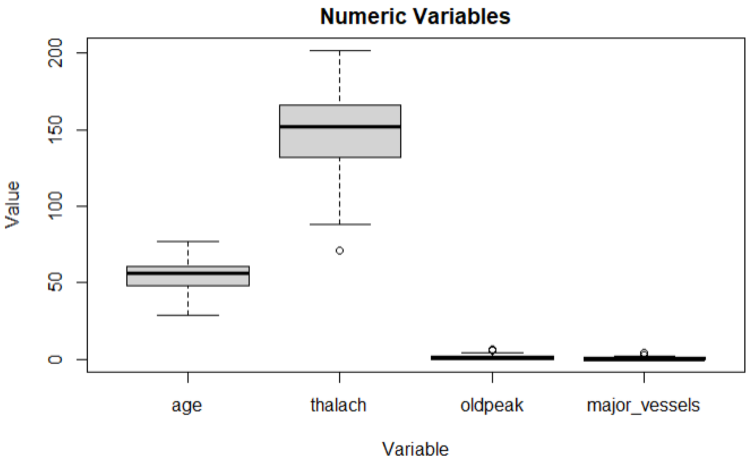

Figure 3: Boxplot for Numeric Features

Box plots provide a visual representation of the data distribution, including any potential outliers, and can be used to identify points located outside the whiskers or outside the box. The age data does not contain any apparent outliers, as both the minimum and maximum values fall within a reasonable range for human age. However, for thalach, there are some values that exceed both the upper quartile (Q3) and the maximum value. These values could potentially be considered outliers. No extreme or outlier values are present for the ST depression caused by exercise relative to rest (oldpeak) or the number of major vessels stained by fluoroscopy (major_vessels).

To mitigate the effect of extraneous data when modeling, Tukey's box plot method will be applied to identify outliers. Outliers will be identified as values that fall under the lower limit or above the upper limit as defined respectively by the first quartile (Q1) minus 1.5 times the interquartile range (IQR) and the third quartile (Q3) plus 1.5 times the IQR. The IQR will be defined as the range of the middle 50% of the data, computed as the difference between Q3 and Q1.

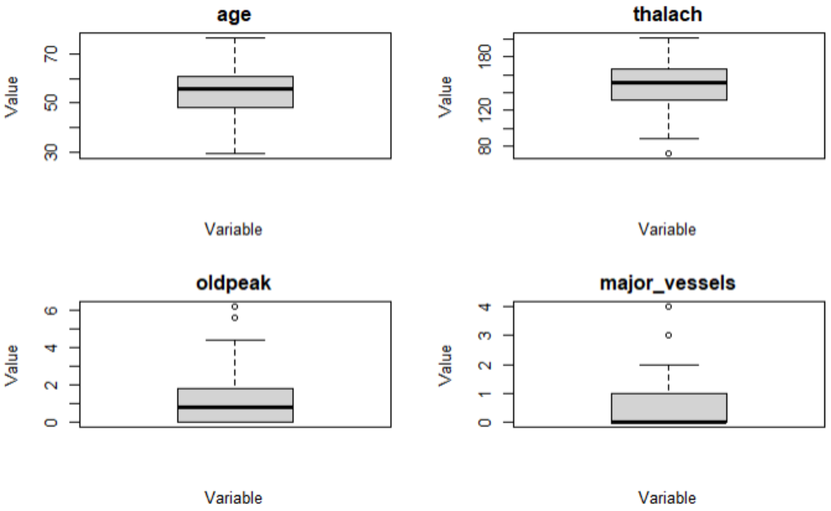

Figure 4: Boxplot for Further Analysis About Outliers in Numeric Features

Each numerical variable will then be examined for any values that fall outside the upper and lower limits, which are considered outliers and may indicate the presence of extreme data values.

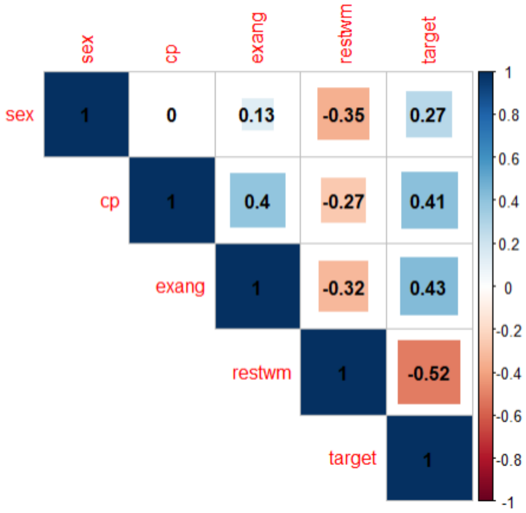

Figure 5: Correlation Matrix Plot with The Correlation Values for Categories Features

Several correlation coefficients were observed between different variables and the likelihood of heart attack. Notably, a positive correlation of 0.27 was found between sex and target, suggesting a potential association between gender and the probability of heart attack. Additionally, a moderate positive correlation of 0.41 was observed between cp and target, indicating a potential link between chest pain type and the likelihood of heart attack. Moreover, a moderate positive correlation of 0.43 was present between exang and target, implying a possible connection between exercise-induced angina and the probability of heart attack. Finally, a strong negative correlation of -0.52 was observed between restwm and target, suggesting a significant association between abnormal resting electrocardiogram and the probability of heart attack.

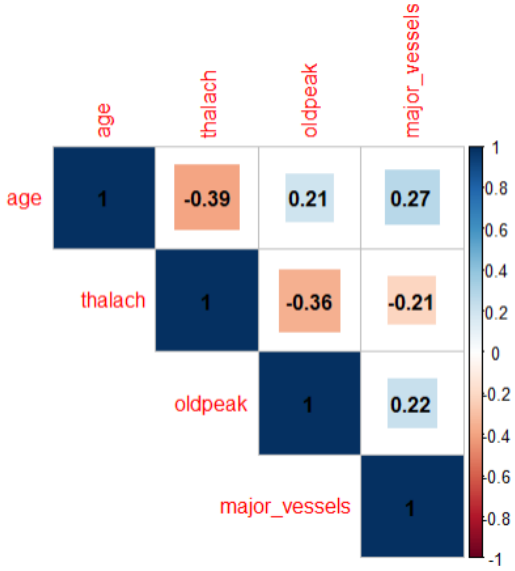

Figure 6: Correlation Matrix Plot with The Correlation Values for Numeric Features

A negative correlation of -0.392 was found between age and thalach, indicating that maximum heart rate tends to decrease as age increases. Additionally, a negative correlation of -0.355 was found between thalach and exercise-induced ST segment depression (oldpeak), suggesting that individuals with higher maximum heart rates tend to have lower levels of ST segment depression induced by exercise. Moreover, a positive correlation of 0.208 was observed between age and oldpeak, indicating that the level of ST segment depression induced by exercise may also increase as age increases. Finally, a positive correlation of 0.271 was observed between age and major_vessels, suggesting that the number of major vessels may also increase as age increases.

Building Decision Tree Models

Using 70% of the data for training provides ample samples to learn the parameters and features of the model, which enables the model to capture patterns and correlations in the data more effectively, thereby improving the model's generalization ability. Using 30% of the data for testing can evaluate the performance of the trained model. This segment of the data did not participate in the training process, so it can be used to verify the model's performance on unseen data, providing an estimate of the model's generalization ability and judging its effectiveness in practical applications. Meanwhile, using less testing data can better check whether the model is overfitting, reducing the risk of overfitting while improving the accuracy of the model. In machine learning, the 70% training and 30% testing ratio is widely accepted and used in many research and practice as it is believed to provide reasonable results in most cases.

To generate a decision tree using rpart with the default settings, refer to the output below:

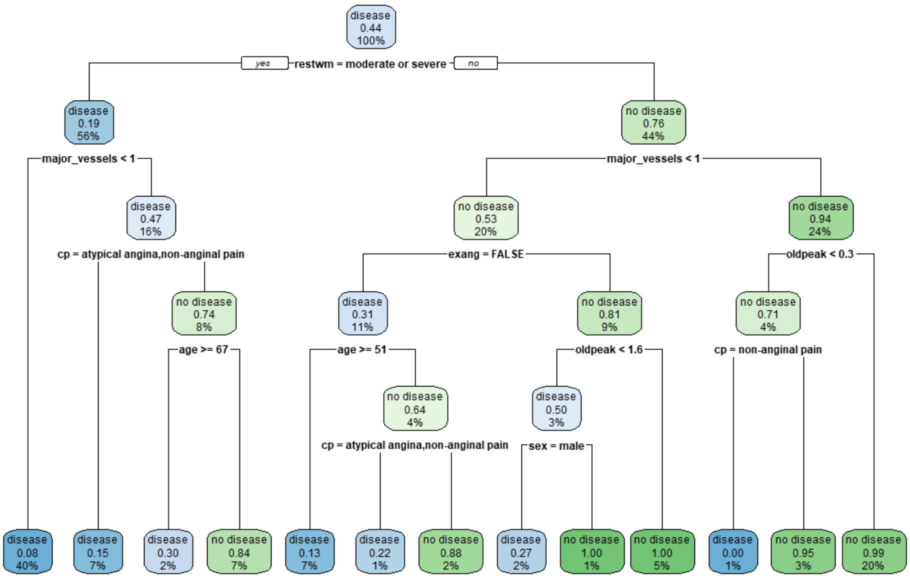

Figure 7: Plot for Decision Tree Model 1

In rpart, the complexity parameter (cp) is a hyperparameter that governs the complexity of the decision tree. It decides whether a split should be attempted or not, based on whether it can improve the overall fit by a factor. In cases where anova split is used, the overall R-squared value must increase cp at each step. The primary role of this parameter is to prune unnecessary splits and save computation time. The default value of cp is 0.05, resulting in an incomplete decision tree by default.

To obtain a complete decision tree, it is necessary to set the value of cp to 0.

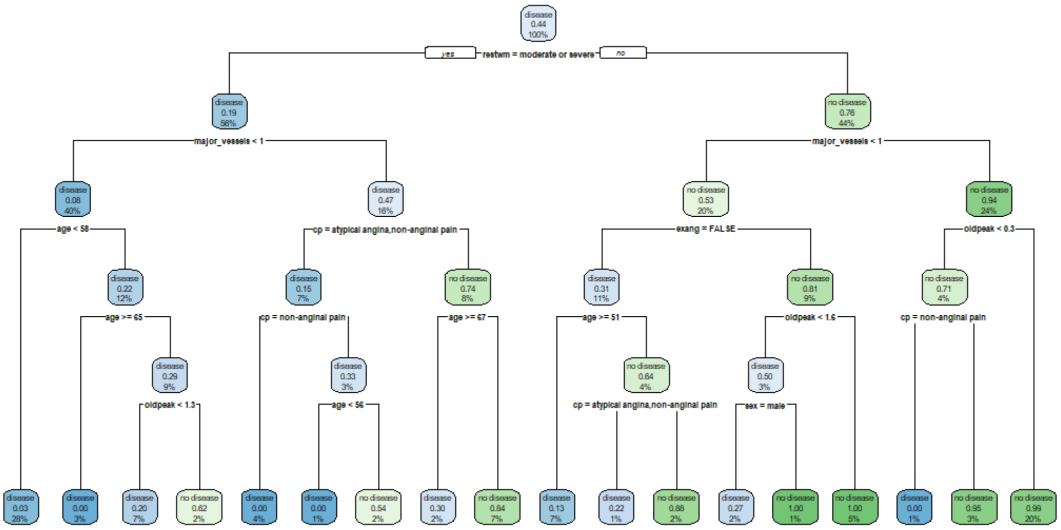

Figure 8: Plot for Decision Tree Model 2 with cp = 0

There have been modifications made to the conditions for splitting nodes, the order in which nodes are split, the introduction of new split points and leaf nodes, as well as changes made to the purity of leaf nodes. While the first split point in both decision trees remains the same, the second decision tree introduces new split points and leaf nodes that were not present in the first decision tree. Additionally, certain leaf nodes in the second decision tree have a different level of purity when compared to their corresponding leaf nodes in the first decision tree.

The minsplit parameter is used in the rpart algorithm to determine the minimum number of samples required in a node before a split can be attempted. When set to 50, the algorithm requires at least 50 samples in each node before considering a split. This parameter can be used to control the complexity of the decision tree, limit its growth, and prevent overfitting. However, it can also affect the depth and accuracy of the tree. A larger minsplit value may result in a shallower tree and lower accuracy on the training set. On the other hand, if the value is set too high, the algorithm may fail to capture specific patterns and rules in the training set due to noise and other details.

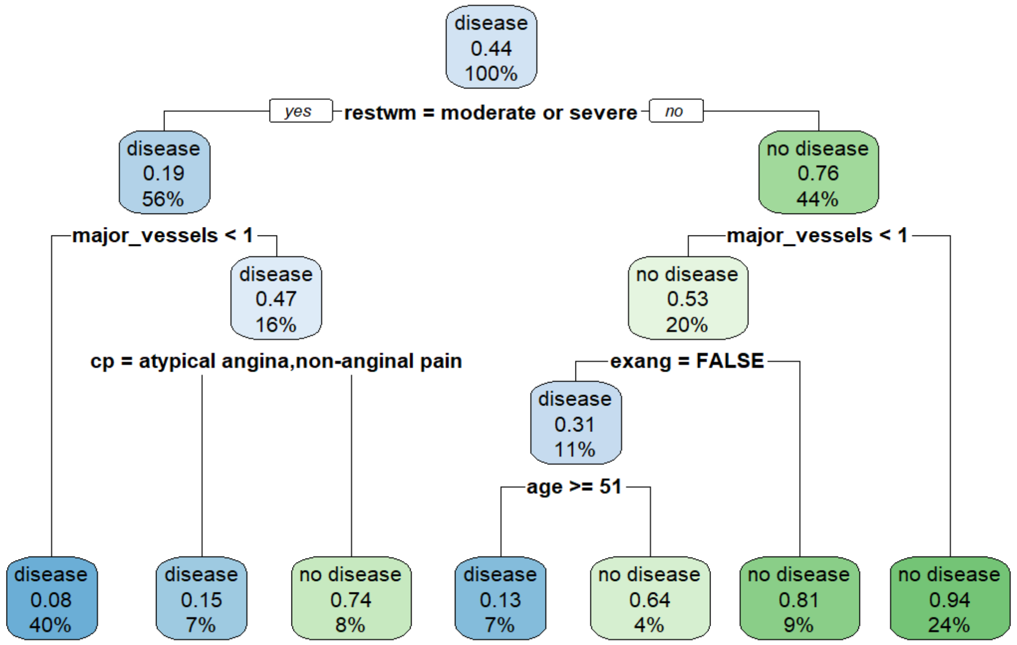

Figure 8: Plot for Decision Tree Model 3 with minsplit = 50

Both models begin with a root node and split based on the "restwm" variable. However, tree_model_03 splits based on additional variables, including "major_vessels," "cp," "exang," and "age," whereas tree_model_01 only splits based on "major_vessels" and "cp." The terminal nodes in tree_model_03 display predicted probabilities for each node, which differ from those in tree_model_01, indicating that the two models have different predicted outcomes for the given dataset.

The rpart function's maxdepth parameter influences how deep a decision tree can be. Smaller values result in simpler trees with fewer branches and nodes, which makes them easier to interpret and reduces computational complexity. By modifying maxdepth, a trade-off between model complexity and performance can be achieved. Experimenting with different values can help identify the optimal depth for a given dataset and problem.

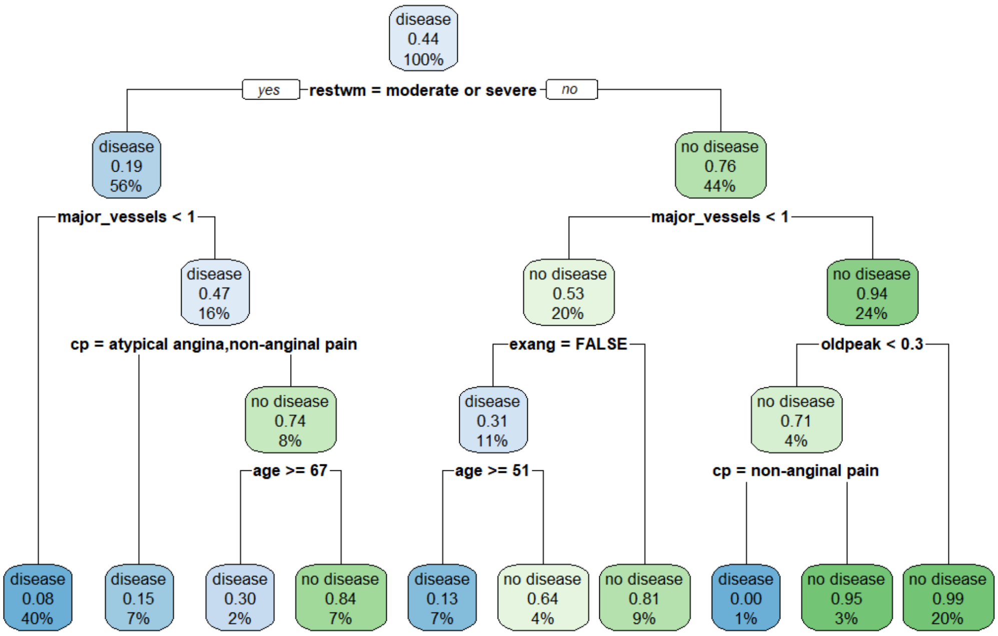

Figure 9: Plot for Decision Tree Model 4 with maxdepth = 4

By setting maxdepth = 4 in tree_model_04, the decision tree has been simplified, resulting in a shallower depth, and a reduced number of splits and nodes. While this can improve interpretability and reduce overfitting, it may also lead to a slightly higher misclassification rate compared to tree_model_01. The reduction in depth makes the tree easier to understand, while limiting the number of splits and nodes helps prevent overfitting and makes the tree less prone to capturing noise or outliers in the data. The "loss" column in the tree output shows that some terminal nodes in tree_model_04 have a higher misclassification count than in tree_model_01.

Model	Accuracy	Precision	F-score	AUC
tree_model_01	0.8448	0.7938	0.8552	0.8456
tree_model_02	0.8520	0.8200	0.8571	0.8525
tree_model_03	0.8267	0.8156	0.8273	0.8456
tree_model_04	0.8339	0.8182	0.8357	0.8342

Table 4: Model Performance Parameters

When evaluating model performance and prediction accuracy, several metrics are commonly used. These include accuracy, precision, F-score, and AUC.

Accuracy measures overall classification accuracy, while precision measures the accuracy of positive predictions. F-score, on the other hand, considers both precision and recall, making it suitable for evaluating imbalanced datasets. Finally, AUC measures the ability of the model to classify positive and negative examples.

Compare Models

After comparing the four decision tree models, we determined that tree_model_02 is the best option because it has the highest accuracy of 0.8520 among all the models, indicating the overall correctness of its predictions. Additionally, it has the highest precision of 0.8200, the highest F-score of 0.8571, and the highest AUC of 0.8525. Precision, F-score, and AUC are useful metrics for evaluating classification models, as they measure the proportion of correctly predicted positive cases, the balance between precision and recall, and the model's ability to distinguish between positive and negative cases, respectively. Therefore, tree_model_02 is the most reliable model in identifying positive cases correctly while minimizing false positives and negatives.

In the decision tree model (tree_model_02), the key attributes that determine the presence or absence of disease are restwm, major_vessels, cp, and age. The restwm attribute distinguishes between disease and no disease cases based on the severity of the symptom. The major_vessels attribute indicates the number of major vessels colored by fluoroscopy, with a value less than 0.5 indicating a higher likelihood of disease. The cp attribute represents chest pain type, with different categories such as "atypical angina, non-anginal pain" and "asymptomatic, typical angina" contributing to the prediction of disease or no disease. Age is used in several splits throughout the tree, with specific thresholds creating different branches and indicating its importance in predicting disease. Among these attributes, major_vessels and cp appear to be particularly predictive, providing valuable insights into the factors influencing the likelihood of cardiovascular disease.

References

Attia, Z. I., et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

Dey, D., et al. (2018). Machine learning in cardiovascular medicine: Are we there yet? Heart, 104(14), 1156-1164.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Mozaffarian, D., et al. (2015). Heart disease and stroke statistics-2015 update: A report from the American Heart Association. Circulation, 131(4), e29-e322.

World Health Organization. (2020). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)