1. Describe the Data: Select a dataset and describe the outcome variable and features for prediction. Explain how the outcome variable relates to the features and what insights can be gained from analyzing the data.

The outcome variable is: “Target”: 0= less chance of heart attack 1= more chance of heart attack.

The features are Age : Age of the patient, Sex : Sex of the patient, exang: exercise induced angina (1 = yes; 0 = no), ca: number of major vessels (0-3), cp : Chest Pain type chest pain type, trtbps : resting blood pressure (in mm Hg), chol : cholestoral in mg/dl fetched via BMI sensor, fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), rest\_ecg : resting electrocardiographic results, thalach : maximum heart rate achieved.

Some insights that can be gained are risk factors for a heart attack based on age, sex, exercise levels, cholesterol levels, and more.

1. Find the Best Value for Minimum Split: Use the decision tree algorithm to find the best value for the minimum split for the dataset. Explain your reasoning for selecting the optimal value.

Minimum split for 30:

The correct classification rate based on the training set is 80.43%

The correct classification rate based on the validation set is 68.91%

Minimum split for 20:

The correct classification rate based on the training set is 86.41%

The correct classification rate based on the validation set is 71.43%

Minimum split for 10:

The correct classification rate based on the training set is 88.04%

The correct classification rate based on the validation set is 69.75%

Based on the findings, a minimum split of 20 is the best choice based on the correct classification rates for both training and validation. It has the second highest training set and the highest validation set. The combination of this makes it the best choice.

1. Find the Node with the Highest and Lowest Probability: After building the decision tree model, identify the node with the highest and lowest probability. Explain what insights can be gained from analyzing these nodes and how they relate to the outcome variable and features.

The node with the highest probability: Node #4, Probability: [0.0, 1.0]

Node 4 has thalach (maximum heart rate achieved) less than or equal to 2.5. It also has no exercise induced angina (exang: exercise induced angina (1 = yes; 0 = no)), the person is less than or equal to 56.5 years of age, and the number of major blood vessels is 0 (ca: number of major vessels (0-3).

The node with the lowest probability: Node #17, Probability: [1.0, 0.0]

Node 17 has thalach (maximum heart rate achieved) greater than 2.5. It also has an old peak (previous peak) of greater than .85, and has chest pain of less than or equal to 1.5 meaning chest pain is fairly typical (Value 1: typical angina and Value 2: atypical angina).

Understanding the insights from both nodes helps in refining treatment strategies based on certain variables listed above for patients with heart disease or at risk of developing it.