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.

This dataset is covering titanic passengers. The outcome variable is survived or not survived. The features for prediction are male or female, age, and fare. The outcome variable relates to the features because there should be a statistical difference in the survival rate of the different demographics on the ship (considering factors like women and children being saved before men, and wealthier passengers getting more privilege).

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.

The best minimum split for this dataset was 63. This seemed to be the best number because it produced a relatively high accuracy with train at 80.14% and validation at 76.57%. Any lower, and there starts to be issues with overfitting. At 63, the accuracy is high while maintaining a optimal number of nodes.

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.

Node 7 has the highest probability of survival with a 96.1% chance. Node 16 is the lowest probability with a 0% chance of survival. This tells us that women who paid a fare of $40.64 or more were 96.1% likely to survive the accident. Men who paid $26.27 or more and are 57 or older has a 0% chance of surviving.