Mohammad Bagher Soroush - 400130273 - ML 4 Report

1 .Explanation of Process Steps

The provided code involves the following steps:

Data Preprocessing:

- Load and inspect the dataset.
- Handle missing values, if any.
- Encode categorical variables and scale numerical ones as necessary.

Model Training:

- Split the data into training and testing sets.
- Train a Decision Tree Classifier using parameters such as max_depth, min_samples_split, and min_samples_leaf.
- We trained the model base on two scenarios and got the results.
- In the second scenario we used different methods for filling the missing values and picked the best one.

Evaluation:

- Evaluate the model's performance using metrics like accuracy, precision, and recall.
- Visualize the decision tree to interpret its splits and structure.

2. Converting Non-Categorical to Categorical Data

To convert non-categorical (numerical) data into categorical data, common techniques include:

Binning/Discretization:

- Divide numerical values into discrete intervals (bins).
- Example: Converting ages into ranges like 0-18, 19-35, etc.

Thresholding:

- Apply thresholds to numerical features to create binary categories.
- Example: Classifying temperatures as "high" or "low" based on a threshold.

One-Hot Encoding:

- Convert categories into binary columns, each representing a category.
- Example: Encoding Color = {Red, Blue, Green} into separate binary features.

3 .Effect of max_depth on the Model

The max_depth parameter controls the maximum depth of the decision tree:

- Low max_depth:
  Results in a simpler model.

Reduces the risk of overfitting but may lead to underfitting.

- High max_depth:

  Produces a more complex model capable of capturing detailed patterns.

  Increases the risk of overfitting, especially on small datasets.

Setting the appropriate max_depth involves balancing bias and variance through validation techniques.

4 .Limitations of min_samples_split and min_samples_leaf

- min_samples_split:
  Controls the minimum number of samples required to split a node.
  High values: Reduces overfitting by limiting splits, leading to simpler trees.
  Low values: Allows more splits, capturing finer details but increasing overfitting risk.

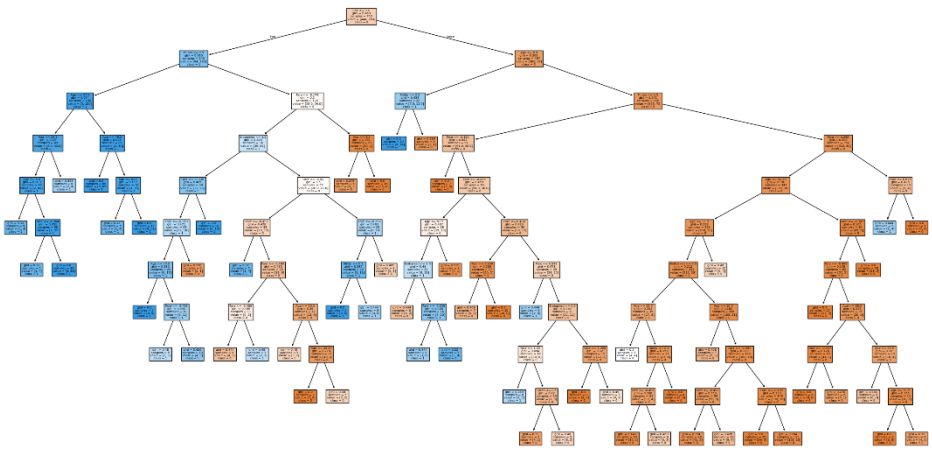
- min_samples_leaf:
  Sets the minimum number of samples required in a leaf node.
  High values: Forces larger leaf nodes, which may smooth out noise but overlook small patterns.
  Low values: Allows smaller leaves, potentially capturing noise and overfitting.

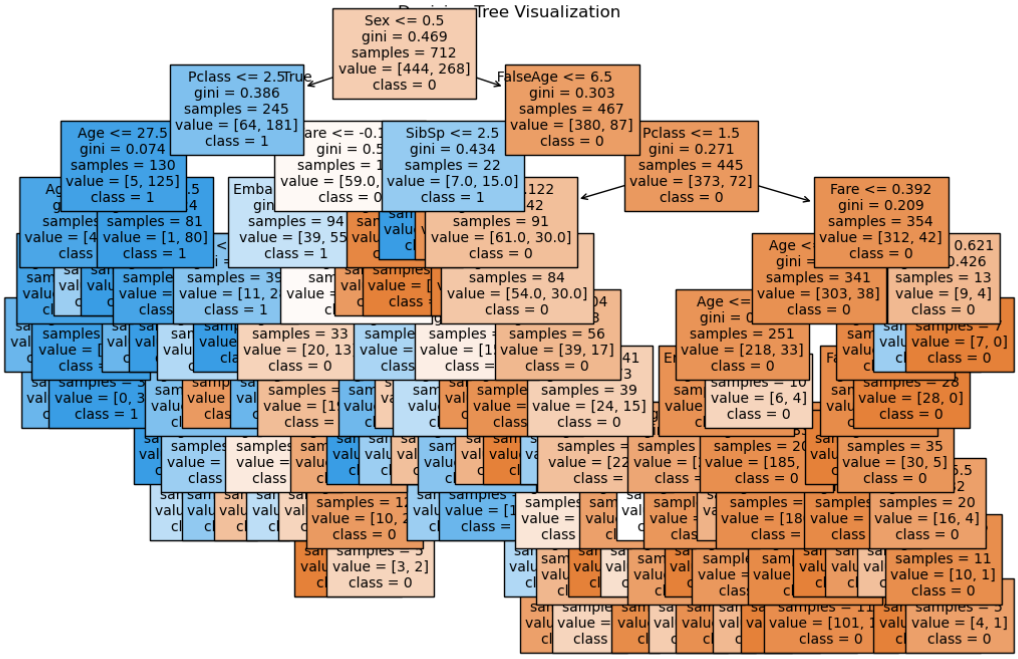
Recommendations:

Choose these parameters carefully based on dataset size and complexity, using techniques like cross-validation.

5 .Decision Tree Visualization

The decision tree structure is visualized below, showing how the model splits data based on feature values:





This tree illustrates the sequence of splits made to classify data, including thresholds, sample sizes, and feature importance.