Assignment 2: Decision Tree

DSCI 6601: Practical Machine Learning

Submission

For this assignment, select a publicly available dataset of your choice that contains at least five features. The goal is to build a Decision Tree model for either classification or regression, depending on the nature of your dataset. You will evaluate the model's performance and analyze how the features contribute to the predictions. Complete the tasks outlined below, and ensure to include both your Python code and the dataset in your submission.

Question for Decision Trees

Based on your chosen dataset:

1. **Dataset Overview:**

   - Provide a basic overview of the dataset, including details such as feature names, target variable, presence of missing values, and identification of categorical features.

2. **Data Preprocessing:**

   - Describe how you would preprocess the dataset before building a Decision Tree model. Discuss steps such as handling missing values, encoding categorical features, and scaling features if necessary.
   - Provide the code for each preprocessing step and explain why these steps are important for decision trees.

3. **Model Implementation:**

   - Implement a Decision Tree classifier using Scikit-learn.
   - Explain how you would split the data into training and testing sets, train the model, and evaluate it using appropriate metrics.
   - Include the code and interpretation of the results.

4. **Feature Importance:**

   - Decision Trees provide feature importance scores that help understand which features contribute the most to the predictions.

- Extract and visualize the feature importance for the dataset.
- Discuss how you would use this information to either refine the model or explain the results to a non-technical audience.
- Provide code and an interpretation of the feature importance scores.

Submission

Your submission should include the following:

- Comprehensive responses to all the tasks outlined above, accompanied by the Python code and outputs from your Jupyter Notebook.

- Clear and concise explanations of each step taken, with well-documented text and comments included in your Jupyter Notebook for every step.

- A thorough analysis of the results within the Jupyter Notebook.

- The dataset used for the assignment.