Problem Statement 3 for Machine Learning Lab: Decision Trees

Objective:

The objective of this lab is to understand and apply the concepts of Decision Trees, focusing on constructing Decision Trees using the Gini Index for both regression and classification purposes. Additionally, students will learn to evaluate the performance of their models using various performance metrics such as the Confusion Matrix, Kappa Statistics, Sensitivity, Specificity, Precision, Recall, F-measure, and the ROC curve.

Datasets:

- **Classification Dataset**: Iris Dataset(https://archive.ics.uci.edu/dataset/53/iris)
    - **Description**: The Iris dataset consists of 150 samples of iris flowers, with 4 features (sepal length, sepal width, petal length, petal width) and a target variable indicating the species (Setosa, Versicolor, Virginica).
- **Regression Dataset**: Boston Housing Dataset(https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset)
    - **Description**: The Boston Housing dataset contains 506 samples with 13 features (e.g., average number of rooms per dwelling, per capita crime rate by town) and a target variable indicating the median value of owner-occupied homes.

Tasks:

1. **Understanding Decision Trees**:
    - Review the theoretical background of Decision Trees.
    - Understand the process of constructing Decision Trees using the Gini Index for classification and mean squared error for regression.
2. **Constructing Decision Trees**:
    - Implement a Decision Tree for a classification task using the Gini Index.
    - Implement a Decision Tree for a regression task using mean squared error.
    - Utilize the CART algorithm to build both classification and regression trees.
3. **Performance Metrics**:
    - Learn about various performance metrics including:
        - **Confusion Matrix**: Understand its components (True Positives, True Negatives, False Positives, False Negatives) and how to compute them.
        - **Kappa Statistics**: Understand the calculation and interpretation of the Kappa coefficient to assess the agreement between predicted and observed classifications.
        - **Sensitivity and Specificity**: Calculate and interpret these metrics to understand the true positive rate and true negative rate.
        - **Precision and Recall**: Calculate and interpret these metrics to understand the positive predictive value and the true positive rate.
        - **F-measure**: Calculate the harmonic mean of Precision and Recall to assess the balance between them.
        - **ROC Curve**: Plot the ROC curve and calculate the Area Under the Curve (AUC) to evaluate the model's performance.

Steps to Follow:

1. **Data Preparation**:
    - **Classification Dataset (Iris)**:
        - Load the Iris dataset.
        - Explore the dataset to understand its structure.
        - Split the dataset into training and testing sets.
    - **Regression Dataset (Boston Housing)**:
        - Load the Boston Housing dataset.
        - Explore the dataset to understand its structure.
        - Split the dataset into training and testing sets.
2. **Implementing Decision Trees**:
    - **Classification Task**:
        - Construct a Decision Tree classifier using the Gini Index.
        - Train the classifier on the training set.
        - Predict the target variable on the testing set.
        - Evaluate the classifier's performance using the specified metrics.
    - **Regression Task**:
        - Construct a Decision Tree regressor using mean squared error.
        - Train the regressor on the training set.
        - Predict the target variable on the testing set.
        - Evaluate the regressor's performance using the specified metrics.
3. **Evaluating Performance**:
    - **Classification Task**:
        - Construct and interpret the Confusion Matrix.
        - Calculate and interpret the Kappa Statistics.
        - Calculate Sensitivity, Specificity, Precision, Recall, and F-measure.
        - Plot the ROC curve and calculate the AUC.
    - **Regression Task**:
        - Calculate the mean squared error.
        - Plot actual vs. predicted values.
4. **Analysis and Interpretation**:
    - Compare the performance of the Decision Tree models using the different metrics.
    - Discuss the strengths and weaknesses of the models based on the evaluation results.
    - Provide insights and recommendations for improving model performance.