Problem Statement 4 for Machine Learning Lab: Random Forest

Objective:

The objective of this lab is to understand and apply the concepts of Random Forests, focusing on constructing and evaluating Random Forest models for both classification and regression tasks. Students will learn to implement Random Forests using Python, train the models on datasets, and evaluate their performance using various metrics such as accuracy, precision, recall, F1-score, mean squared error (MSE), and R-squared.

Datasets:

- **Classification Dataset**: UCI Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic) Dataset
    - **Description**: The dataset consists of 569 samples with 30 features and a target variable indicating the diagnosis (malignant or benign).
    - **Link**: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- **Regression Dataset**: UCI Machine Learning Repository - California Housing Dataset
    - **Description**: The dataset contains 20,640 samples with 8 features and a target variable indicating the median house value for California districts.
    - **Link**: https://www.kaggle.com/datasets/camnugent/california-housing-prices

Tasks:

1. **Understanding Random Forests**:
    - Review the theoretical background of Random Forests.
    - Understand the process of constructing Random Forests for both classification and regression tasks.
2. **Constructing Random Forest Models**:
    - Implement a Random Forest classifier.
    - Implement a Random Forest regressor.
    - Utilize the Random Forest Classifier and Random Forest Regressor from the sklearn library.
3. **Performance Metrics**:
    - Learn about various performance metrics including:
        - **Classification Metrics**: Accuracy, Precision, Recall, F1-score, ROC curve, and AUC.
        - **Regression Metrics**: Mean Squared Error (MSE), R-squared.

Steps to Follow:

1. **Data Preparation**:
    - **Classification Dataset (Breast Cancer Wisconsin)**:
        - Load the Breast Cancer Wisconsin dataset from the provided link.
        - Explore the dataset to understand its structure.
        - Split the dataset into training and testing sets.
    - **Regression Dataset (California Housing)**:
        - Load the California Housing dataset from the provided link.
        - Explore the dataset to understand its structure.

        - Split the dataset into training and testing sets.
2. **Implementing Random Forest Models**:
    - **Classification Task**:
        - Construct a Random Forest classifier.
        - 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 Random Forest regressor.
        - 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**:
        - Calculate accuracy, precision, recall, and F1-score.
        - Plot the ROC curve and calculate the AUC.
    - **Regression Task**:
        - Calculate the mean squared error and R-squared.
        - Plot actual vs. predicted values.
4. **Analysis and Interpretation**:
    - Compare the performance of the Random Forest 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.