Programming Assignment 2: Classification Task and Performance Evaluation (10 points)

Shen-Shyang Ho (Dr.)

October 2, 2023

- In this assignment, you will be using the dataset assigned to you in Assignment 1.
- You will be assigned three classification methods from the following classification methods: **Naive Bayes Classifier, Support Vector Machine (SVM), Decision Tree, Neural Network, Random Forest, Adaboost**
- Scikit-learn (https://scikit-learn.org/stable/user_guide.html) will be used in this assignment.
    1. Naive Bayes Classifier: **GaussianNB** with default parameters.
    2. Support Vector Machine (SVM): **LinearSVC** with default parameters.
    3. Decision Tree: **DecisionTreeClassifier** with parameter **max_depth=10** and default values for the other parameters.
    4. Neural Network: **MLPClassifier** with parameter $hidden\_layer\_sizes = (10, 10, 10,)$ (i.e., 3 hidden layers with 10 nodes each) and default values for the other parameters.
    5. Random Forest: **RandomForestClassifier** with default parameters.
    6. Adaboost: **AdaBoostClassifier** with default parameters.

1. Use images from ALL FOUR classes. Convert the images to grayscale pixel intensity histograms. (These will be the vector representations of the images). This will be your dataset for Part 2. (0.25 point)
2. Perform standardization on the dataset. (see https://scikit-learn.org/stable/modules/preprocessing.html (0.25 point)
3. Split the dataset into a training set and a test set: For each class, perform a training/test split of 80/20. (0.25 point)
4. (Model Selection) Perform a standard 5-fold cross-validation and a stratified 5-fold cross-validation on the **training set** for k-Nearest Neighbor Classifiers such that $k = 1, 3, 5, 7, 10, 20$. (2.5 points)
    - Plot a graph (x-axis: k; y-axis: mean validation/training error (%)) containing four error curves (2 validation error curves and 2 training error curves - label them clearly using a legend to define the curves). Which k has the lowest mean error for each curve? Comment about (1) the model complexity for k-Nearest Neighbor classifier in relation to k, and (2) when/whether there is overfitting/underfitting. (1.5 points)
    - Use the k value with the lowest mean validation error for your k-Nearest Neighbor classifier from the stratified 5-fold cross-validation. What is the test error? (0.25 point)

5. (Performance Comparison) Perform stratified 5-fold cross-validation on the 4-class classification problem using the three classification methods (available on canvas) assigned to you. Plot the confusion matrices for the three approaches (clearly label the classes) using the test set (See Figure 1). (If you use code from any website, please do proper referencing. You will get 0 point for this assignment without proper referencing) (3.75 points)

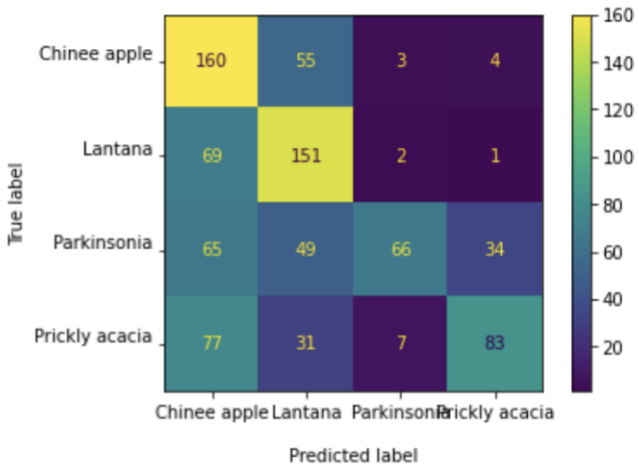

Figure 1:

- Based on the confusion matrices (on the test set), which do you think is the best method? Why? (0.50 point)
- Based on the mean validation accuracies (from the 5-fold cross-validation) for the three methods. Which is the best method? (0.25 point)
- Compute the accuracies for the three methods on the test set. Which is the best method? (0.25 point)
- Compute the F-measure for the three methods on the test set. Which is the best method? (0.25 point)