DecisionTrees & RandomForests

DeadLine: 14 Farvardin 1402 Assignment 1

1 Classification on Covertype Data Set

In this problem set, you will be working with a dataset that includes tree observations from four areas of the Roosevelt National Forest in Colorado. The dataset contains 581,012 samples with 55 columns. The last column is the label (7 different classes) and the first 54 columns are related to the input features. Your task is to perform classification on this dataset.

2 Instructions

1. Load the data and separate the labels (corresponding to the Cover_Type column) from the training features. y corresponds to the Cover_Type column, and X corresponds to the data in the first 54 columns.

2. Shuffle the data and divide it into training and test data with a 70:30 ratio. Name the training data X_train and y_train, and the test data X_test and y_test.

3. Perform pre-processing on the data if necessary.

4. Use 5-fold cross-validation to pick the best model parameters for the Decision Tree and Random Forest models. In this process, a range of values for the model parameters is selected, and then k-fold cross-validation is performed on the training data for each set of parameters. The average validation performance is computed across the k-folds for each set of parameters, and the set of parameters that gives the best average performance is selected. By trying different parameter settings and selecting the one that gives the best average performance, we can optimize the model and obtain the best possible results on unseen data. Do not use the test data to perform cross-validation.

5. Perform classification using Decision Tree and Random Forest models with the best parameters obtained using 5-fold cross-validation. Mention the best parameters for these models in your report.

6. In your report file, describe the evaluation metrics: accuracy, precision, recall, and f1_score.

7. Provide these 4 evaluation metrics along with a confusion matrix for both training and test data.

8. Analyze all the obtained confusion matrices in your report file.

3 Additional Guidance

- Make sure your code is in .ipynb format.

- You can use pandas.read_csv command to read the data.

- Along with your code, please include a report file that thoroughly analyzes your results.

- Use appropriate visualizations and statistics to support your analysis and conclusions.