Assignment: ML (3)

Dataset:

Use Pen-Digits datasets (train dataset & test dataset) with provided splits to solve questions.

Decision tree

1- Generate a scatterplot matrix to show the relationships between the variables and a heatmap to determine correlated attributes, then write a summary of what you noticed.
2- Ensure data is in the correct format for downstream processes (e.g., remove redundant information, convert categorical to numerical values, address missing values, etc.)
3- Fit a decision tree to the training data. Plot the tree, interpret the results, and display accuracy and Confusion Matrix.
4- Try different ways to improve the decision tree algorithm (e.g., use different splitting strategies, prune tree after splitting). Does pruning the tree improves the accuracy?

Bagging

(Bagging is to generate a set of bootstrap datasets, create estimators for each bootstrap dataset, and finally utilize majority voting (soft or hard) to get the final decision.)

1- Apply bagging strategy to classify test set samples by using SVM and Decision Tree algorithm as base estimators. Display accuracy and Confusion Matrix.
2- Apply Random Forest algorithm (the baseline), then fine tune this baseline. For the number of estimators, Try 5 different values within the interval of [10, 200]. Plot accuracy vs. number of estimators.

Boosting

1- Use GradientBoosting classifier to classify test set samples. There are 2 important hyperparameters in GradientBoosting, i.e., the number of estimators, and learning rate. First, tune number of estimators parameter by trying 4 values in the interval of [10, 200]. Then by using the tuned value for number of estmators, tune the learning rate parameter by trying 4 values within the range of [0.1, 0.9]. Display accuracy and Confusion Matrix separately for the best value of both parameters (Number of estimators and learning rate).
2- Build XGBoost classifier with the same parameters that you obtained in the last one. Provide accuracy and Confusion Matrix.
3- Comment on Bagging and Boosting approaches.