**ASSIGNMENT 1**

**Description of the dataset:**

The dataset I chose is Phoneme. As the name suggests, this dataset differentiates between nasal (class 1) and oral (class 2) vowel sounds. There are five features and one target in this ML task. The five features are amplitudes of the first five harmonics. The positive value corresponds to a local maximum of the spectrum, and the negative value corresponds to the minimum value of the range. There are 5404 instances in the dataset. The output is classified as 1 or 2 if it's nasal or oral sound, respectively.

**Configurations of the methods:**

* **Decision tree with default parameters:** For the cross-validation, the min\_samples\_leaf is 1, the evaluation measure is roc\_auc (since the task is binary classification), and the number of folds for cross-validation is 10.
* **Decision tree with tuned min\_samples\_leaf using GridSearchCV:** For the parameter tuning, the min\_samples\_leaf values taken are 2,4,6,8,10. Through internal cross-validation, the sklearn has chosen the value 10. The evaluation measure is roc\_auc, and the number of folds for cross-validation is 10.
* **Random Forest:** The evaluation measure is roc\_auc, and the number of folds is 10.
* **Bagged decision tree:** The evaluation measure is roc\_auc, and the number of folds is 10.
* **AdaBoosted decision tree:** The evaluation measure is roc\_auc, and the number of folds is 10.

**Table:**

|  |  |
| --- | --- |
| **Method** | **AUC score** |
| Decision tree with default parameters | 0.843683018812787 |
| Decision tree with tuned min\_samples\_leaf using GridSearchCV | 0.9028943026215821 |
| Random Forest decision tree | 0.9642744033267385 |
| Bagging decision tree | 0.9519098283385421 |
| AdaBoosted decision tree | 0.8912110938033699 |

**Results and Conclusion:** The decision tree with tuned parameters performed better than the decision tree with default parameters. Overall, the Random Forest decision tree has served better than the other four decision trees, with an AUC score of 0.964. Out of the five decision trees, the decision tree with default parameters has the lowest AUC score of 0.84. Perhaps, by doing the error analysis, exploring more models, more fine-tuning the parameters, the AUC score can be improved.