CS 6375: Machine Learning Homework-3

1. Tree Classifiers

1. Decision Tree:

Set	Accuracy	F1 Score	criterion	splitter	max_depth	max_features
c300_d100	0.555	0.5527	gini	random	500	Log2
c300_d1000	0.569	0.5681	entropy	best	500	sqrt
c300_d5000	0.6085	0.607	gini	Random	500	sqrt
c500_d100	0.685	0.6865	gini	Best	1000	sqrt
c500_d1000	0.6405	0.6463	Entropy	Random	500	sqrt
c500_d5000	0.6737	0.675	Entropy	Random	1000	sqrt
c1000_d100	0.66	0.653	Entropy	Random	1000	sqrt
c1000_d1000	0.727	0.7278	gini	Best	500	sqrt
c1000_d5000	0.805	0.807	Entropy	Random	500	sqrt
c1500_d100	0.81	0.8173	Entropy	Random	1000	sqrt
c1500_d1000	0.8765	0.879	Entropy	Random	500	sqrt
C1500_d5000	0.9277	0.9285	Entropy	Best	1000	sqrt
C1800_d100	0.92	0.9207	Entropy	Random	500	sqrt
C1800_d1000	0.9475	0.9479	gini	Random	500	sqrt
C1800_d5000	0.9701	0.9702	Entropy	Best	1000	sqrt

- More detailed logs of all the hyperparameter combinations are found in the OutputLogs folder.

2. Bagging:

Set	Accuracy	F1 Score	criterion	splitter	max_depth	max_features
c300_d100	0.63	0.637	Entropy	Best	2	sqrt
c300_d1000	0.6865	0.698	gini	Random	2	sqrt
c300_d5000	0.7839	0.791	gini	random	10	sqrt
c500_d100	0.685	0.67	gini	Best	2	sqrt
c500_d1000	0.7815	0.785	gini	Best	2	sqrt
c500_d5000	0.8539	0.8583	gini	Best	10	sqrt
c1000_d100	0.85	0.838	Entropy	Random	50	sqrt
c1000_d1000	0.9125	0.9121	Entropy	Random	10	sqrt
c1000_d5000	0.9573	0.9578	gini	Best	10	sqrt
c1500_d100	0.935	0.9377	gini	Best	2	sqrt
c1500_d1000	0.9895	0.9895	gini	Best	10	sqrt
C1500_d5000	0.9959	0.9959	Entropy	Best	10	sqrt
C1800_d100	1.0	1.0	Entropy	Random	500	Log2
C1800_d1000	0.999	0.999	gini	best	10	sqrt
C1800_d5000	0.999	0.999	gini	Random	10	sqrt

- More detailed logs of all the hyperparameter combinations are found in the OutputLogs folder.

3. Random Forest:

Set	Accuracy	F1 Score	max_features	criterion	n_estimators	max_depth
c300_d100	0.855	0.8625	Log2	entropy	1000	50
c300_d1000	0.917	0.9177	sqrt	entropy	1000	50
c300_d5000	0.9384	0.9396	sqrt	entropy	1000	50
c500_d100	0.95	0.9505	Log2	gini	1000	50
c500_d1000	0.973	0.9733	Log2	gini	1000	100
c500_d5000	0.9733	0.9775	Log2	entropy	1000	100
c1000_d100	1.0	1.0	Log2	gini	1000	50
c1000_d1000	0.998	0.998	Log2	entropy	1000	100
c1000_d5000	0.9989	0.9989	Log2	entropy	1000	50
c1500_d100	1.0	1.0	sqrt	gini	100	50
c1500_d1000	1.0	1.0	sqrt	gini	100	100
C1500_d5000	1.0	1.0	Log2	gini	100	50
C1800_d100	1.0	1.0	Sqrt	Gini	100	50
C1800_d1000	1.0	1.0	Sqrt	Gini	100	50
C1800_d5000	1.0	1.0	Sqrt	Gini	100	50

- More detailed logs of all the hyperparameter combinations are found in the OutputLogs folder.

4. Boosting:

Set	Accuracy	F1 Score	loss	n_estimators	criterion	Learning_rate
c300_d100	0.86	0.864	deviance	200	Friedman_mse	0.1
c300_d1000	0.989	0.989	exponential	200	Friedman_mse	0.5
c300_d5000	0.999	0.999	exponential	200	Friedman_mse	0.5
c500_d100	0.91	0.913	Exponential	100	Friedman_mse	0.5
c500_d1000	0.9965	0.9965	Exponential	200	Friedman_mse	0.5
c500_d5000	0.999	0.999	deviance	200	Friedman_mse	0.5
c1000_d100	0.965	0.965	exponential	100	Friedman_mse	0.5
c1000_d1000	0.996	0.996	Deviance	200	Friedman_mse	0.5
c1000_d5000	0.9997	0.9997	Deviance	100	Friedman_mse	0.5
c1500_d100	1.0	1.0	deviance	100	Friedman_mse	0.5
c1500_d1000	1.0	1.0	Deviance	200	Friedman_mse	0.5
C1500_d5000	1.0	1.0	Deviance	200	Friedman_mse	0.1
C1800_d100	0.99	0.99	Deviance	100	Friedman_mse	0.1
C1800_d1000	1.0	1.0	Deviance	100	Friedman_mse	0.1
C1800_d5000	0.999	0.999	Deviance	100	Friedman_mse	0.1

- More detailed logs of all the hyperparameter combinations are found in the OutputLogs folder.

5.

- Among the four classifiers, the one which yields the best results is evident from the tables above which is the GradientBoostingClassifier. It gives the best generalization accuracy/F1 score.

  Boosting gives the best result because it creates a strong classifier learning from many weak classifiers. It does weighted averaging over many classifiers, every time improving over previous classifier and trying to adjust for the error from the previous model.

- With the increase in the amount of training data, we can see from the above results, that the accuracy/F1 scores for all the four classifiers is increasing.

- With the increase in the number of clauses, we can see the general trend of increasing accuracy/F1 score for all the four classifiers.

6. MNIST Data:

Decision Tree-MNIST:

Accuracy	criterion	splitter	max_depth
0.8838	entropy	random	500

Bagging - MNIST:

Accuracy	criterion	splitter	max_depth	max_features
0.9523	entropy	best	500	sqrt

Random Forest-MNIST:

Accuracy	max_features	n_estimators	criterion	max_depth
0. 9719	sqrt	1000	gini	100

Boosting - MNIST:

Accuracy	loss	n_estimators	criterion	learning_rate
0. 9515	deviance	50	Friedman_mse	0.5

- In the MNIST dataset, among the four classifiers, we can notice from the above results that the random forest has the best accuracy overall. It is because of the fact that it's built over a large collection of de-correlated tress and it adds randomness to the model, while growing trees. Also, it splits a node based on the best feature available among random subset, resulting in wide diversity and overall better generalization of the model.