Ishva Patel
Prof. Gogate
CS 6375.003
October 13, 2023

Project 2 Report

Question 5. Record the classification accuracy and F1 score for each dataset and classifier (recall that we have four classifiers and we are using best/tuned parameter settings for the classifier) in a table.

	Tree Classifier Accuracy & F1 Score	**Bagging Classifier Accuracy & F1 Score**	**Random Forest Accuracy & F1 Score**	**Gradient Boosting Accuracy & F1 Score**
c1000_d100	Accuracy: 0.74 F1: 0.74257	Accuracy: 0.90 F1: 0.90384	Accuracy: 0.99 F1:0.989898	Accuracy: 0.995 F1:0.99497
c1000_d1000	Accuracy: 0.7525 F1: 0.75287	Accuracy: 0.8695 F1: 0.88259	Accuracy: 0.997 F1: 0.9970	Accuracy: 0.996 F1: 0.996007
c1000_d5000	Accuracy: 0.8166 F1: 0.82082	Accuracy: 0.7779 F1: 0.81375	Accuracy: 0.9991 F1: 0.99909	Accuracy: 0.999 F1:0.999000
c1500_d100	Accuracy: 0.795 F1: 0.80751	Accuracy: 0.81 F1: 0.840336	Accuracy: 1.0 F1: 1.0	Accuracy: 1.0 F1:1.0
c1500_d1000	Accuracy: 0.886 F1: 0.88878	Accuracy: 0.9145 F1: 0.90671	Accuracy: 0.999 F1: 0.99899	Accuracy:1.0 F1:1.0
C1500_d5000	Accuracy: 0.9227 F1: 0.92341	Accuracy: 0.8732 F1: 0.8870479	Accuracy: 1.0 F1: 1.0	Accuracy:1.0 F1:1.0
c1800_d100	Accuracy: 0.92 F1: 0.9230769	Accuracy: 0.995 F1: 0.9950	Accuracy:1.0 F1:1.0	Accuracy:1.0 F1: 1.0
c1800_d1000	Accuracy:0.949 F1: 0.9500	Accuracy:0.979 F1: 0.97941	Accuracy: 1.0 F1: 1.0	Accuracy:1.0 F1:1.0
c1800_d5000	Accuracy: 0.9657 F1: 0.965982	Accuracy: 0.9281 F1:0.9328978	Accuracy:1.0 F1: 1.0	Accuracy:1.0 F1:1.0
c300_d100	Accuracy: 0.565 F1: 0.54922	Accuracy:0.0.58 F1:0.621621	Accuracy:0.865 F1: 0.87323	Accuracy:0.87 F1: 0.87254
c300_d1000	Accuracy:0.5785	Accuracy:0.6505	Accuracy:0.909	Accuracy: 0.933

	F1:0.56568	F1: 0.697271	F1: 0.909990	F1: 0.9324
c300_d5000	Accuracy: 0.6622 F1: 0.67469183	Accuracy:0.5885 F1: 0.62251	Accuracy: 0.9353 F1: 0.9369	Accuracy:0.9664 F1:0.9667
c500_d100	Accuracy: 0.585 F1:0.59512	Accuracy: 0.685 F1: 0.71493	Accuracy: 0.935 F1: 0.93532	Accuracy:0.925 F1: 0.924
c500_d1000	Accuracy: 0.639 F1: 0.6443	Accuracy: 0.6785 F1: 0.7367990	Accuracy: 0.9715 F1: 0.91719	Accuracy: 0.9735 F1:0.973
c500_d5000	Accuracy:0.6845 F1:0.69431	Accuracy:0.6812 F1: 0.7452046	Accuracy:0.917 F1:0.9720	Accuracy: 0.9879 F1:0.987695

Question 5 part 1) Which classifier (among the four) yields the best overall generalization accuracy/F1 score? Based on your ML knowledge, why do you think the "classifier" achieved the highest overall accuracy/F1 score

The classifier that had the best overall classifier is the gradient boosting classifier. This is because, for a majority of the datasets, the gradient boosting classifier had a high F1 Score and accuracy. Gradient boosting is a good classifier for the data because it is able to handle different data by being an ensemble learner. By using various different types of weak learners it is able to increase the accuracy of the data by reducing bias of the model. In addition to that the loss function that is utilized helps with the model's performance. In addition to that, there is the learning rate that is utilized by the model. This helps with reducing overfitting the model to the training data.

Question 5 part 2) What is the impact of increasing the amount of training data on the accuracy/F1 score of each of the four classifiers

For the Single Decision Tree classifier, as we increased the amount of training data the accuracy also went up, this is seen when the amount of clauses remains the same, we can see that for the accuracy increases. Similarly, the F1 score of the decision tree classifier is also going up as the amount of data is available.
For the Bagging classifier, the accuracy of values peaks at certain values, and then it begins to decrease. This indicates that there is an ideal amount of data, that would give the bagging classifier have an optimal size of training data. Similarly, the F1 score follows the same pattern, as the accuracy, which means it peaks at a certain size of data and then decreases again.
For the Random Forest Classifier, as the amount of training data increases so does the accuracy. Looking at a place where the number of features remains constant we can see that there is an increase in the accuracy and the F1 score for the most part. Of course, there are some exceptions, but for the most part, the general pattern is that the accuracy and F1 score increases.
For the Gradient boosting classifier, as the amount of training data increases so does the accuracy and the f1.

Question 5 part 3) What is the impact of increasing the number of features on the accuracy/F1 score of each of the four classifiers

For the decision tree classifier, as the number of features increases the accuracy also increases. The number of features allows for more learning to be done. Which increases the accuracy of the classifier. The accuracy increases significantly as the number of features increases.
The bagging classifier is similar to the decision tree classifier, the accuracy increases as the number of features increases. The impact of increasing the number of features on the accuracy and F1 score is that it makes it more accurate.
In the random forest classifier, as the number of features increases so does the accuracy and the F1 score of the classifier. When looking at the table where the number of training data is the same and the number of features increases, the accuracy for the most part is increasing.
In the gradient boosting classifier, as the number of features increases so does the accuracy and F1 scores increase.

Question 7. Evaluate the four tree and ensemble classifiers you used above on the MNIST dataset (do not compute F1 score on MNIST, just classification accuracy). Which classifier among the four yields the best classification accuracy on the MNIST dataset and why?

The accuracy scores of the four classifiers were:

	Tree Classifier	**Bagging Classifier**	**Random Forest**	**Gradient Boosting**
Accuracy	0.8772	0.9391	0.9704	0.9458

The highest accuracy on the MNIST dataset was from random Forests with an accuracy of 0.9704. The reason that the accuracy was the highest compared to the rest of the classifiers was probably because of the way the classification was being made. Since there are multiple trees it minimizes the bias that might occur from a single decision tree. There is a depreciation in the bias because the random forest will average the values of trees to find a value that works for the classification, this process helps reduce the bias, making the values more accurate. In addition to that, the forest of decision trees also makes it so data isn't overfitted, which makes the testing set more accurate. The whole process of being able to average/find the majority across multiple trees helps classify the test data more accurately.