Assignment 3

Complete the assignment individually. If you worked in a group for assignment 1, you can use the same dataset that you worked with. But this time, you have to work alone.

How to submit

1. Create a folder, named like this: “lastname_firstname”
2. In the folder, include the report. The report has to be a single pdf file. I will not accept any other file format. The naming convention is like this: firstname_lastname.pdf. Anyone who does not abide by this convention rule will get a penalty of 10 points.
3. Include JUST the coding files (just the coding file and absolutely nothing else (that means .py or ipynb). Any file/folder extraneous and 10 points will be the penalty for that.
4. Penalty for late submission: Every day 10 points.
5. **Deadline: Wednesday March 30, 11:59 PM**

Tasks

1. Use the following classifiers to get accuracy, precision, recall and f1-score of your classifier on your data. **Make sure to use 5 or 10 fold cross validation**
    a. Use AdaBoost (18)
    b. Use RandomForest (18)
    c. Use NaiveBayes (18)
    d. Use BaggingClassifier (18)
    e. Use Decision Tree Classifier(12)
2. For each of the classifiers, do the following
    a. *Use top 3 features* selected by the two feature selection technique from your assignment 2, **with scaling** (use z-score scaling)
    b. *Use top 3 features* selection by the two feature selection technique from your assignment 2, **without scaling**
    c. *Use no* feature selection techniques (i.e., use all features) **with scaling**
    d. *Use no* feature selection techniques (i.e. use all features) **without scaling**
    e. **<u>Compare the performances</u>**. Which feature selection, which scaler and which classifier gives you the best performance? (6)
    f. Which performance metric you think is the most important in your use case (precision, recall, accuracy?) Why? (10)

3. For each of the classifiers, check the following links
    a. AdaBoost
        i. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
        ii. Try to tweak the parameter n_estimators (values of 50, 100 and 150) and see if you can get the performance to improve
    b. RandomForest
        i. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
        ii. Try to tweak the parameter n_estimators (values of 100, 50 and 150) and see if you can get the performance to improve
    c. BaggingClassifier
        i. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html
        ii. Do the same with parameters n_estimators (values of 10, 20, 30)
    d. Decision Tree
        i. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
        ii. Do the same for the parameter "criterion" (gini and entropy)
    e. Naïve Bayes
        i. https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB

4. Include all the results in the report, compiled in a table like the following:

Name	**Scaling**	**Feature selector**	**parameter**	**Recall**	**Precision**	**Accuracy**	**F1score**
AdaBoost	z-score	PCA top 3	n_estimators = 100	70.00	90.00	70.00	0.77

Combinations

Feature Selector	Scaler
Feature selector 1 top 3	Z-score Scaling No Scaling
Feature selector 2 top 3	
All features	

Parameters	Total Experiments
AdaBoost -> n_estimators 3 values	AdaBoost: 18
RandomForest -> n_estimators 3 values	RandomForest : 18
BaggingClassifier -> n_estimators 3 values	BaggingClassifier: 18
Decision Tree -> criterion -> 2 values	DecisionTree: 12