CECS 451
Assignment 6
Total: 34 Points

General Instruction

- Submit uncompressed file(s) in the Dropbox folder via BeachBoard (Not email).

1. Using `scikit learn`, evaluate the classification accuracy of the decision tree, bagging, AdaBoost, and Random forest.
   (a) Load the *Breast cancer data* using `sklearn.datasets.load_breast_cancer`.
   (b) (2 points) Print out the names of the features (`X`) and the name of the target (`y`).
   (c) (2 points) Allocate the half of the data to *Train* (`X_train, y_train`) and the remaining half to *Test* (`X_test, y_test`).
   (d) The common goal of the classifiers is predicting target using features.
   (e) The classifiers should be trained using *Train* set and be tested using *Test* set.
   (f) Use the 'entropy' index as the criterion and fix the maximum depth of trees as 2.
   (g) (5 points) Write a program that generates a decision tree from `X_train, y_train` and predict `y_pred` from `X_test`. You can compute accuracy of the classifier by comparing `y_pred` and `y_test`. Please print out the accuracy and the confusion matrix.
   (h) (5 points) Visualize the tree using `sklearn.tree.plot_tree`. Each node of trees should include feature name.
   (i) (5 points) Similarly, write a program that generates multiple decision trees using the bagging. This method should record its prediction accuracy at `bagging_score` by varying the parameter `n_estimators`. Draw a 2D line plot whose X-axis is `n_estimators` and Y-axis `bagging_score`, and the plot should have more than 20 data points of different X-axis values.
   (j) (5 points) Similarly, write a program that generates multiple decision trees using the AdaBoost. Draw a 2D line plot whose X-axis is `n_estimators` and Y-axis `boost_score`, and the plot should have more than 20 data points of different X-axis values.
   (k) (10 points) Similarly, write a program that generates multiple decision trees using the random forest. Fix `n_estimators=100`, and draw a 2D line plot whose X-axis is `max_features` and Y-axis `forest_score`. The plot should have more than 20 data points of different pair of X-axis values.
   (l) Submit your `Assn6.ipynb` which includes all the plots.