Random Forest Documentation

Required libraries:

- Numpy
- Pandas
- Scikit-learn

How to run:

1. Git clone https://github.com/Elliott-Hendrickson/CPTS440FinalProjectWSU.git
2. Navigate to the “RandomForest” folder
3. Run “python test.py”

DecisionTreeRegressor.py -

This file contains the node and decision tree classes.

- Node class
    - Holds all the values that will be used by the tree
- DecisionTree class
    - Inputs include minSapmplesSplit and maxDepth which are both integers. Used to determine when to stop splitting
    - Build tree function
        - Inputs include a data set (2d-array) and current depth
        - Functions purpose is to build the tree
        - Returns the root node
    - getBestSplit function
        - Inputs include the feature data set, target dataset, and whole initial data set
        - Finds which feature-threshold combo has the highest variance
        - Returns the node which has the highest variance
    - varianceReduction function
        - Inputs include the target data set, and left and right datasets
        - Finds and returns the variance used by getBestSplit
    - Fit function
        - Inputs include feature data set and target dataset
        - Turns the inputs into 2d-array
        - Calls the buildTree function
    - makePrediction function
        - Recursive function to traverse through the tree to find the leaf node closest to the input node
    - Predict function
        - Inputs include a feature set
        - Returns a list of predictions
    - Print_tree function
        - Prints out the tree

RandomForest.py -

This file contains the RandomForest class.

- RandomForest class
    - Initiated with nTrees, maxDepth, minSapmplesSplit, nFeatures. All are ints
    - Fit function
        - Inputs include feature and target data sets
        - Builds n specified amount of trees
    - Bootstrap function
        - Inputs include feature and target data sets
        - Randomizes which features are included in a data set
    - Predict function
        - Inputs include a feature set
        - Iterates through the trees in the forest to find predictions
        - Averages the predictions from all the trees
        - Returns a list of predictions

test.py -

This file is what was used for testing and is the file that should be ran.

- First it reads in the data set
- Then it chooses the features
- Then it iterates through n number of times on a data set, building decision trees and random forests to make predictions.
- Then it finds the average RMSE for the decision tree and the random forest run n-number of times