CS 60050 MACHINE LEARING

Assignment 1

Question 1. Decision Tree Regression

Provided the data that has the compressive strength in MPa when given attributes mentioned below. The dataset provided contains 1030 instances. A total of 9 attribute breakdowns, 8 quantitative input variables, and 1 quantitative output variable.

A Decision Tree Regression model has to be built that predicts the compressive strength in MPa provided the Attributes given below.

Attribute Name	**Definition**
cement	kg in a m3 mixture
slag	kg in a m3 mixture
flyash	kg in a m3 mixture
water	kg in a m3 mixture
superplasticizer	kg in a m3 mixture
coarseaggregate	kg in a m3 mixture
fineaggregate	kg in a m3 mixture
age	in days

Results of the model built:

1. Out of 10 random splits it is found that best test accuracy and the depth of the tree is as follows:

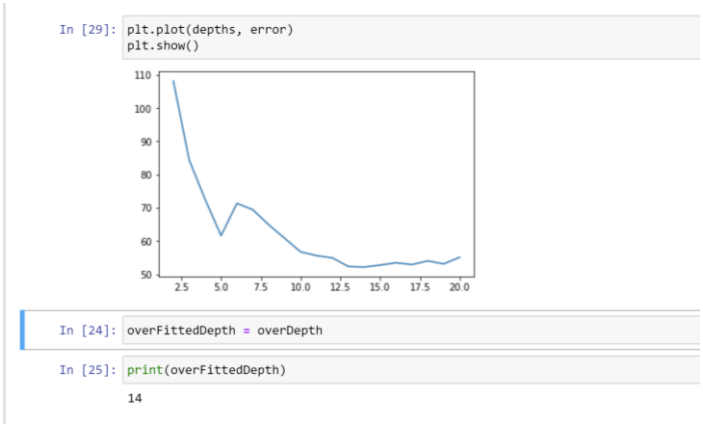
```
In [18]: import random
         best_split = 0
         error = float("inf")
         depth = 0
         for i in range(10):
             splits = random.randint(1,10)
             regressor = DecisionTreeRegressor(min_samples_split=splits, max_depth=4)
             regressor.fit(X_train,y_train)
             Y_pred = regressor.predict(X_test)
             error1 = error_term(Y_pred, y_test)
             if(error1 < error):
                 error = error1
                 best_split = splits
                 depth = regressor.tree_depth()

In [19]: print("Best Decision tree is with depth %d,error %d and min_split %d" %(depth,error, best_split))
         Best Decision tree is with depth 5,error 68 and min_split 10
```

2. The tree obtained after post pruning is given below:

```
age <= 14.0 ? 68.43045799075574
 left:cement <= 350.0 ? 50.51662715790667
  left:superplasticizer <= 6.7 ? 28.59450414687433
    left:age <= 7.0 ? 15.194560361374922
        left:cement <= 186.2 ? 8.904088352786303
                left:slag <= 13.6 ? 6.791272935652174
                                left:17.845
                                right:8.23913043478261
                right:fineaggregate <= 734.0 ? 6.549703083507936
                                left:20.80181818181818
                                right:14.273617021276598
        right:slag <= 0.0 ? 8.521568055555559
                left:cement <= 165.0 ? 2.2630126041666667
                                left:16.88
                                right:23.094666666666665
                right:31.995
    right:age <= 3.0 ? 36.221172660311204
        left:cement <= 214.9 ? 15.073957718441552
                left:fineaggregate <= 785.4 ? 2.1118238751147844
                                left:16.70777777777778
                                right:12.940000000000001
                right:slag <= 0.0 ? 9.634048001700679
                                left:21.15625
                                right:27.42833333333333
        right:cement <= 238.1 ? 26.046264535010593
                left:fineaggregate <= 792.7 ? 9.101534765625
                                left:31.23375
                                right:25.200000000000003
                right:fineaggregate <= 800.9 ? 26.662041322314053
                                left:44.260000000000005
                                right:33.89000000000001
  right:water <= 181.1 ? 41.3903139117633
    left:age <= 3.0 ? 48.74406619188595
        left:flyash <= 0.0 ? 14.135314727608495
                left:coarseaggregate <= 944.7 ? 7.658200000000001
                                left:33.699999999999996
                                right:41.370000000000005
                right:26.79166666666666
        right:cement <= 446.0 ? 21.889293888888908
                left:index <= 109.0 ? 24.84273812003969
                                left:49.34285714285714
                                right:39.29555555555556
                right:54.676
    right:fineaggregate <= 699.0 ? 46.83105360291225
        left:36.23714285714286
        right:age <= 1.0 ? 16.665388698412695
                left:6.27
                right:coarseaggregate <= 1047.8 ? 15.458442942176875
                                left:24.24083333333333
                                right:13.004999999999999
 right:cement <= 355.9 ? 74.44783914583451
  left:index <= 519.0 ? 37.0490230224063
    left:cement <= 251.8 ? 31.045416781648115
        left:superplasticizer <= 6.4 ? 19.187001684358663
                left:age <= 56.0 ? 21.422019230769237
                                left:31.900384615384613
                                right:41.209999999999994
                right:index <= 322.0 ? 20.707144500570564
```

```
                                left:40.2565625
                                right:49.36176470588235
        right:fineaggregate <= 594.0 ? 31.01197954199995
                left:38.368
                right:slag <= 114.0 ? 35.37038077661258
                                left:52.18607142857143
                                right:65.692
    right:cement <= 255.5 ? 31.52585035938496
        left:slag <= 86.0 ? 40.56699188328312
                left:cement <= 164.2 ? 11.718893830996226
                                left:13.346315789473682
                                right:20.27285714285714
                right:slag <= 184.0 ? 11.869917216845217
                                left:27.970714285714283
                                right:35.009729729729735
        right:slag <= 100.5 ? 36.256986778997046
                left:cement <= 313.0 ? 10.500926657793602
                                left:29.725806451612907
                                right:36.209999999999994
                right:cement <= 316.1 ? 22.143319273334882
                                left:43.455609756097566
                                right:57.556666666666665
right:water <= 181.1 ? 68.68439655712876
    left:slag <= 151.2 ? 30.177356995635733
        left:coarseaggregate <= 801.0 ? 16.016311847633148
                left:41.37
                right:cement <= 446.0 ? 15.126662303101597
                                left:59.38909090909091
                                right:67.59941176470588
        right:slag <= 189.0 ? 25.92302550667325
                left:index <= 137.0 ? 11.479732426303853
                                left:71.94000000000001
                                right:79.0111111111111
                right:63.13
    right:coarseaggregate <= 884.9 ? 58.90481942009161
        left:65.17
        right:cement <= 500.1 ? 29.131363039430532
                left:age <= 56.0 ? 13.221779325259519
                                left:38.94705882352941
                                right:46.21941176470588
                right:62.35666666666666
```

The plot for depth vs error is given below:



We can see that with increasing depth the accuracy increases or the error decreases. From the above plot, it is seen that the depth at which the model overfits is 14.