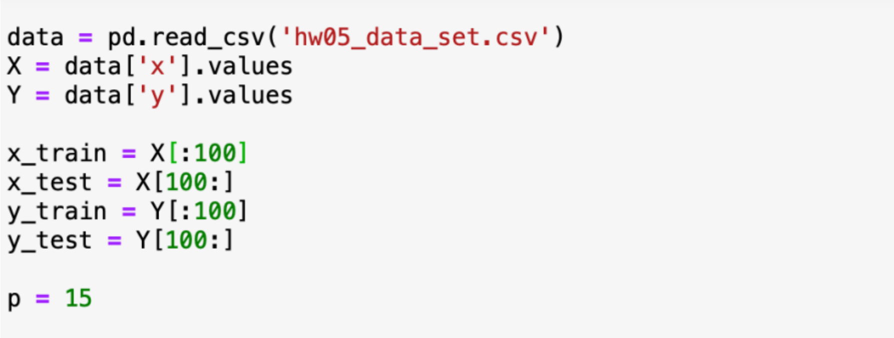
ENGR421: Introduction to Machine Learning
Fall 2020 – Homework 5
Ege Yelken – 61742

The aim of this assignment was to implement a tree regression algorithm on a univariate data set. I have followed the steps below:

1. Divided the data set into two by assigning the first 100 data points to the training set and the rest to the test set and set the pre-pruning parameter p as 15.

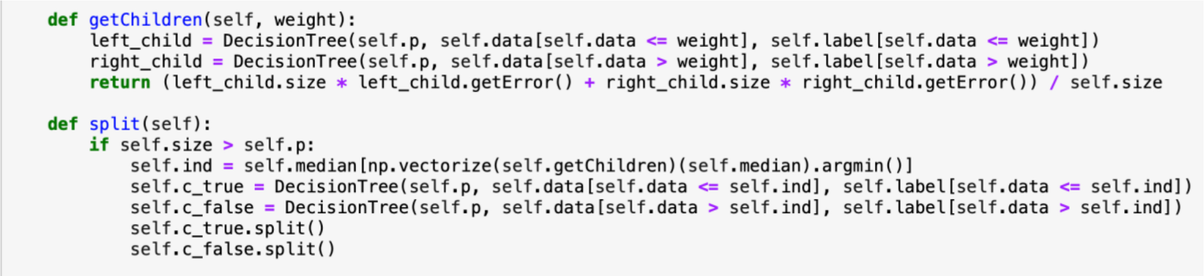


```
data = pd.read_csv('hw05_data_set.csv')
X = data['x'].values
Y = data['y'].values

x_train = X[:100]
x_test = X[100:]
y_train = Y[:100]
y_test = Y[100:]

p = 15
```

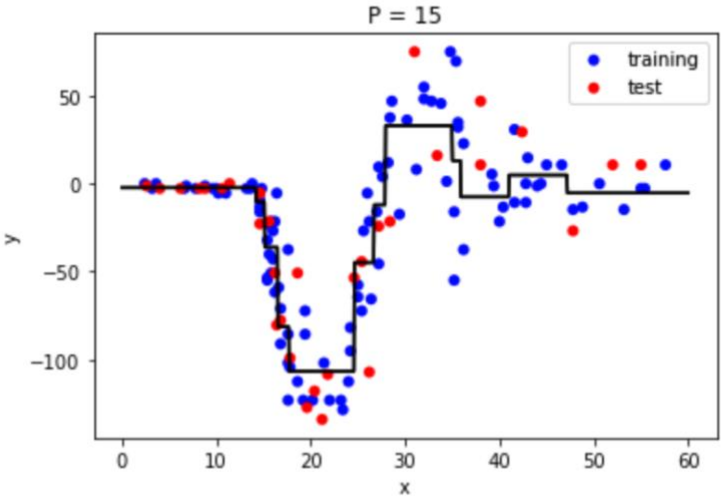
2. For training and testing, I have implemented the decision tree that splits the data by assigning the points that are smaller than or equal to the corresponding weight to the left child and the rest to the right child for each node.

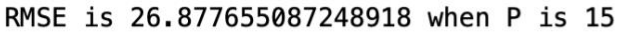


```
    def getChildren(self, weight):
        left_child = DecisionTree(self.p, self.data[self.data <= weight], self.label[self.data <= weight])
        right_child = DecisionTree(self.p, self.data[self.data > weight], self.label[self.data > weight])
        return (left_child.size * left_child.getError() + right_child.size * right_child.getError()) / self.size

    def split(self):
        if self.size > self.p:
            self.ind = self.median[np.vectorize(self.getChildren)(self.median).argmin()]
            self.c_true = DecisionTree(self.p, self.data[self.data <= self.ind], self.label[self.data <= self.ind])
            self.c_false = DecisionTree(self.p, self.data[self.data > self.ind], self.label[self.data > self.ind])
            self.c_true.split()
            self.c_false.split()
```

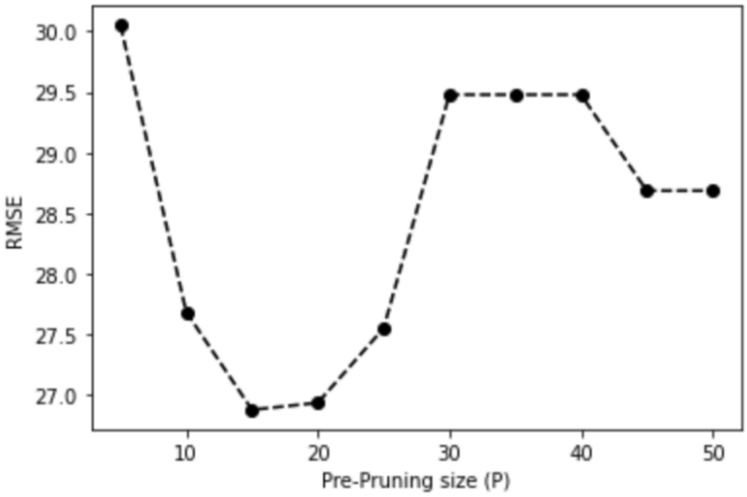
3. I trained the regression tree with the pruning parameter as 15 and calculated the RMSE.





```
RMSE is 26.877655087248918 when P is 15
```

4. I trained the tree again with the p as {5, 10, 15 … 55} and plotted with the corresponding RMSE values.


RMSE
Pre-Pruning size (P)