PRML LAB 2

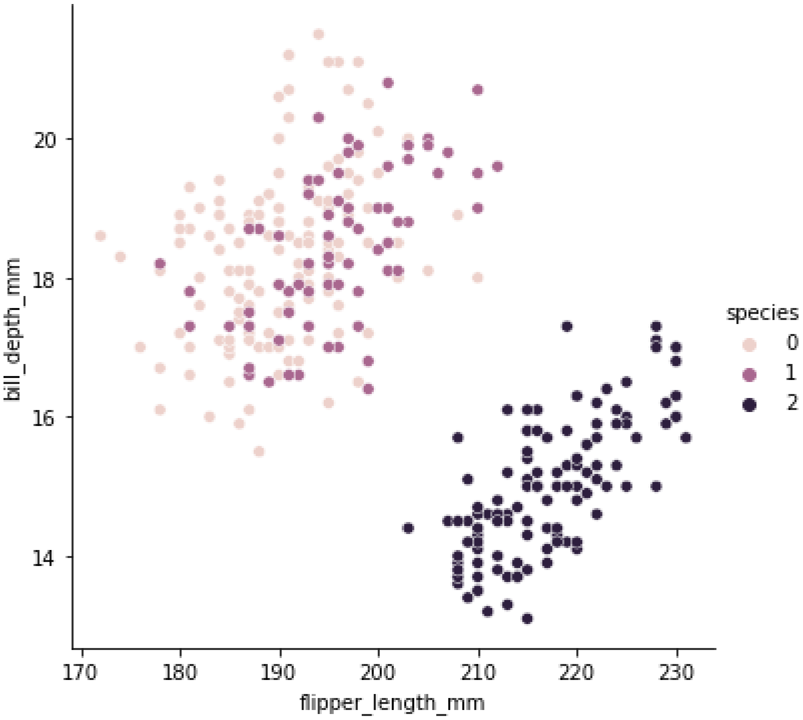
Harsh Sharma (B20CS017)

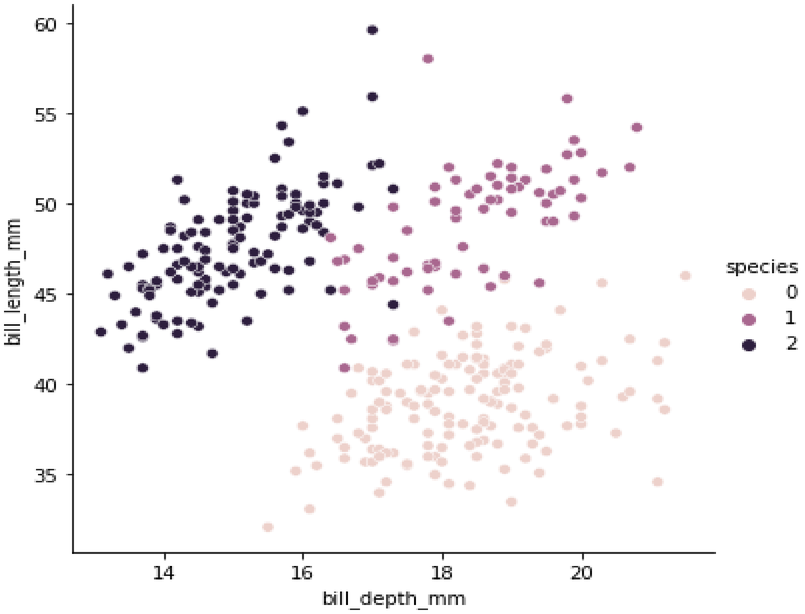
Question 1

a) Preprocessing and data visualization

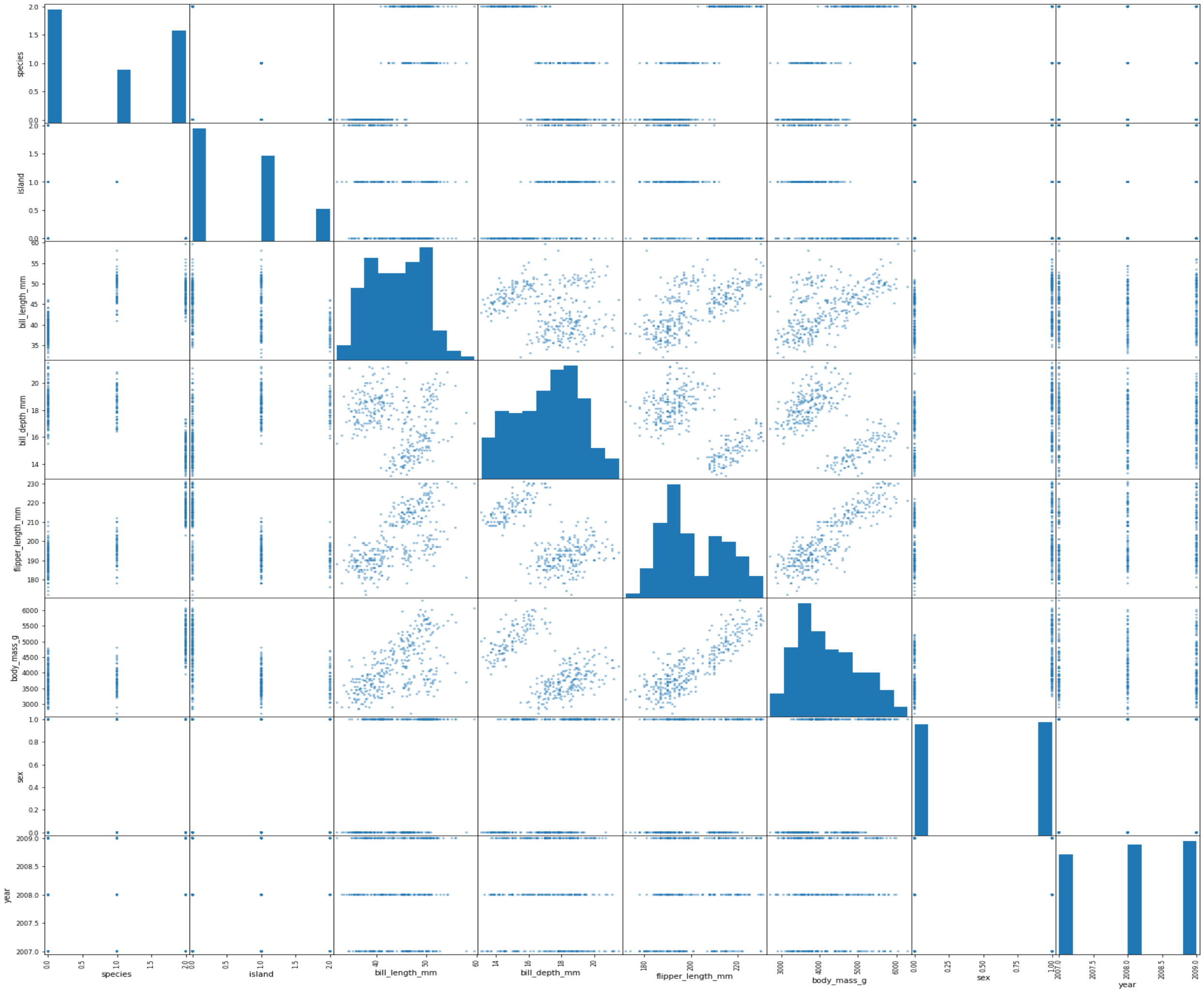
- In preprocessing , all the rows containing nan values have been removed .
- For categorical encoding features having dtype equal to object are encoded using label encoder .
- For data visualization I have used a scatter_matrix from pandas as well as seaborn's pair_plot function which plots the graph between all the combinations of two features .

Seaborn's pair_plot





Plots from scatter_matrix



b) Cost Function

- Cost function is calculated by using the gini index which is calculated by subtracting the sum of the squared probabilities of each class from one .
- It favors larger partitions.
- For a perfectly classified gini index would be zero .
- Information gain is calculated by subtracting the weighted sum of the gini index of the children after splitting from the parent gini.

c) Conversion of continuous variables to categorical

- In the con_to_cat() function all continuous variables are changed to categorical .
- For conversion every continuous feature is traversed and a threshold value is chosen for which information gain is maximum .
- After choosing the threshold value , the value greater than it in the feature is converted to 0 and less than equal to it converted to 1.

d) Best_Splitting

- In the best_split() function the dataset is splitted into children that is left and right subtree and parent node .
- A threshold value is calculated on the basis of maximum information gain .
- On the basis threshold value left and right subtree are created.

e) Training

- This is recursively done by calling best_split() function on a training dataset of size 80 percent of actual dataset until classification is completed, that is the gini index becomes zero or maximum depth is reached or min samples split is reached to create a decision tree .
- When max depth and min samples reach a certain value or there is no information gain, the tree stops growing and the node becomes a leaf node, with the value of the leaf node determined by voting.

f) Classification (i.e at test time)

- The trained model uses a predict function to predict the output of testing data .
- The feature value is compared to the threshold value of root and if it is less than that then it would traverse left tree else right tree .
- The traversal is done recursively until leaf node comes .
- After traversal classification is completed and all the predicted values are stored in an array .

g) Accuracy

- Predicted values are compared to the actual output of testing data to obtain accuracy .
- For class -wise accuracy the predicted value of each class is compared to the actual one's .
- The overall accuracy of the decision tree is 97.01492537313433 percent and class-wise accuracy is 98.03921568627452 .

Question 2

a) Preprocessing and splitting of dataset

- In preprocessing , all the rows containing nan values have been removed .

- The dataset is splitted into a 70:10:20 ratio which represents ratio of training ,validation and testing data respectively .

b) Hyper Parameter Tuning

For hyper parameter tuning four parameters are varied .These are the following features:

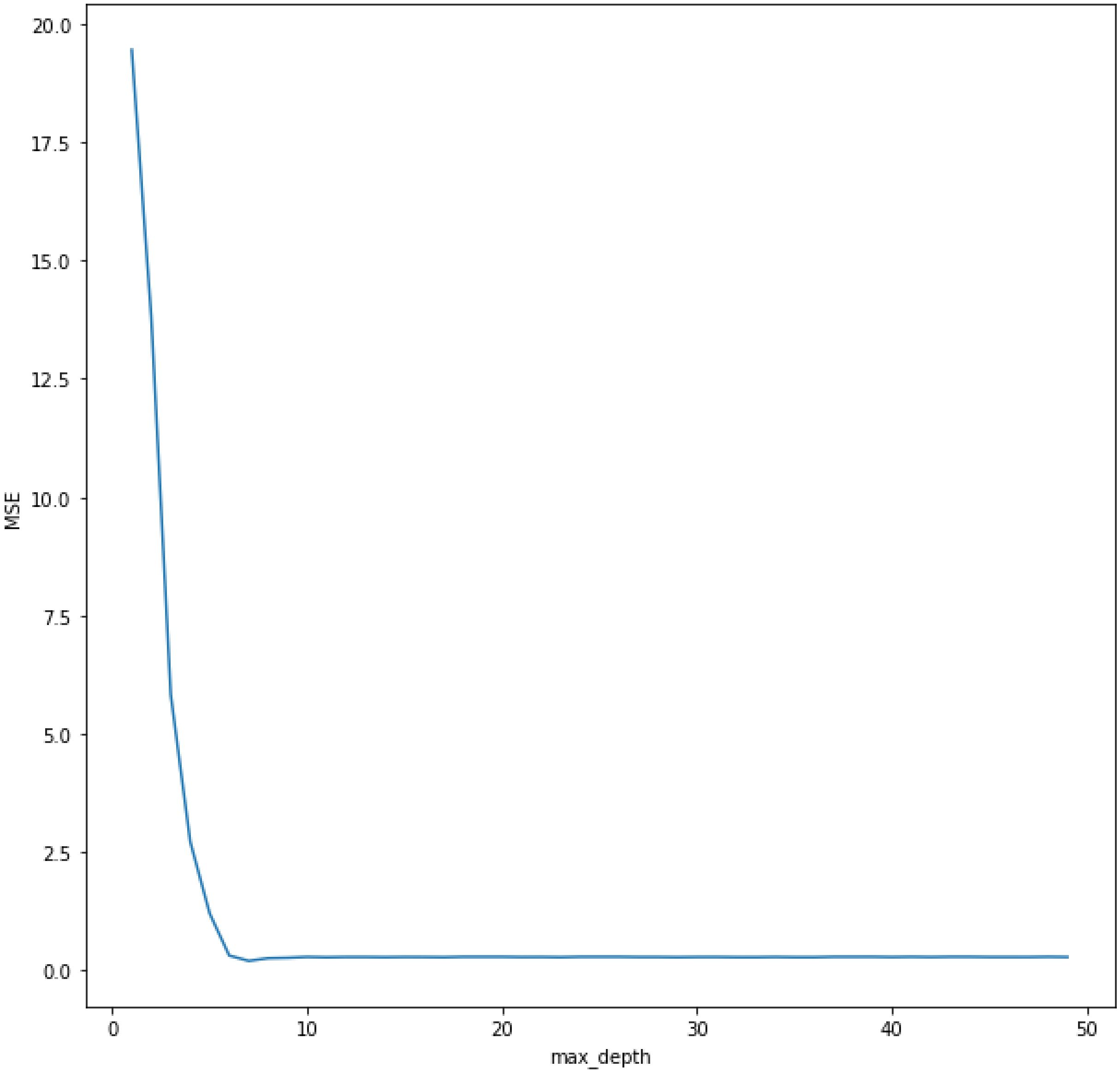
1. MAX_DEPTH

2. MAX_LEAF_NODES

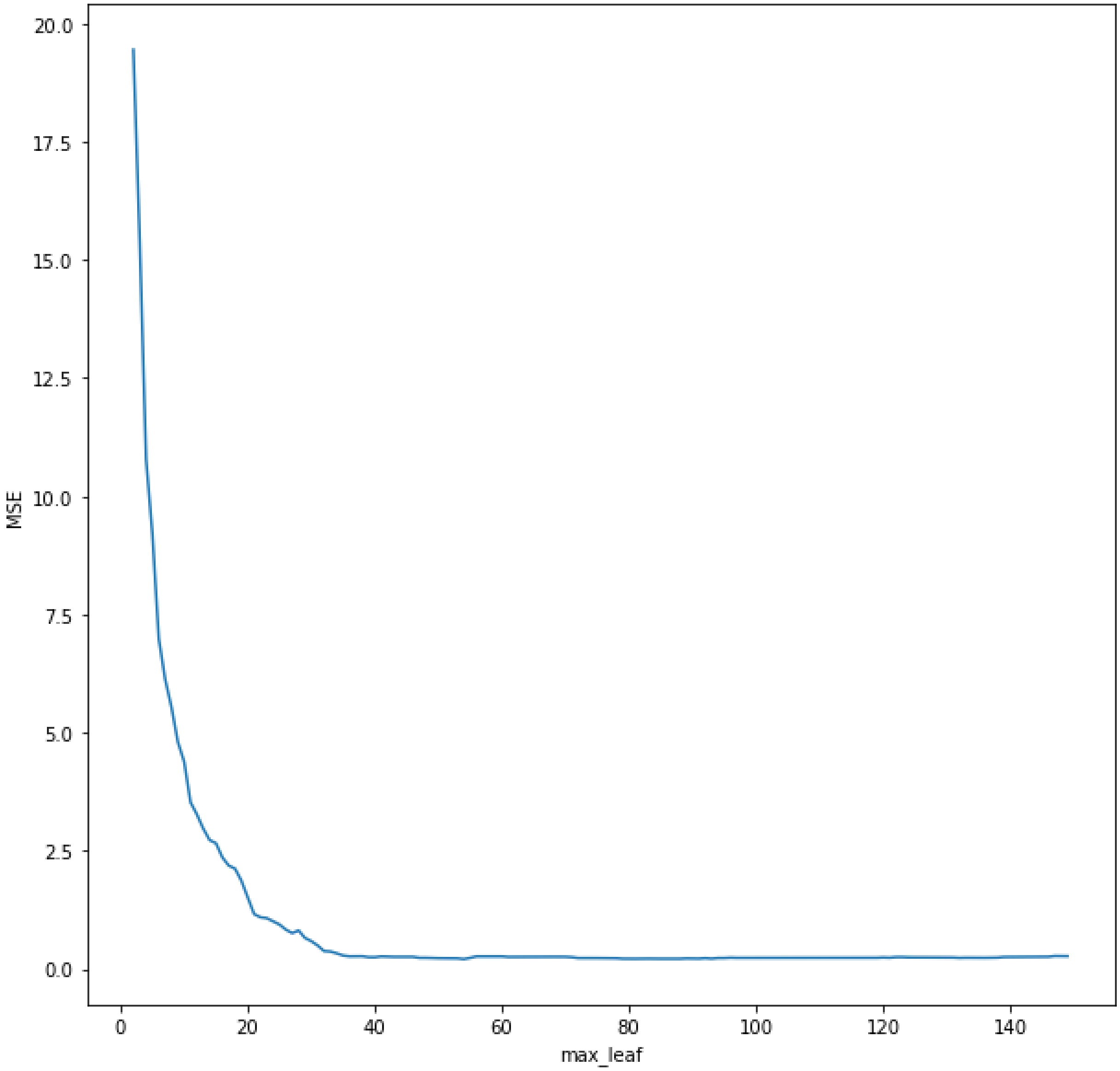
3. MIN_SAMPLES_SPLIT

4. MIN_SAMPLES_LEAF

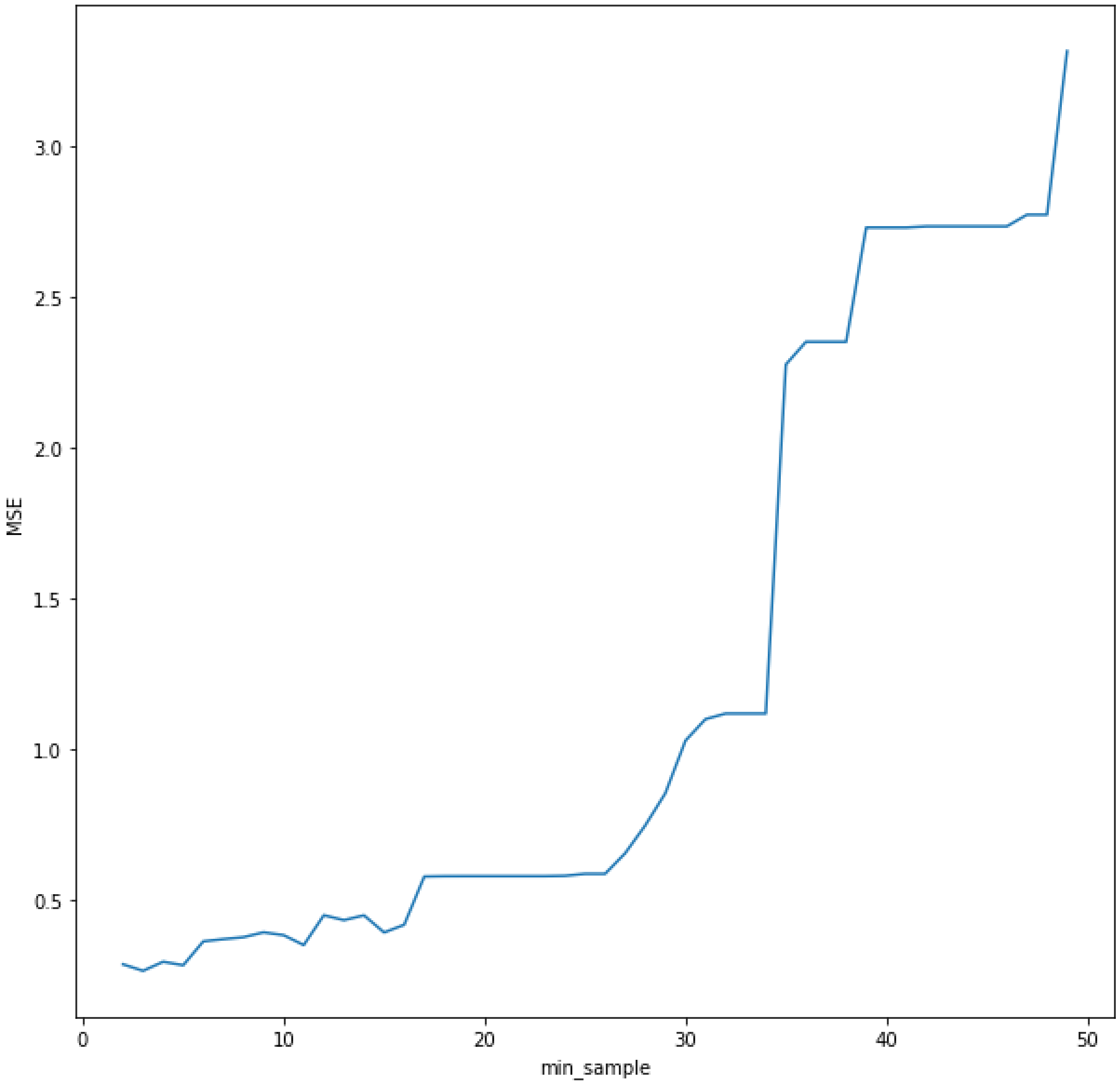
1) MAX_DEPTH :-The maximum depth of the tree .If the tree's max depth is none, the tree will grow until all of the leaf nodes are pure or all of the leaves contain samples fewer than min samples split. Overfitting occurs when max depth is set to none or surpasses a certain amount, and the model's performance suffers as a result. When max depth is less than a certain amount, underfitting occurs, lowering the model's performance. As a result, the mse graph declines at first and subsequently increases, but the growth is quite sluggish.



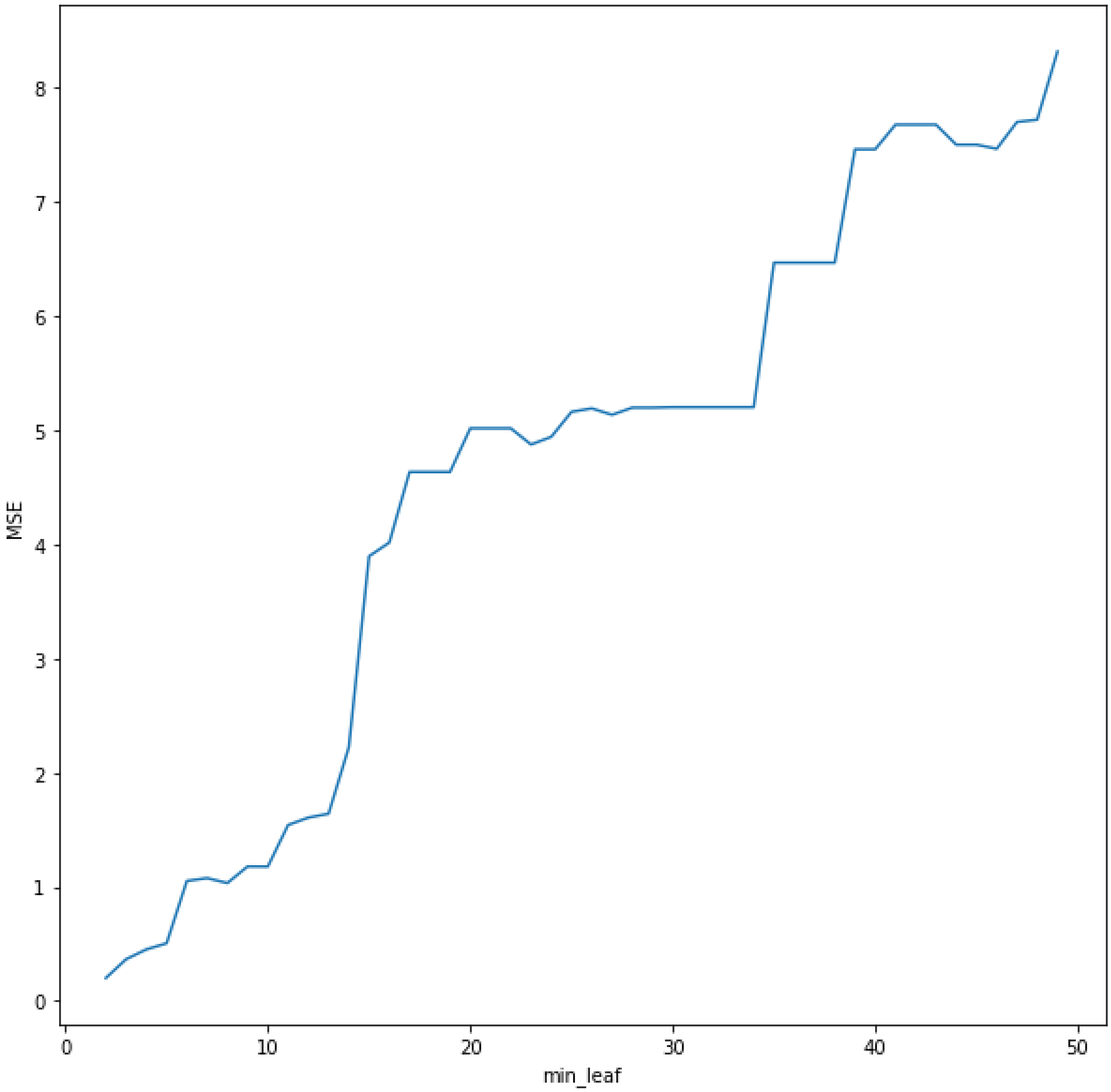
2) MAX_LEAF_NODES :- It is the number of maximum leaf nodes that a decision tree can have .As a result, if max leaf nodes is less than a certain amount, the MSE value rises due to underfitting. The tree's growth is halted early due to a lack of max leaf nodes. As a result, MSE falls as max leaf nodes increase, then becomes constant as the tree grows completely.



3) MIN_SAMPLES_SPLIT :- Minimum samples required for splitting is MIN_SAMPLES_SPLIT. The default value of min_samples_split is 2 . The performance will suffer if we increase min samples split because nodes will stop splitting at an early stage and tree growth will halt. As a result, increasing the value of min samples split will raise MSE.



4) MIN_SAMPLES_LEAF :- The minimum number of samples required to be at a leaf node.The default value of min_samples_leaf is 1. Generally min_samples_leaf is used to avoid overfitting. By increasing min_samples_leaf we can stop the growth of trees after a particular value of samples in nodes hence overfitting can be avoided. But with increase in min_samples_leaf growth of trees may stop at an early stage and underfitting occurs.



To determine the values of the above parameters in order to obtain the optimal tree, I ran four single variations of each parameter, so that when I run four nested loops simultaneously altering all four parameters, I can approximate the range of all four parameters. This will help to speed up the execution of all four nested loops while also lowering MSE. From the graphs created by running four single loops, I've picked the approximate ranges of four parameters for which the MSE is lowest. The MSE value I received after adjusting all four parameters at the same time was 0.16866309928660567.

c) 5-Fold Cross Validation and Plotting Decision Tree

5-fold cross-validation is performed using the optimal hyper-parameters decided in the above part .The data is divided into five equal folds (partitions). For training, four folds are employed, while one fold is used for testing. The operation is carried out five times. This is accomplished by utilizing the scikit learn library's built-in capabilities. Inbuilt functions were also used to calculate mean squared error. The decision tree was then plotted.

Decision Tree