Q.3: Decision Tree

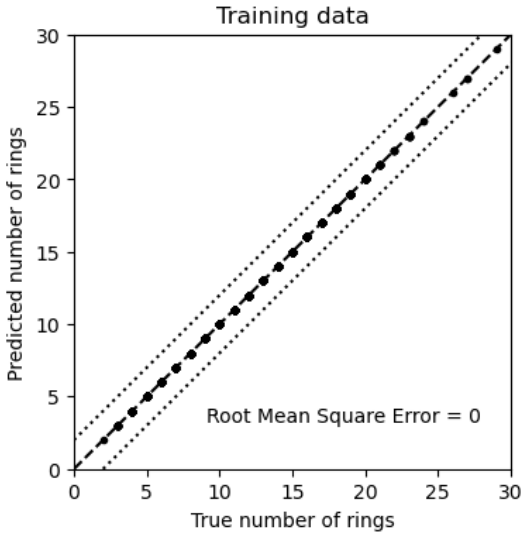
Decision Tree on Abalone dataset

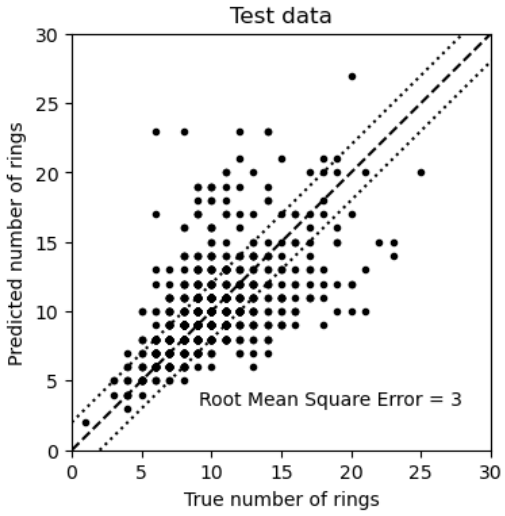
The **DecisionTreeRegressor** is an algorithm used to estimate a continuous variable instead of a discrete one.

Testing score: [-0.05001266528192927, 0.20131855871953352, 0.06866661803868834, 0.14899358230388193, 0.12614396022862273]

Training score: [1.0, 1.0, 1.0, 1.0, 1.0]

This model **overfits** the dataset and that is why, validation error is very high.





The Decision Tree overfits the training set, i.e. its parameters are fine tuned to reproduce the results of the training set but generalized badly to data not seen previously.

GridSearchCV on RAW Abalone data

Using decision tree, with max_depth of 4, we got the accuracy of 0.26238260321462337

printing the decision tree using graphviz for Raw abalone data

```
|--- Shell_weight <= 0.14
|   |--- Diameter <= 0.22
|   |   |--- Shell_weight <= 0.02
|   |   |   |--- Whole_weight <= 0.02
|   |   |   |   |--- class: 3
|   |   |   |--- Whole_weight >  0.02
|   |   |   |   |--- class: 4
|   |   |--- Shell_weight >  0.02
|   |   |   |--- Length <= 0.25
|   |   |   |   |--- class: 4
|   |   |   |--- Length >  0.25
|   |   |   |   |--- class: 5
|   |--- Diameter >  0.22
|   |   |--- Shell_weight <= 0.09
|   |   |   |--- Sex <= 1.50
|   |   |   |   |--- class: 7
|   |   |   |--- Sex >  1.50
|   |   |   |   |--- class: 9
|   |   |--- Shell_weight >  0.09
|   |   |   |--- Sex <= 0.50
|   |   |   |   |--- class: 8
|   |   |   |--- Sex >  0.50
|   |   |   |   |--- class: 7
|--- Shell_weight >  0.14
|   |--- Shell_weight <= 0.25
|   |   |--- Sucked_weight <= 0.43
|   |   |   |--- Shell_weight <= 0.19
|   |   |   |   |--- class: 8
|   |   |   |--- Shell_weight >  0.19
|   |   |   |   |--- class: 9
|   |   |--- Sucked_weight >  0.43
|   |   |   |--- Shell_weight <= 0.18
|   |   |   |   |--- class: 10
|   |   |   |--- Shell_weight >  0.18
|   |   |   |   |--- class: 9
|   |--- Shell_weight >  0.25
|   |   |--- Shell_weight <= 0.39
|   |   |   |--- Sucked_weight <= 0.44
|   |   |   |   |--- class: 10
|   |   |   |--- Sucked_weight >  0.44
|   |   |   |   |--- class: 10
|   |   |--- Shell_weight >  0.39
|   |   |   |--- Sucked_weight <= 0.61
|   |   |   |   |--- class: 10
|   |   |   |--- Sucked_weight >  0.61
|   |   |   |   |--- class: 11
```

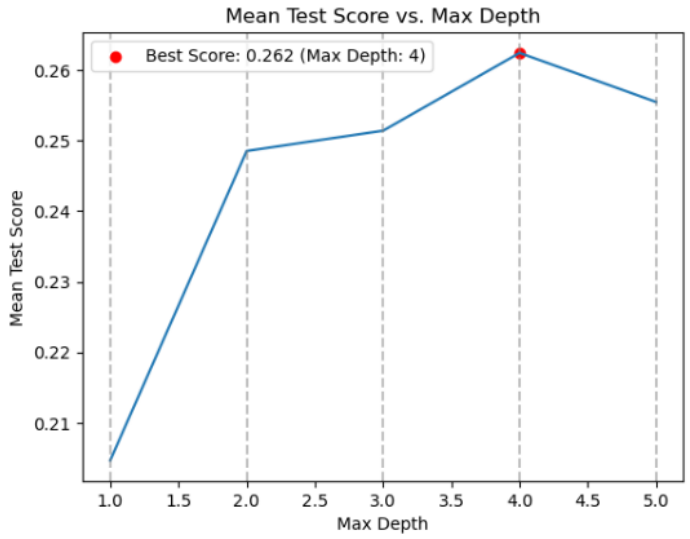
Summary of the decision tree

The decision tree for the abalone dataset has a total of 15 nodes and a maximum depth of 5. The first split is based on the "shell weight" feature, with a threshold of 0.14 g. If the shell weight is less than or equal to 0.14 g, the tree continues to split based on the "diameter" feature, with a threshold of 0.22 cm.

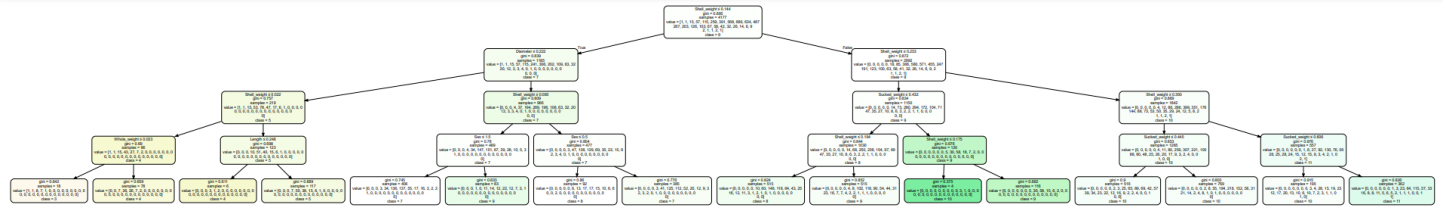
If the diameter is less than or equal to 0.22 cm, the tree further splits on "shell weight" and "whole weight" features. If the shell weight is less than or equal to 0.02 g and the whole weight is less than or equal to 0.02 g, the predicted age of the abalone is 3 years. If the shell weight is less than or equal to 0.02 g and the whole weight is greater than 0.02 g, the predicted age is 4 years. If the shell weight is greater than 0.02 g and the length is less than or equal to 0.25 cm, the predicted age is 4 years. If the length is greater than 0.25 cm, the predicted age is 5 years.

further tree continues to use the feature's threshold and classifies the abolones based on their ring classes.

Plotting the Max depth vs Mean test score for RAW abalone dataset



Plotting decision tree:



from the above graph plot, we can see that the maximum accuracy which is 26.23% is achieved when there is a max depth of 4

Tuning the hyperparameter and finding the best hyper-parameter that maximizes the accuracy for Raw abalone dataset
Best parameters: {'max_depth': 5, 'max_features': 'log2', 'max_leaf_nodes': 20, 'min_samples_leaf': 4, 'min_weight_fraction_leaf': 0.0, 'splitter': 'best'}
Best accuracy score: 0.267408818726184

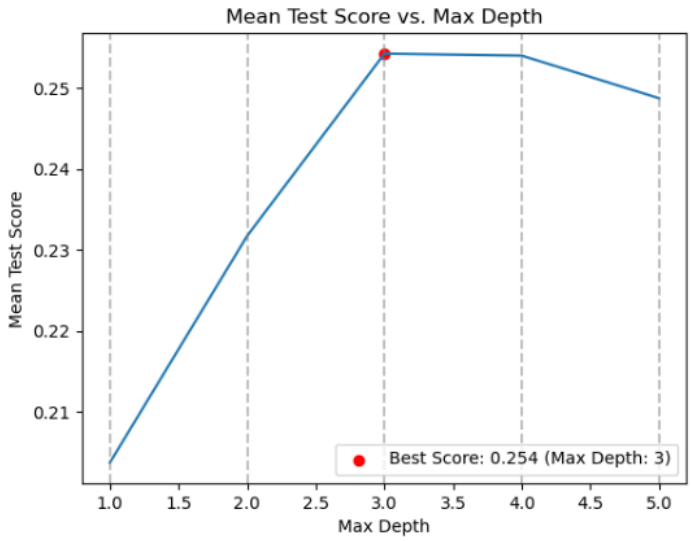
GridSearchCV on PCA pre-processed Abalone data

Best parameters: {'max_depth': 3}

Best accuracy score: 0.2542440477895883
After using 3 Principal components of PCA, we got the accuracy of 25.42% with 3 nodes. since we are using only first 3 PC of PCA, there is some loss of information. even with reduced dimensions, the accuracy is at par with the accuracy of the raw dataset.

printing the decision tree using graphviz for PCA Pre-processed abalone data and the Max depth vs Mean test score for PCA pre-processed abalone dataset

```
|--- PC2 <= -0.25
|    |--- PC3 <= 0.09
|    |    |--- PC1 <= 0.45
|    |    |    |--- class: 7
|    |    |--- PC1 >  0.45
|    |    |    |--- class: 10
|    |--- PC3 >  0.09
|    |    |--- PC2 <= -0.68
|    |    |    |--- class: 4
|    |    |--- PC2 >  -0.68
|    |    |    |--- class: 5
|--- PC2 >  -0.25
|    |--- PC2 <= 0.02
|    |    |--- PC1 <= 0.02
|    |    |    |--- class: 8
|    |    |--- PC1 >  0.02
|    |    |    |--- class: 9
|    |--- PC2 >  0.02
|    |    |--- PC2 <= 0.38
|    |    |    |--- class: 10
|    |    |--- PC2 >  0.38
|    |    |    |--- class: 11
```



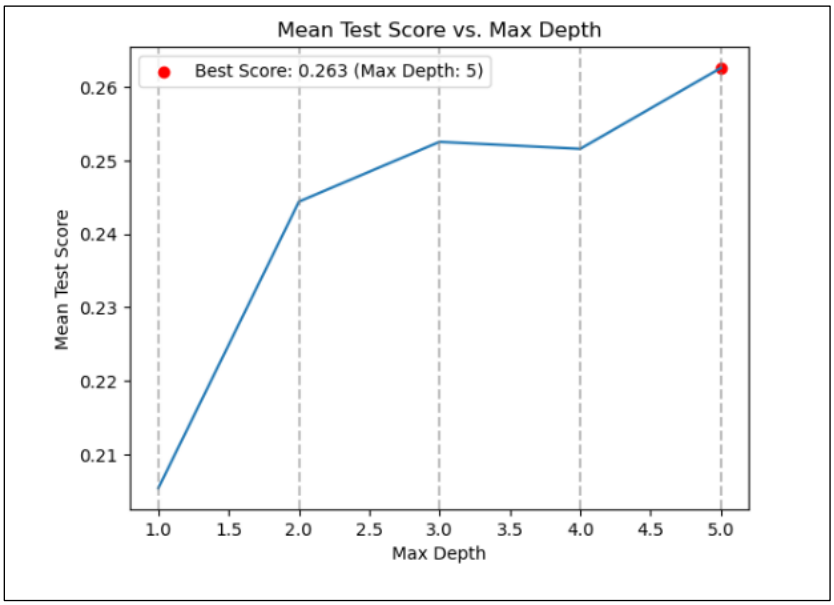
Best parameters: {'max_depth': 5, 'max_features': 'sqrt', 'max_leaf_nodes': 15, 'min_samples_leaf': 5, 'min_weight_fraction_leaf': 0.0, 'splitter': 'best'}
Best accuracy score: 0.25639773085408135

Using LDA as a preprocessing step on Abalone dataset

Best parameters: {'max_depth': 5}
Best accuracy score: 0.2626232702059995

Printing the decision tree using graphviz usind LDA preprocessed Abalone dataset and the Max depth vs Mean test score for LDA pre-processed abalone dataset

```
|--- LDA1 <= -0.60
|   |--- LDA1 <= -3.28
|   |   |--- LDA1 <= -4.13
|   |   |   |--- LDA2 <= -2.77
|   |   |   |   |--- LDA2 <= -3.89
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- LDA2 >  -3.89
|   |   |   |   |   |--- class: 3
|   |   |   |--- LDA2 >  -2.77
|   |   |   |   |--- LDA3 <= -0.49
|   |   |   |   |   |--- class: 4
|   |   |   |   |--- LDA3 >  -0.49
|   |   |   |   |   |--- class: 4
|   |   |--- LDA1 >  -4.13
|   |   |   |--- LDA3 <= -0.17
|   |   |   |   |--- LDA1 <= -3.63
|   |   |   |   |   |--- class: 5
|   |   |   |   |--- LDA1 >  -3.63
|   |   |   |   |   |--- class: 6
|   |   |   |--- LDA3 >  -0.17
|   |   |   |   |--- LDA1 <= -3.64
|   |   |   |   |   |--- class: 7
|   |   |   |   |--- LDA1 >  -3.64
|   |   |   |   |   |--- class: 5
|   |--- LDA1 >  -3.28
|   |   |--- LDA1 <= -1.45
|   |   |   |--- LDA1 <= -2.01
|   |   |   |   |--- LDA1 <= -3.25
|   |   |   |   |   |--- class: 7
|   |   |   |   |--- LDA1 >  -3.25
|   |   |   |   |   |--- class: 6
|   |   |   |--- LDA1 >  -2.01
|   |   |   |   |--- LDA2 <= -0.08
|   |   |   |   |   |--- class: 7
|   |   |   |   |--- LDA2 >  -0.08
|   |   |   |   |   |--- class: 7
|   |   |--- LDA1 >  -1.45
|   |   |   |--- LDA2 <= 0.28
|   |   |   |   |--- LDA3 <= -1.12
|   |   |   |   |   |--- class: 9
|   |   |   |   |--- LDA3 >  -1.12
|   |   |   |   |   |--- class: 8
|   |   |   |--- LDA2 >  0.28
|   |   |   |   |--- LDA3 <= -1.32
|   |   |   |   |   |--- class: 9
|   |   |   |   |--- LDA3 >  -1.32
|   |   |   |   |   |--- class: 7
|--- LDA1 >  -0.60
|   |--- LDA2 <= 0.09
|   |   |--- LDA2 <= -1.45
|   |   |   |--- LDA3 <= -0.02
|   |   |   |   |--- LDA2 <= -5.73
|   |   |   |   |   |--- class: 17
|   |   |   |   |--- LDA2 >  -5.73
|   |   |   |   |   |--- class: 11
|   |   |   |--- LDA3 >  -0.02
|   |   |   |   |--- LDA1 <= 1.20
|   |   |   |   |   |--- class: 13
|   |   |   |   |--- LDA1 >  1.20
|   |   |   |   |   |--- class: 16
|   |   |--- LDA2 >  -1.45
|   |   |   |--- LDA3 <= -0.31
|   |   |   |   |--- LDA1 <= 0.57
|   |   |   |   |   |--- class: 10
|   |   |   |   |--- LDA1 >  0.57
|   |   |   |   |   |--- class: 11
|   |   |   |--- LDA3 >  -0.31
|   |   |   |   |--- LDA2 <= -0.19
|   |   |   |   |   |--- class: 13
|   |   |   |   |--- LDA2 >  -0.19
|   |   |   |   |   |--- class: 10
|   |--- LDA2 >  0.09
|   |   |--- LDA1 <= 0.23
|   |   |   |--- LDA2 <= 0.63
|   |   |   |   |--- LDA2 <= 0.58
|   |   |   |   |   |--- class: 8
|   |   |   |   |--- LDA2 >  0.58
|   |   |   |   |   |--- class: 10
|   |   |   |--- LDA2 >  0.63
|   |   |   |   |--- LDA3 <= 0.23
|   |   |   |   |   |--- class: 9
```



Tuning the hyperparameter and finding the best hyper-parameter that maximizes the accuracy for LDA pre-processed abalone dataset

Best parameters: {'max_depth': 5,
'max_features': 'log2',
'max_leaf_nodes': 20,
'min_samples_leaf': 5,
'min_weight_fraction_leaf': 0.0,
'splitter': 'best'}

Best accuracy score: 0.2585531329685128

Implementation of Decision trees on wine dataset starts here

Using Raw Wine Data

Decision tree accuracy and best hyperparameter using raw wine data

Best parameters: {'max_depth': 5, 'max_features': 'log2', 'max_leaf_nodes': 20, 'min_samples_leaf': 4, 'min_weight_fraction_leaf': 0.0, 'splitter': 'best'}
Best accuracy score: 0.5320946290045596

Decision tree accuracy and best hyperparameter using PCA pre-processed wine data

Best parameters: {'max_depth': 5, 'max_features': 'log2', 'max_leaf_nodes': 20, 'min_samples_leaf': 1, 'min_weight_fraction_leaf': 0.0, 'splitter': 'best'}
Best accuracy score: 0.5034671640907208

Decision tree accuracy and best hyperparameter using LDA pre-processed wine data

Best parameters: {'max_depth': 5, 'max_features': 'log2', 'max_leaf_nodes': None, 'min_samples_leaf': 2, 'min_weight_fraction_leaf': 0.0, 'splitter': 'best'}
Best accuracy score: 0.5460980635992183

Conclusion:

LDA is performing better than PCA because the goal of the dataset is to classify the abalones into different age groups based on their physical characteristics. LDA takes into account the class information while PCA does not. Therefore, LDA is better suited for this classification problem.

Moreover, the abalone dataset has a low number of features compared to its sample size. This means that the dataset may not have a high degree of redundancy, which is necessary for PCA to work well. LDA is less affected by the degree of redundancy in the data because it explicitly takes into account the class information.

LDA is suited for the abalone dataset because it is specifically designed for classification problems and takes into account the class information, which is essential for the task of predicting the age of abalones based on their physical characteristics.