Out[194]: Click here to toggle on/off the raw code.

Abalone Datasets

Out[157]:

```
array([[0.51351351, 0.5210084 , 0.0840708 , ..., 0.15030262, 0.1323239 ,
        0.14798206],
       [0.37162162, 0.35294118, 0.07964602, ..., 0.06624075, 0.06319947,
        0.06826109],
       [0.61486486, 0.61344538, 0.11946903, ..., 0.17182246, 0.18564845,
        0.2077728 ],
       ...,
       [0.70945946, 0.70588235, 0.18141593, ..., 0.3527236 , 0.37788018,
        0.30543099],
       [0.74324324, 0.72268908, 0.13274336, ..., 0.35642233, 0.34298881,
        0.29347285],
       [0.85810811, 0.84033613, 0.17256637, ..., 0.63517149, 0.49506254,
        0.49177877]])
```

Out[158]:

	PC1	PC2	PC3
0	-0.230816	-0.026563	-0.006786
1	-0.497671	0.043791	0.003049
2	-0.068857	-0.081454	0.011720
3	-0.230997	-0.012962	0.004214
4	-0.532797	0.057362	-0.000513
...	...	...	...
4172	0.100632	-0.034549	-0.011468
4173	0.128141	-0.023082	-0.028686
4174	0.273938	0.019037	-0.025086
4175	0.262282	-0.027659	-0.045737
4176	0.739028	0.130322	-0.046922

4177 rows × 3 columns

Out[159]:

Rings	0	1	2
15	-0.791003	-0.235208	0.359351
7	-2.355522	0.336978	0.214024
9	0.766719	-0.246564	1.129422
10	-0.611434	0.098075	0.230542
7	-2.674301	0.527509	0.102575
...	...	...	...
11	0.921330	-0.612381	-0.272399
10	0.425796	-0.894428	-0.034727
9	1.064523	-0.385654	-0.787231
10	0.840757	-1.513723	-0.864217
12	0.843580	0.352389	-2.262564

4177 rows × 3 columns

Wine Datasets

Out[160]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcoh
0	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.
1	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.
2	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.
3	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
4	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.
...	...	...	...	...	...	...	...	...	...	...	.
95	7.1	0.260	0.29	12.4	0.044	62.0	240.0	0.9969	3.04	0.42	9.
96	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.
97	8.6	0.265	0.36	1.2	0.034	15.0	80.0	0.9913	2.95	0.36	11.
98	9.8	0.360	0.46	10.5	0.038	4.0	83.0	0.9956	2.89	0.30	10.
99	6.0	0.340	0.66	15.9	0.046	26.0	164.0	0.9979	3.14	0.50	8.

100 rows × 13 columns

Out[161]:

```
((6497, 12), (6497,))
```

Out[162]:

	PC1	PC2
0	-0.298897	-0.337622
1	-0.241913	-0.084556
2	-0.225052	-0.036821
3	-0.290807	-0.123910
4	-0.290807	-0.123910
...	...	...
6492	0.793106	0.054298
6493	0.789921	0.151012
6494	0.790787	0.111938
6495	0.808678	0.031407
6496	0.751111	0.102306

6497 rows × 2 columns

```
(6497, 2)
```

Out[163]:

	0	1
quality		
6	0.752078	-1.466209
6	1.445150	0.392049
6	-0.123015	0.911451
6	0.288961	-0.721769
6	0.288961	-0.721769
...	...	...
5	0.512278	-0.224430
6	-0.514707	-0.597340
6	-0.231160	-0.831907
5	0.630811	0.158871
6	-0.668993	-2.296580

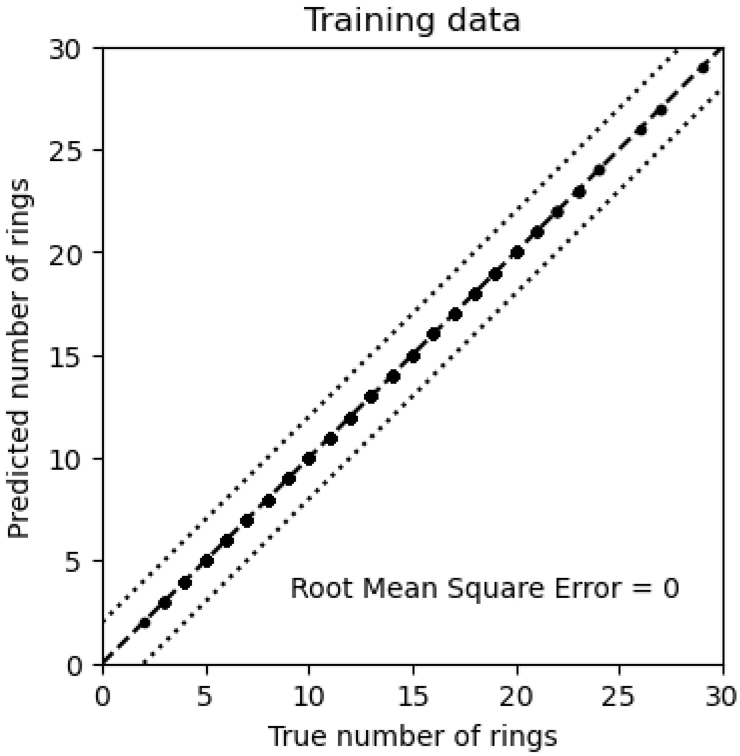
6497 rows × 2 columns

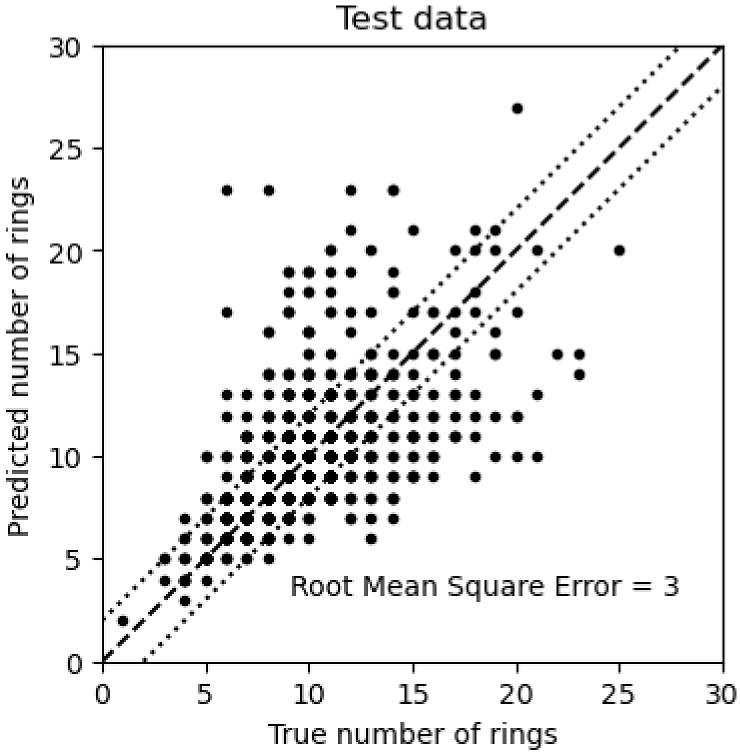
Decision Tree on Abalone dataset

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

```
Testing score:  [-0.05001266528192927, 0.20131855871953352, 0.068666618038688
34, 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.

```
Out[168]: [0.2743675926265253,
           0.32115270603998214,
           0.3433132428439525,
           0.40126800234349214,
           0.4278684960176351,
           0.4376204461182531,
           0.4583198698246772,
           0.36523602923095344,
           0.35501436701103506]
```

GridSearchCV on RAW Abalone data

```
Best parameters: {'max_depth': 4}
Best accuracy score: 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
```

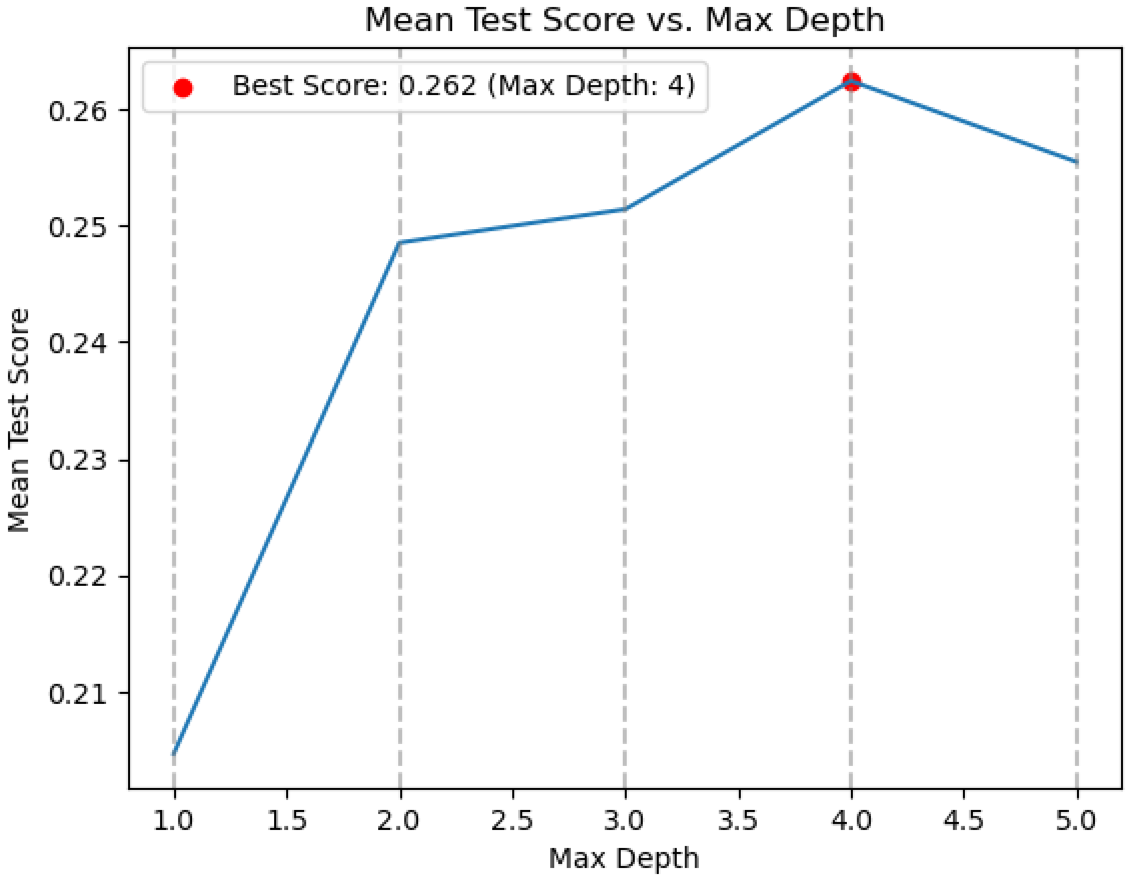
Summary of the above decision tree

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



```
Out[191]: 'Source.gv.pdf'
```

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': 2
0, 'min_samples_leaf': 4, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
Best accuracy score: 0.267408818726184
```

GridSearchCV on PCA pre-processed Abalone data

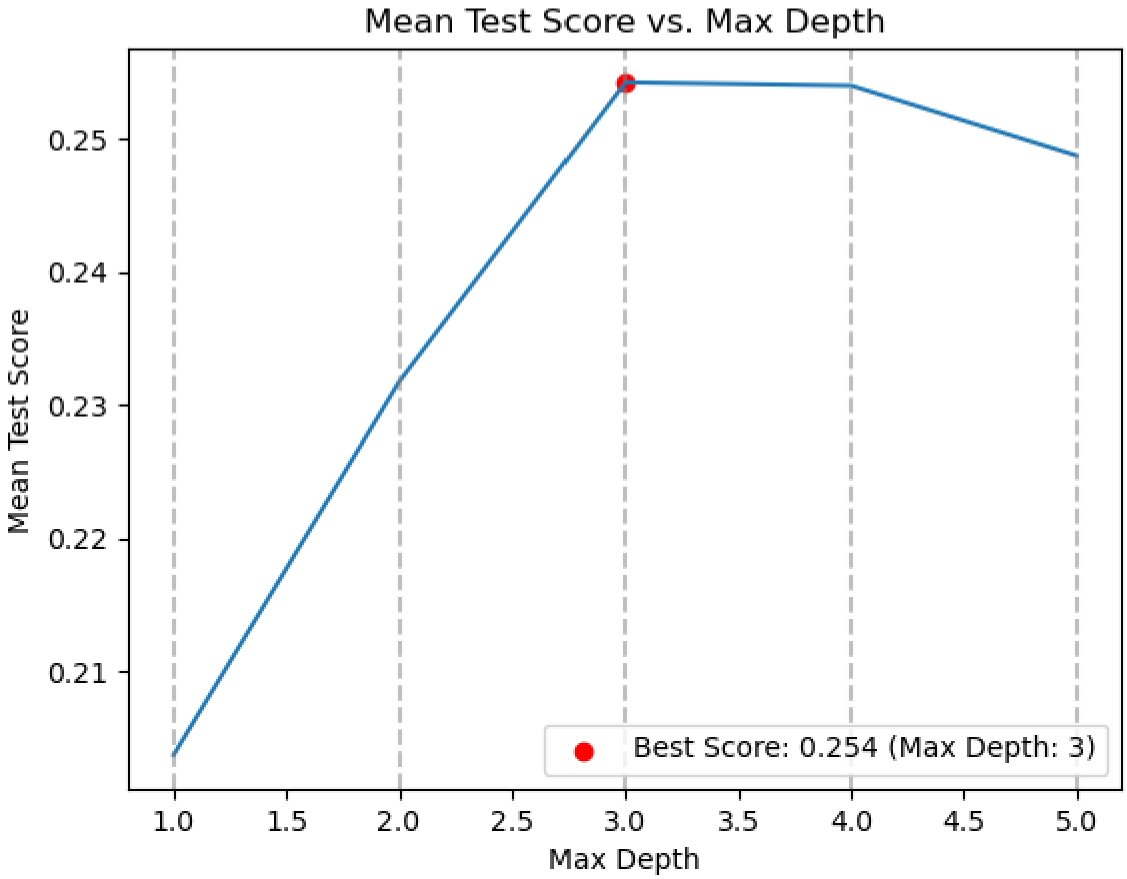
```
Best parameters: {'max_depth': 3}
Best accuracy score: 0.2542440477895883
```

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

```
|--- 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
```

Plotting the Max depth vs Mean test score for PCA pre-processed abalone dataset



```
Out[175]: 'Source.gv.pdf'
```

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

```
Best parameters: {'max_depth': 4, 'max_features': 'sqrt', 'max_leaf_nodes': 1
5, 'min_samples_leaf': 2, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
Best accuracy score: 0.25831389851875197
```

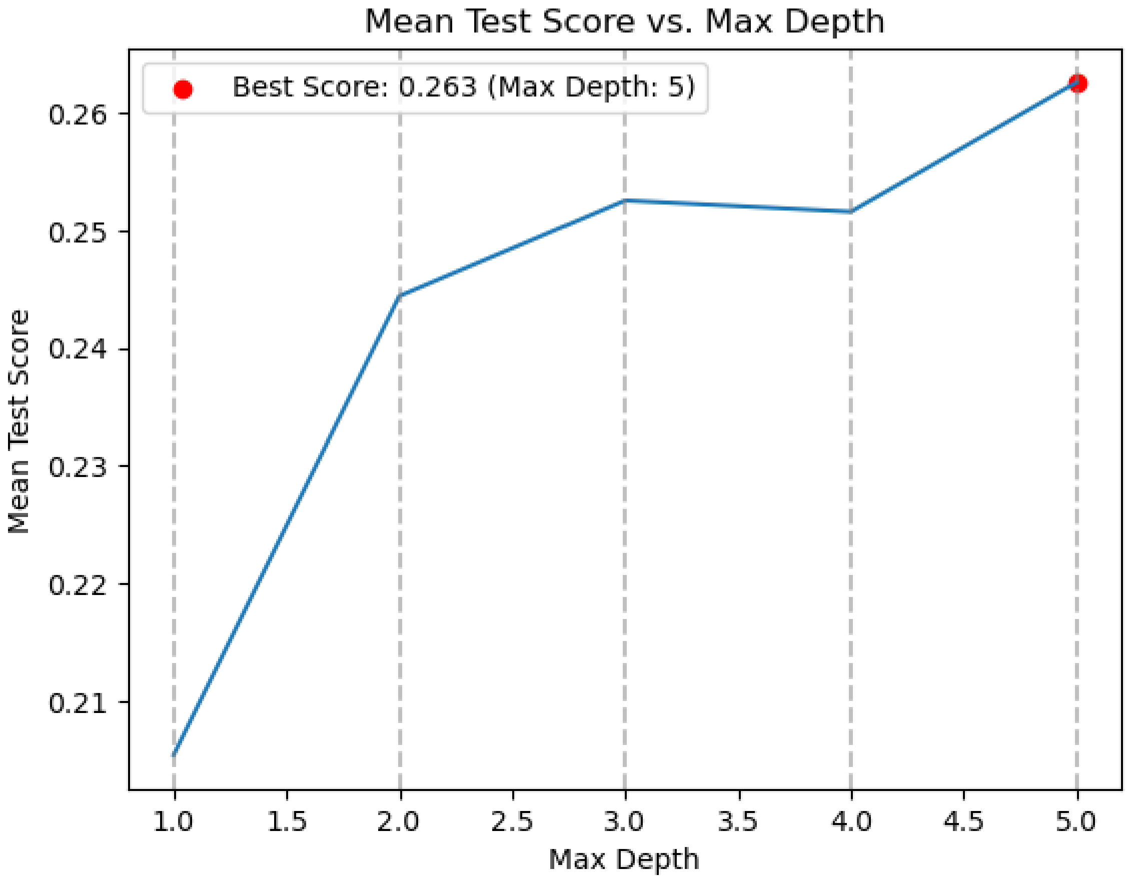
Using LDA as a preprocessing step on Abalone dataset

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

Printing the decision tree using graphviz usind LDA preprocessed Abalone dataset

```
|--- 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
|   |   |   |   |--- LDA3 >  0.23
|   |   |   |   |   |--- class: 8
|   |   |--- LDA1 >  0.23
|   |   |   |--- LDA2 <= 0.79
|   |   |   |   |--- LDA1 <= 1.30
|   |   |   |   |   |--- class: 9
|   |   |   |   |--- LDA1 >  1.30
|   |   |   |   |   |--- class: 11
|   |   |   |--- LDA2 >  0.79
|   |   |   |   |--- LDA3 <= -1.45
|   |   |   |   |   |--- class: 11
|   |   |   |   |--- LDA3 >  -1.45
|   |   |   |   |   |--- class: 9
```

Plotting the Max depth vs Mean test score for LDA pre-processed abalone dataset



```
Out[179]: 'Source.gv.pdf'
```

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': 2
0, 'min_samples_leaf': 5, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
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': 2
0, 'min_samples_leaf': 4, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
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': 2
0, 'min_samples_leaf': 1, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
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': N
one, 'min_samples_leaf': 2, 'min_weight_fraction_leaf': 0.0, 'splitter': 'bes
t'}
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.