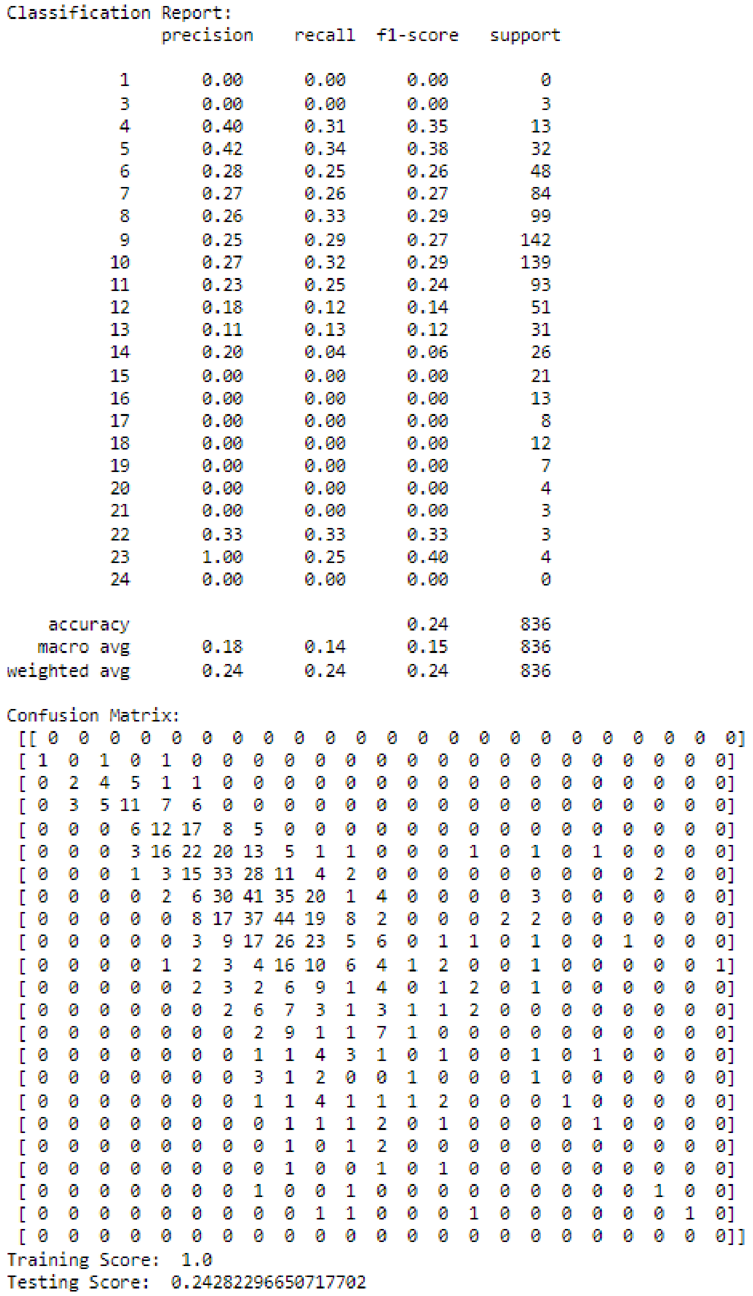
Q.5 Gradiant Boosting

Gradient Boosting on Abalone dataset

Accuracy on raw abalone dataset: 24.28

Plotting confusion matrix for abalone raw dataset



```
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         0
           3       0.00      0.00      0.00         3
           4       0.40      0.31      0.35        13
           5       0.42      0.34      0.38        32
           6       0.28      0.25      0.26        48
           7       0.27      0.26      0.27        84
           8       0.26      0.33      0.29        99
           9       0.25      0.29      0.27       142
          10       0.27      0.32      0.29       139
          11       0.23      0.25      0.24        93
          12       0.18      0.12      0.14        51
          13       0.11      0.13      0.12        31
          14       0.20      0.04      0.06        26
          15       0.00      0.00      0.00        21
          16       0.00      0.00      0.00        13
          17       0.00      0.00      0.00         8
          18       0.00      0.00      0.00        12
          19       0.00      0.00      0.00         7
          20       0.00      0.00      0.00         4
          21       0.00      0.00      0.00         3
          22       0.33      0.33      0.33         3
          23       1.00      0.25      0.40         4
          24       0.00      0.00      0.00         0

    accuracy                           0.24       836
   macro avg       0.18      0.14      0.15       836
weighted avg       0.24      0.24      0.24       836

Confusion Matrix:
 [[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  2  4  5  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  3  5 11  7  6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  6 12 17  8  5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  3 16 22 20 13  5  1  1  0  0  0  1  0  1  0  1  0  0  0  0]
 [ 0  0  0  1  3 15 33 28 11  4  2  0  0  0  0  0  0  0  0  0  2  0  0]
 [ 0  0  0  0  2  6 30 41 35 20  1  4  0  0  0  0  3  0  0  0  0  0  0]
 [ 0  0  0  0  0  8 17 37 44 19  8  2  0  0  0  2  2  0  0  0  0  0  0]
 [ 0  0  0  0  0  3  9 17 26 23  5  6  0  1  1  0  1  0  0  1  0  0  0]
 [ 0  0  0  0  1  2  3  4 16 10  6  4  1  2  0  0  1  0  0  0  0  0  1]
 [ 0  0  0  0  0  2  3  2  6  9  1  4  0  1  2  0  1  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  2  6  7  3  1  3  1  1  2  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  2  9  1  1  7  1  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  1  4  3  1  0  1  0  0  1  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  3  1  2  0  0  1  0  0  0  1  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  1  4  1  1  1  2  0  0  0  1  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  1  1  2  0  1  0  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  0  1  2  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  0  0  1  0  1  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0  0  1  0  0]
 [ 0  0  0  0  0  0  0  0  0  1  1  0  0  0  1  0  0  0  0  0  0  1  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]]
Training Score:  1.0
Testing Score:  0.24282296650717702
```

The accuracy on the abalone raw dataset using Gradient Boosting classifier is less than Random Forests when using similar parameters, possibly due to the effect of outliers. It takes longer to train with Gradient Boosting than Random Forests.

Lack of strong linear relationships: Gradient boosting relies on creating ensembles of weak learners, which are typically decision trees, to model complex non-linear relationships between the input features and the target variable. However, if the input features do not exhibit strong linear relationships with the target variable, the decision trees may not be able to capture the complex non-linear relationships in the data.

Insufficient number of features:
The Abalone dataset contains only 8 input features, which may not be sufficient to capture all the complexity in the data. If the input features do not provide enough information to accurately predict the target variable, the model may not perform well.

Few other things that can be deduced from the above confusion matrix.

Overfitting: Gradient boosting can be prone to overfitting if the hyperparameters are not tuned properly. Overfitting occurs when the model learns to fit the training data too closely, which can lead to poor generalization performance on new, unseen data. This can happen if the model is too complex relative to the size of the training data, or if the learning rate is set too high, which causes the model to over-emphasize the contribution of individual trees.

Randomness in the data: The Abalone dataset contains some randomness due to the nature of the abalone shells and the way they grow. This can make it difficult for any model to accurately predict the age of an abalone based solely on physical measurements, which may contribute to the poor performance of gradient boosting on this dataset.

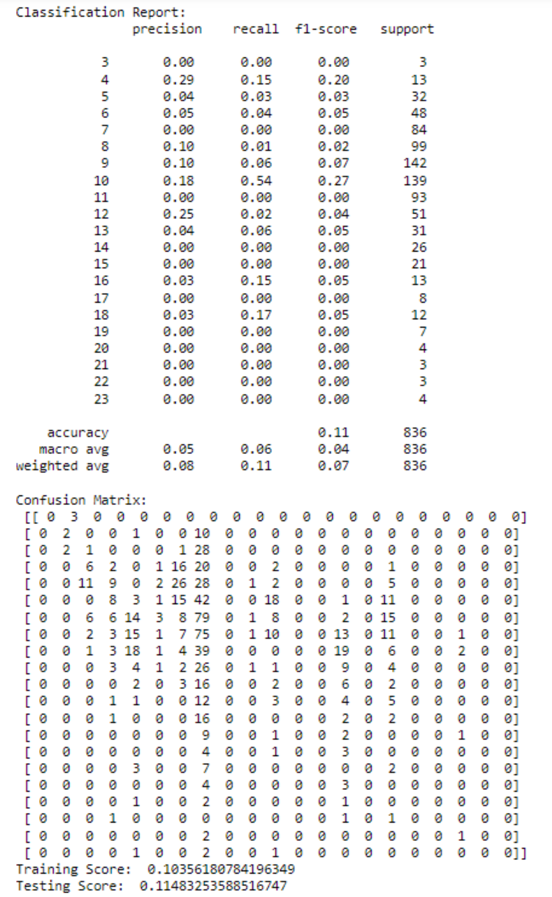
Changing the parameter n_estimator=100, lr=0.1, and max depth 3 for raw abalone dataset: Accuracy score: 0.2548
Upon using more optimum parameters for Gradient Boosting, the accuracy increases. This low accuracy may be due to the fact that the features are highly correlated.

Gradient Boosting on Wine dataset

Accuracy score on Raw wine dataset: 59.92
The accuracy of Gradient boosting on the wine - raw dataset is more than that of Random forests and this may be due to the fact that the dataset has outliers and is not balanced. When the dataset contains imbalanced classes, Random Forests may produce biased predictions towards the majority class, as each tree is built independently and can be influenced by the class imbalance, while Gradient Boosting Classifier can adjust the weights of the samples to balance the classes.

Gradient Boosting on Abalone - PCA dataset

Accuracy score: 0.1148



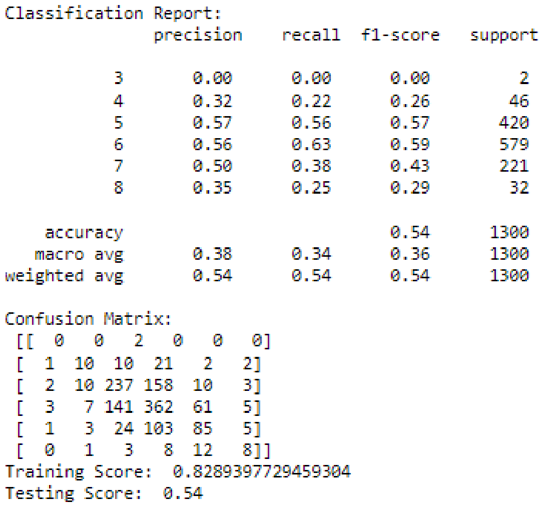
```
Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.29      0.15      0.20        13
           5       0.04      0.03      0.03        32
           6       0.05      0.04      0.05        48
           7       0.00      0.00      0.00        84
           8       0.10      0.01      0.02        99
           9       0.10      0.06      0.07       142
          10       0.18      0.54      0.27       139
          11       0.00      0.00      0.00        93
          12       0.25      0.02      0.04        51
          13       0.04      0.06      0.05        31
          14       0.00      0.00      0.00        26
          15       0.00      0.00      0.00        21
          16       0.03      0.15      0.05        13
          17       0.00      0.00      0.00         8
          18       0.03      0.17      0.05        12
          19       0.00      0.00      0.00         7
          20       0.00      0.00      0.00         4
          21       0.00      0.00      0.00         3
          22       0.00      0.00      0.00         3
          23       0.00      0.00      0.00         4

    accuracy                           0.11       836
   macro avg       0.05      0.06      0.04       836
weighted avg       0.08      0.11      0.07       836

Confusion Matrix:
 [[ 0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  2  0  0  1  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  2  1  0  0  0  1 28  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  6  2  0  1 16 20  0  0  2  0  0  0  0  1  0  0  0  0  0]
 [ 0  0 11  9  0  2 26 28  0  1  2  0  0  0  0  5  0  0  0  0  0]
 [ 0  0  0  8  3  1 15 42  0  0 18  0  0  1  0 11  0  0  0  0  0]
 [ 0  0  6  6 14  3  8 79  0  1  8  0  0  2  0 15  0  0  0  0  0]
 [ 0  0  2  3 15  1  7 75  0  1 10  0  0 13  0 11  0  0  1  0  0]
 [ 0  0  1  3 18  1  4 39  0  0  0  0  0 19  0  6  0  0  2  0  0]
 [ 0  0  0  3  4  1  2 26  0  1  1  0  0  9  0  4  0  0  0  0  0]
 [ 0  0  0  0  2  0  3 16  0  0  2  0  0  6  0  2  0  0  0  0  0]
 [ 0  0  0  1  1  0  0 12  0  0  3  0  0  4  0  5  0  0  0  0  0]
 [ 0  0  0  1  0  0  0 16  0  0  0  0  0  2  0  2  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  9  0  0  1  0  0  2  0  0  0  0  1  0  0]
 [ 0  0  0  0  0  0  0  4  0  0  1  0  0  3  0  0  0  0  0  0  0]
 [ 0  0  0  0  3  0  0  7  0  0  0  0  0  0  0  2  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  4  0  0  0  0  0  3  0  0  0  0  0  0  0]
 [ 0  0  0  0  1  0  0  2  0  0  0  0  0  1  0  0  0  0  0  0  0]
 [ 0  0  0  1  0  0  0  0  0  0  0  0  0  1  0  1  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  1  0  0]
 [ 0  0  0  0  1  0  0  2  0  0  1  0  0  0  0  0  0  0  0  0  0]]
Training Score:  0.10356180784196349
Testing Score:  0.11483253588516747
```

Note:

The accuracy on PCA dataset upon using Gradient Boosting is lesser than Random forests. Overall it can be seen that PCA hurts the performance of a tree boosting classifier as data has been lost while reducing the number of dimensions.

Gradient Boosting on Wine - PCA dataset

Accuracy score on wine dataset using PCA preprocessing: 0.5400



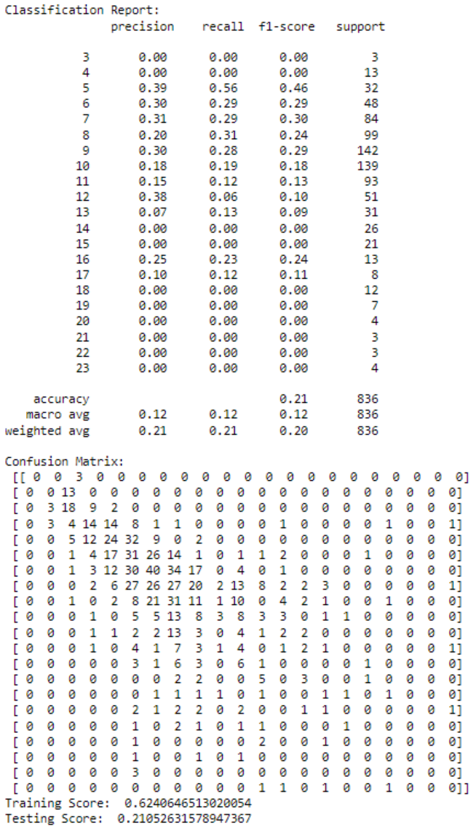
```
Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         2
           4       0.32      0.22      0.26        46
           5       0.57      0.56      0.57       420
           6       0.56      0.63      0.59       579
           7       0.50      0.38      0.43       221
           8       0.35      0.25      0.29        32

    accuracy                           0.54      1300
   macro avg       0.38      0.34      0.36      1300
weighted avg       0.54      0.54      0.54      1300

Confusion Matrix:
 [[  0   0   2   0   0   0]
 [  1  10  10  21   2   2]
 [  2  10 237 158  10   3]
 [  3   7 141 362  61   5]
 [  1   3  24 103  85   5]
 [  0   1   3   8  12   8]]
Training Score:  0.8289397729459304
Testing Score:  0.54
```

Note:

The training score is 82% whereas the same classifier has a training score of approximately 70% on raw data without PCA reduction. So in this case, PCA helps in improving the accuracy but there is a considerable amount of overfitting.

Gradient Boosting on Abalone - LDA dataset

Accuracy score: 0.2105



```
Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00        13
           5       0.39      0.56      0.46        32
           6       0.30      0.29      0.29        48
           7       0.31      0.29      0.30        84
           8       0.20      0.31      0.24        99
           9       0.30      0.28      0.29       142
          10       0.18      0.19      0.18       139
          11       0.15      0.12      0.13        93
          12       0.38      0.06      0.10        51
          13       0.07      0.13      0.09        31
          14       0.00      0.00      0.00        26
          15       0.00      0.00      0.00        21
          16       0.25      0.23      0.24        13
          17       0.10      0.12      0.11         8
          18       0.00      0.00      0.00        12
          19       0.00      0.00      0.00         7
          20       0.00      0.00      0.00         4
          21       0.00      0.00      0.00         3
          22       0.00      0.00      0.00         3
          23       0.00      0.00      0.00         4

    accuracy                           0.21       836
   macro avg       0.12      0.12      0.12       836
weighted avg       0.21      0.21      0.20       836

Confusion Matrix:
 [[ 0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  3 18  9  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  3  4 14 14  8  1  1  0  0  0  0  1  0  0  0  0  1  0  0  1]
 [ 0  0  5 12 24 32  9  0  2  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  1  4 17 31 26 14  1  0  1  1  2  0  0  0  1  0  0  0  0]
 [ 0  0  1  3 12 30 40 34 17  0  4  0  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  2  6 27 26 27 20  2 13  8  2  2  3  0  0  0  0  0  1]
 [ 0  0  1  0  2  8 21 31 11  1 10  0  4  2  1  0  0  1  0  0  0]
 [ 0  0  0  1  0  5  5 13  8  3  8  3  3  0  1  1  0  0  0  0  0]
 [ 0  0  0  1  1  2  2 13  3  0  4  1  2  2  0  0  0  0  0  0  0]
 [ 0  0  0  1  0  4  1  7  3  1  4  0  1  2  1  0  0  0  0  0  1]
 [ 0  0  0  0  0  3  1  6  3  0  6  1  0  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  2  2  0  0  5  0  3  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  1  1  1  1  0  1  0  0  1  1  0  1  0  0  0]
 [ 0  0  0  0  0  2  1  2  2  0  2  0  0  1  1  0  0  0  0  0  1]
 [ 0  0  0  0  0  1  0  2  1  0  1  1  0  0  0  1  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  0  0  0  0  2  0  0  1  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  0  1  0  1  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  1  1  0  1  0  0  1  0  0  0]]
Training Score:  0.6240646513020054
Testing Score:  0.21052631578947367
```

Note:

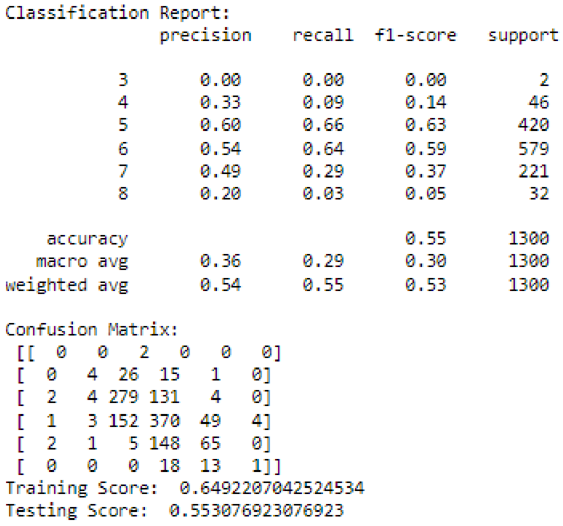
Training score is high as when the dataset has a small number of samples, Gradient boosting can overfit and since most features in the abalone dataset is highly correlated, dimensionality reduction has a positive effect on efficient computation.

But testing score is very low as there is considerable loss of data and Gradient boosting works better with more features.

The mean accuracy using Random Forests is 0.27 whereas for Gradient boosting, it is lower. This is possible if there are too many outliers/high correlation in the dataset, which is true for this case.

Gradient Boosting on Wine - LDA dataset

Accuracy score: 0.5531



```
Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         2
           4       0.33      0.09      0.14        46
           5       0.60      0.66      0.63       420
           6       0.54      0.64      0.59       579
           7       0.49      0.29      0.37       221
           8       0.20      0.03      0.05        32

    accuracy                           0.55      1300
   macro avg       0.36      0.29      0.30      1300
weighted avg       0.54      0.55      0.53      1300

Confusion Matrix:
 [[  0   0   2   0   0   0]
 [  0   4  26  15   1   0]
 [  2   4 279 131   4   0]
 [  1   3 152 370  49   4]
 [  2   1   5 148  65   0]
 [  0   0   0  18  13   1]]
Training Score:  0.6492207042524534
Testing Score:  0.553076923076923
```

Note:

There is less overfitting in the training data after using LDA and Gradient boosting techniques. The test accuracy is also close but not very high. Compared to random forests, the accuracy is similar on Wine - LDA dataset.