ECE 657A Assignment 2

Submitted By :
Amandeep Kaur (21044104)
Bhupesh Dod (21046099)

Question 3: Decision Tree Solution:

Solution:

The libraries are imported, the datasets are loaded. Combined the two wine datas into a single dataset. Performed the Exploratory Data Analysis and searched for missing values and no missing values were found. Now, apply standardization on data and split into test and training sets.

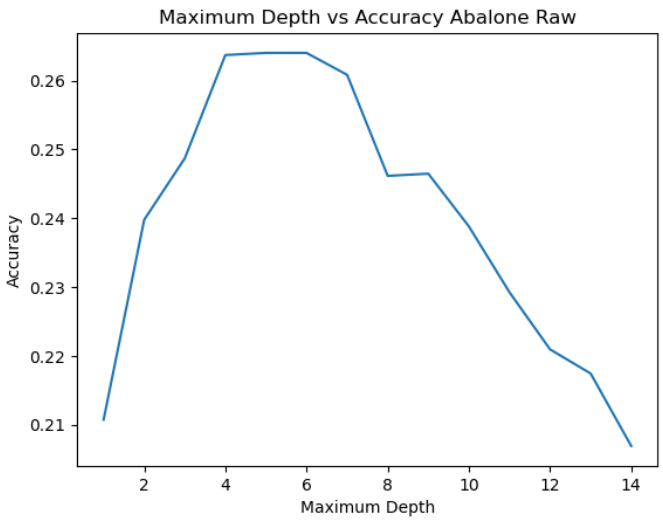
The library used for decision tree is sklearn.tree.

3.1 Abalone Data

3.1a Abalone Raw

The decision tree is run on Abalone Raw data with default settings and the accuracy comes up to be 19.62%. Which is less than knn but slightly more than MultiNomial and Complement Naive Bayes without 5 Fold.

Now, apply 'GridSearchCV' to tune the hyper parameters, for the scope of this report, the parameter to be tuned is the 'max_depth' of decision tree. Upon, applying this technique, the best parameter came up to be max_depth = 5, with accuracy = 24.69%. **Abalone Raw accuracy = 24.69%, max_depth = 5.** This is a significant increase from the default parameters. It is also evident from the graph that, best setting is depth=5.

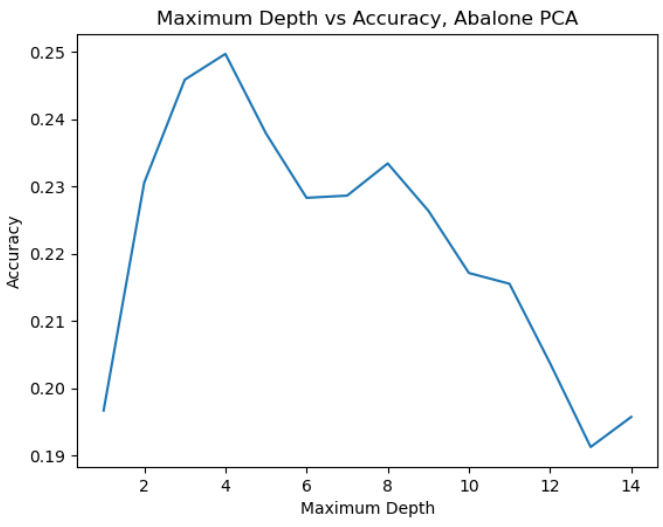


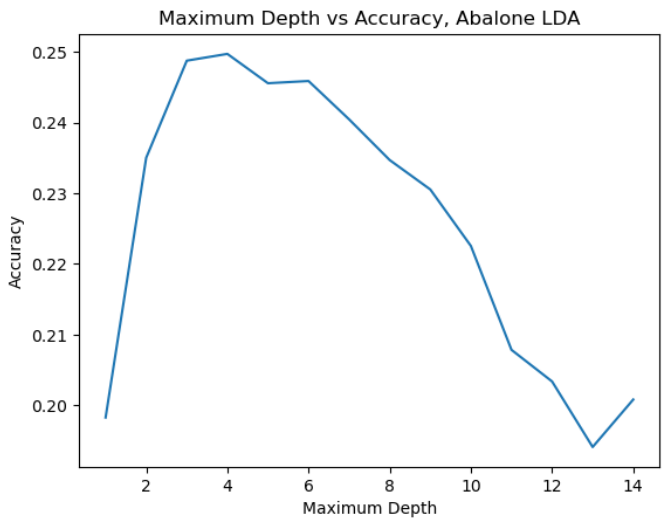
The above graph shows the accuracy vs maximum depth of Decision Tree for Abalone Raw data. The adjacent picture shows the decision tree as a text of rules to divide the nodes. The first split is made at Shell Weight <= 0.14 or Shell Weight > 0.14. This structure is easily interpretable because the data is in it's original form and has not been scaled and no dimensionality is reduced. However, for if PCA or LDA is applied on a dataset, it loses it's interpretability and if a decision tree is formed on such data, the rule on node split or impurity would be vague and not interpretable.

```
|--- Shell_weight <= 0.14
|   |--- Shell_weight <= 0.06
|   |   |--- Shell_weight <= 0.02
|   |   |   |--- Whole_weight <= 0.02
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Whole_weight >  0.02
|   |   |   |   |--- truncated branch of depth 2
|   |   |--- Shell_weight >  0.02
|   |   |   |--- Viscera_weight <= 0.02
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Viscera_weight >  0.02
|   |   |   |   |--- truncated branch of depth 2
|   |--- Shell_weight >  0.06
|   |   |--- Shell_weight <= 0.11
|   |   |   |--- Shucked_weight <= 0.08
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Shucked_weight >  0.08
|   |   |   |   |--- truncated branch of depth 2
|   |   |--- Shell_weight >  0.11
|   |   |   |--- Shucked_weight <= 0.23
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Shucked_weight >  0.23
|   |   |   |   |--- truncated branch of depth 2
|--- Shell_weight >  0.14
|   |--- Shell_weight <= 0.25
|   |   |--- Shucked_weight <= 0.43
|   |   |   |--- Shell_weight <= 0.19
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Shell_weight >  0.19
|   |   |   |   |--- truncated branch of depth 2
|   |   |--- Shucked_weight >  0.43
|   |   |   |--- Whole_weight <= 0.98
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Whole_weight >  0.98
|   |   |   |   |--- truncated branch of depth 2
|   |--- Shell_weight >  0.25
|   |   |--- Shell_weight <= 0.36
|   |   |   |--- Shucked_weight <= 0.43
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Shucked_weight >  0.43
|   |   |   |   |--- truncated branch of depth 2
|   |   |--- Shell_weight >  0.36
|   |   |   |--- Shucked_weight <= 0.55
|   |   |   |   |--- truncated branch of depth 2
|   |   |   |--- Shucked_weight >  0.55
|   |   |   |   |--- truncated branch of depth 2
```

3.1b Abalone PCA

Apply PCA on Abalone dataset with 95% variance explained. When default parameters of Decision Tree are used to train the Abalone PCA data, the accuracy comes down to 18.47%. The accuracy has decreased from abalone raw. Apply 'GridSearchCV' to tune the depth parameter. The best setting for depth comes as 4 and the corresponding accuracy as 27.94%. **Abalone PCA accuracy = 27.94%, max_depth=4.** The graph represents the change in accuracy with respect to max depth of the tree.





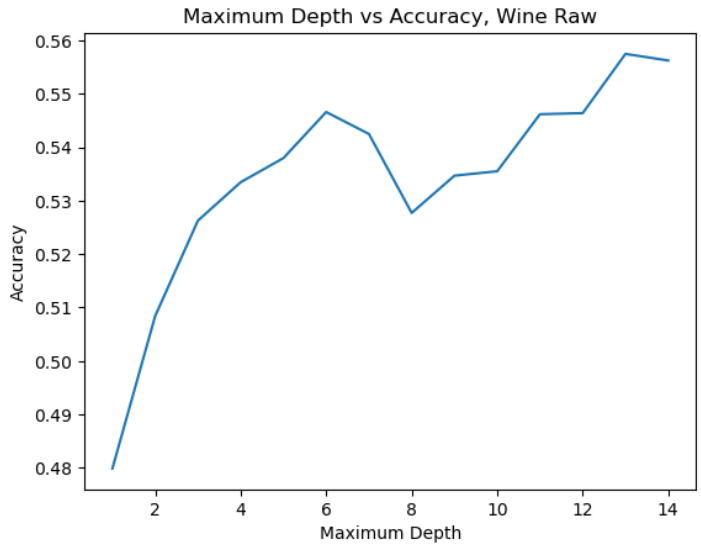
3.1c Abalone LDA

For abalone LDA, the accuracy with default set of parameters is 19.14%, more than the previous two. The best accuracy after tuning the depth parameter comes up to 24.69% with depth = 4. The above graph shows the accuracy vs depth graph for LDA Abalone. **Abalone LDA accuracy is 24.69%, max_depth = 4.**

3.2 Wine Data

3.2a Wine Raw

The decision tree is run on Wine Raw data with default parameters and the accuracy comes up to be 57.35%. Which is less than knn but more than MultiNomial and Complement Naive Bayes. After tuning the depth parameter, the best accuracy emerges as 58.40% with max_depth = 13. **Wine Raw accuracy is 58.40%, max_depth = 13.** The graph shows max_depth vs accuracy. The Decision tree text shows the split at alcohol <= 10.35 and goes on further. The tree is easily interpretable as the data is in it's original form.




```
|--- alcohol <= 10.35
|   |--- volatile acidity <= 0.24
|   |   |--- alcohol <= 8.85
|   |   |   |--- fixed acidity <= 7.25
|   |   |   |   |--- truncated branch of depth 5
|   |   |   |--- fixed acidity >  7.25
|   |   |   |   |--- truncated branch of depth 5
|   |   |--- alcohol >  8.85
|   |   |   |--- residual sugar <= 1.35
|   |   |   |   |--- truncated branch of depth 8
|   |   |   |--- residual sugar >  1.35
|   |   |   |   |--- truncated branch of depth 12
|   |--- volatile acidity >  0.24
|   |   |--- alcohol <= 9.85
|   |   |   |--- fixed acidity <= 10.85
|   |   |   |   |--- truncated branch of depth 12
|   |   |   |--- fixed acidity >  10.85
|   |   |   |   |--- truncated branch of depth 4
|   |   |--- alcohol >  9.85
|   |   |   |--- volatile acidity <= 0.56
|   |   |   |   |--- truncated branch of depth 12
|   |   |   |--- volatile acidity >  0.56
|   |   |   |   |--- truncated branch of depth 10
|--- alcohol >  10.35
|   |--- alcohol <= 11.44
|   |   |--- free sulfur dioxide <= 16.50
|   |   |   |--- sulphates <= 0.58
|   |   |   |   |--- truncated branch of depth 12
|   |   |   |--- sulphates >  0.58
|   |   |   |   |--- truncated branch of depth 11
|   |   |--- free sulfur dioxide >  16.50
|   |   |   |--- index <= 3997.00
|   |   |   |   |--- truncated branch of depth 12
|   |   |   |--- index >  3997.00
|   |   |   |   |--- truncated branch of depth 12
|   |--- alcohol >  11.44
|   |   |--- index <= 3774.00
|   |   |   |--- alcohol <= 12.55
|   |   |   |   |--- truncated branch of depth 12
|   |   |   |--- alcohol >  12.55
|   |   |   |   |--- truncated branch of depth 12
|   |   |--- index >  3774.00
|   |   |   |--- alcohol <= 12.78
|   |   |   |   |--- truncated branch of depth 9
|   |   |   |--- alcohol >  12.78
|   |   |   |   |--- truncated branch of depth 7
```

3.2b Wine PCA

The accuracy of Wine PCA with default parameters is 59.02%. After parameter tuning, the accuracy is 58.65% at max_depth = 14. **Wine PCA accuracy is 58.65% at max_depth = 14.**

3.3c Wine LDA

Decision tree Accuracy for Wine LDA with default params is 57.54%. After parameter tuning, it is 56.44% with max_depth = 17. **Wine LDA accuracy is 59.38% at max_depth = 17.**