Cara Evangeline Chandra Mohan
190539421

Data Mining - Lab Assignment 3

1.Compare the Number of Leaves and Size of Tree of both trees (i.e. with and without pruning) and explain any differences observed

	Number of Leaves	**Size of Tree**
With Pruning	61	93
Without Pruning	121	175

When a tree is pruned, the size of the tree decreases because irrelevant branches are removed which is not useful in classifying the tuples. Irrelevant branches cause overfitting of training dataset and leads to increase of error in classifying of test dataset and so either pre-pruning or post pruning is done.
When a tree is not pruned all branches which has no value to classify an instance gets added.

2.Compare the Test Accuracy of both trees. Which tree shows a better performance? Explain your observation based on the notion of tree pruning.

	Accuracy
With Pruning	90.5712
Without Pruning	86.6379

The tree which is pruned shows a better performance.
A tree is built based on either information gain or gini index. When information gain or gini index is below (for information gain) or above (for gini index) a certain threshold the adding of further more branches is stopped which is known as pre-pruning. Test error is calculated for every branch that is added, if the error is not gradually decreasing or if the error is increasing, the adding of furthermore branches is stopped. In post pruning the tree is completely built and error is calculated and accordingly the branches are removed.

3.Which class is being heavily misclassified? Why has this happened?

Virginica is highly misclassified.

Virginica class tuples are reduced 5 times than before which takes it down to 5 training tuples and 5 testing tuples. When the boundary region is plotted by using the fitting of training dataset, versicolor tuples dominates and hence the class 1(versicolor) region occupies a wider area which in turn leads to misclassifying of virginica class tuples.

Cara Evangeline Chandra Mohan
190539421

4.Obtain the accuracy for this class from the test dataset and identify the other class that it is being confused with.

	Actual Setosa	**Actual Versicolor**	**Actual Virginica**
Predicted Setosa	24	0	0
Predicted Versicolor	1	25	3
Predicted Virginica	0	0	2

Accuracy of class Virginica = 2/(3+2) = 0.4
3 Virginica class tuples has been misclassified as Versicolor.

5.What is the new accuracy for class 2 (virginica)? Compare this accuracy with the accuracy obtained in the previous section and explain any discrepancies.

	Actual Setosa	**Actual Versicolor**	**Actual Virginica**
Predicted Setosa	24	0	0
Predicted Versicolor	1	5	14
Predicted Virginica	0	0	11

Accuracy of class Virginica = 11/(14+11) = 0.44
Here, number of Versicolor class tuples is reduced down to 5 test tuples and number of Virginica class tuples is 25 but the boundary region is fitted using the previous problem statement where virginica tuple has 5 training and 5 test tuples. And hence Versicolor class has 100% accuracy and Virginica class has 44% accuracy because of the increased Virginica class tuples.

6.What prior should you use to get maximum accuracy in region B? What accuracy do you get by using this value?

```
gnb_A_with_uniform_priors = GaussianNB(priors=np.array([5/11,1/11,5/11]))
gnb_A_with_uniform_priors.fit(XtrImbalanced_A, YtrImbalanced_A)

print_classifier_report(XtrImbalanced_A, YtrImbalanced_A, XteImbalanced_B,
YteImbalanced_B, gnb_A_with_uniform_priors, True)
```

Cara Evangeline Chandra Mohan
190539421

	Actual Setosa	**Actual Versicolor**	**Actual Virginica**
Predicted Setosa	24	0	0
Predicted Versicolor	1	2	6
Predicted Virginica	0	3	19

Number of Setosa class tuples = 25
Number of Versicolor class tuples = 5
Number of Virginica class tuples = 25

The prior to get the maximum accuracy in region B is [25/55, 5/55, 25/55] and Accuracy of the classifier is: 0.818

7.Compare the performance of both classifiers in the 2-feature scenario with the performance in the 200-feature scenario and explain any differences you might observe.

For dimensions =2

Train accuracy (Logistic Regression): 0.78
Test accuracy (Logistic Regression): 0.72

Train accuracy (Naive Bayes): 0.78
Test accuracy (Naive Bayes): 0.72

For dimensions = 200

Train accuracy (Logistic Regression): 1.00
Test accuracy (Logistic Regression): 0.78

Train accuracy (Naive Bayes): 1.00
Test accuracy (Naive Bayes): 1.00

Increasing the number of features to 200, leads to overfitting of data and here in the example given, random points are considered for testing and training and there is a high probability that both test and train data are almost alike leading to a perfect accuracy of 100% in case of dimensions = 200 for naive bayes classifier.