Computer Assignment: Decision Tree

Submitted by- Vandit Sharma **Roll No.-** 17EC10060

Aim- Train and test a decision tree classifier on the Fisher Iris dataset

Results-

Case-I: criterion = 'entropy' without parameter tuning

- Accuracy on training data: 100.0
- Accuracy on test data: 100

Case-II: criterion = 'entropy' with parameter tuning

A deep tree captures more information about the data which causes overfitting in decision trees. So, in order to reduce overfitting, we need to reduce the max_depth of the tree. In *Case-I,* the max_depth of the tree was found to be 5. So here, I varied the depth to 4, 3 and then 2.

By default the min_samples_leaf value is set to 1. This might cause overfitting since a number of small branches might be created exclusively for one sample. Hence, min_samples_leaf can also be used to control over-fitting by defining that each leaf has more than one element. This will ensure that the tree cannot overfit the training dataset. Here, I varied the min_samples_leaf value to 2, 10 and then 50.

Finally, I observed the results when both these parameters were varied at the same time.

S. No.	max_depth	min_samples_leaf	Accuracy on training data	Accuracy on test data
1	4	default(1)	99.167	96.667
2	3	default(1)	96.667	93.333
3	2	default(1)	96.667	93.333
4	None	2	96.667	93.333
5	None	10	96.667	93.333
6	None	50	66.667	66.667
7	3	10	96.667	93.333
8	2	50	66.667	66.667

Table 1: Effect of parameter tuning on accuracy with *criterion = 'entropy'*

Case-III: criterion = 'gini' without parameter tuning

- Accuracy on training data: 100.0
- Accuracy on test data: 93.333

Case-IV: criterion = 'gini' with parameter tuning

A deep tree captures more information about the data which causes overfitting in decision trees. So, in order to reduce overfitting, we need to reduce the max_depth of the tree. In *Case-I,* the max_depth of the tree was found to be 4. So here, I varied the depth to 3, 2 and then 1.

By default the min_samples_leaf value is set to 1. This might cause overfitting since a number of small branches might be created exclusively for one sample. Hence, min_samples_leaf can also be used to control over-fitting by defining that each leaf has more than one element. This will ensure that the tree cannot overfit the training dataset. Here, I varied the min_samples_leaf value to 2, 10 and then 50.

Finally, I observed the results when both these parameters were varied at the same time.

S. No.	max_depth	min_samples_leaf	Accuracy on training data	Accuracy on test data
1	3	default(1)	97.5	96.667
2	2	default(1)	96.667	93.333
3	1	default(1)	66.667	66.667
4	None	2	96.667	93.333
5	None	10	96.667	93.333
6	None	50	66.667	66.667
7	2	10	96.667	93.333
8	1	50	66.667	66.667

Table 2: Effect of parameter tuning on accuracy with *criterion = 'gini'*

Observations-

- A good accuracy of over 90 percent was observed for almost all cases.
- There was not much difference in accuracy between cases with criterion = 'entropy' and criterion = 'gini'
- Very low accuracy was achieved in case of min_samples_leaf = 50. This is because keeping such a high value prevents the tree from learning sufficiently, which leads to a higher number of wrong predictions.