Assignment 2 Report Part -2

Name: Sai Saketh Aluru
Roll No. 16CS30030

Implementing a decision tree classifier for the simplified 20 newsgroup dataset.

Given data:

	archive	name	atheism	resources	alt	last	modified	december	version	atheist	...	illustrator	dps	toward	eastman	hackers	jennifer	schaertel
0	1	1	1	1	1	1	1	1	1	1	...	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
12	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
14	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

The accuracies obtained on training and test data with increasing depth of the decision tree is:

Depth of the tree	Training accuracy	Test accuracy
1	73.83309759547383,	73.83309759547383
2	82.09236569274269	81.47100424328147
3	83.69462770970782	81.61244695898161
4	85.67389255419415	83.5926449787836
5	87.8416588124411	81.75388967468176
6	90.85768143261075	82.46110325318246
7	92.17719132893497	82.31966053748232

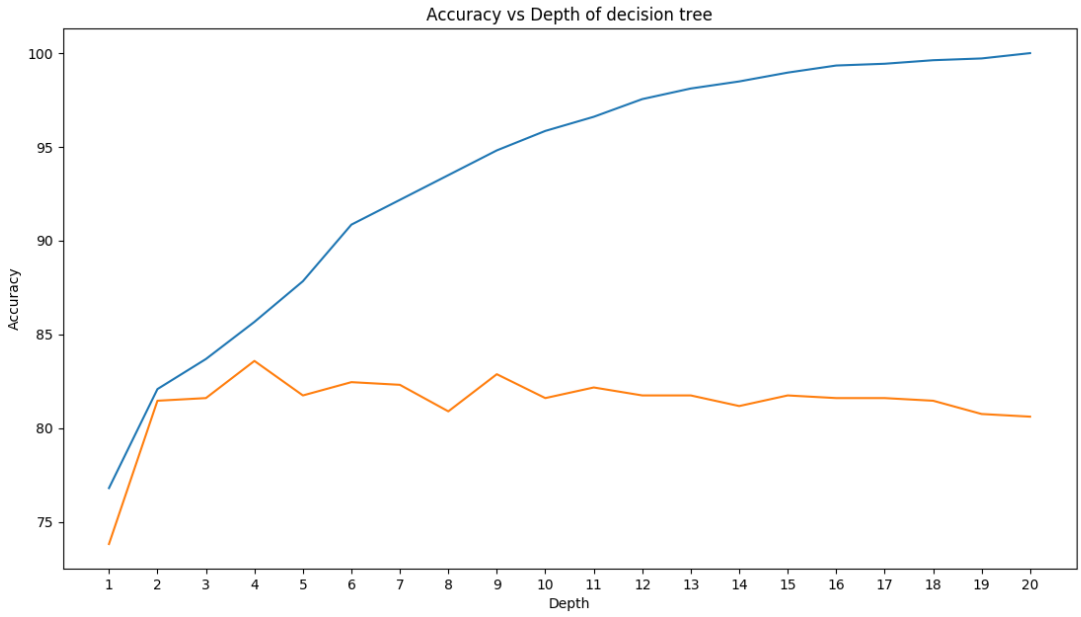
8	93.49670122525919	80.90523338048091
9	94.81621112158341	82.88543140028288
10	95.85296889726673	81.61244695898161
11	96.60697455230914	82.17821782178218
12	97.54948162111216	81.75388967468176
13	98.11498586239397	81.75388967468176
14	98.49198868991517	81.18811881188118
15	98.96324222431669	81.75388967468176
16	99.34024505183788	81.61244695898161
17	99.43449575871819	81.61244695898161
18	99.62299717247879	81.47100424328147
19	99.71724787935909	80.76379066478076
20	100.0	80.62234794908062

Note: The tree containing only a root node is specified as the tree with depth 1

We observe that the training accuracy increases steadily as the depth increases till a depth of 20 after which we reach the maximum tree size and a fully trained decision tree with training accuracy 100%. But we can see that initially, test accuracy increases with respect to the depth as the classifier is becoming better but after a depth of 4, we can see the test accuracy decreasing or varying unsteadily, which is due to the decision overfitting the data.

The best test accuracy occurs at a height of 4, at which the training accuracy is 85.67% and the test accuracy is 83.59%.

The obtained plot of the training and test accuracies w.r.t the depth of the decision tree is:



Note: Blue - training accuracy, Orange - Test accuracy

Using scikit-learn's decision trees and Information gain as the split criteria, the obtained accuracy values are:

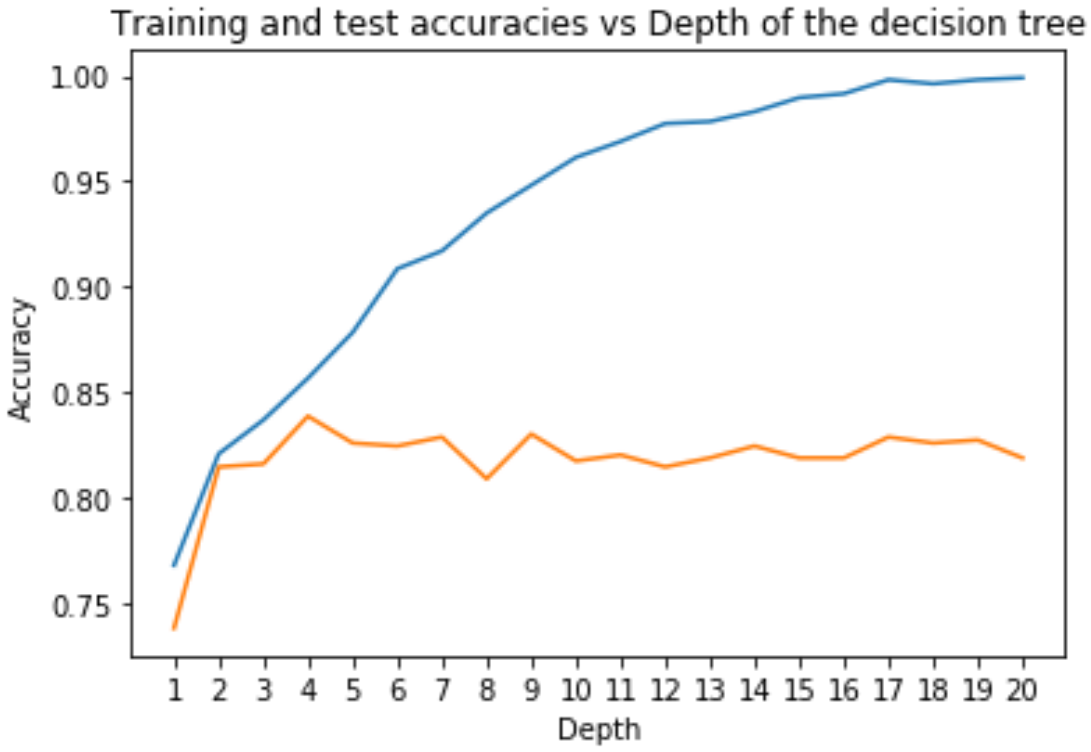
Depth	Training Accuracy	Test Accuracy
1	76.81432610744581	73.83309759547383
2	82.0923656927427	81.47100424328148
3	83.69462770970783	81.61244695898161
4	85.67389255419415	83.87553041018387
5	87.84165881244109	82.6025459688826
6	90.85768143261075	82.46110325318247
7	91.70593779453345	82.88543140028288
8	93.49670122525919	80.9052333804809
9	94.81621112158342	83.02687411598303
10	95.94721960414703	81.75388967468176
11	96.60697455230914	82.03677510608204
12	97.92648444863337	81.47100424328148
13	97.92648444863337	81.89533239038189
14	98.49198868991518	82.46110325318247

15	98.96324222431668	81.89533239038189
16	99.52874646559849	81.89533239038189
17	99.52874646559849	82.88543140028288
18	99.8114985862394	82.6025459688826
19	99.90574929311969	82.74398868458275
20	100	81.89533239038189

Note: The tree containing only a root node is specified as the tree with depth 1

The best accuracy occurs at a height of 4, similar to our decision tree from before. Here as well, we can see that the test accuracy increases first and then decreases, signifying overfitting.

The obtained plot of the training and test accuracies with respect to the depth of the decision tree is:



Note: Blue - training accuracy, Orange - Test accuracy

In the plot as well, we can see that overfitting occurs after a height of 4 as the test accuracy decreases and is varying unsteadily.

The tree formed at height 10 is

```
writes = 0
| god = 0
        | that = 0
                | bible = 0
                        | atheist = 0
                                | keith = 0 : 2
                                | keith = 1 : 1
                        | atheist = 1 : 1
                | bible = 1 : 1
        | that = 1
                | wrote = 0
                        | people = 0
                                | religious = 0 : 2
                                | religious = 1 : 1
                        | people = 1
                                | windows = 0 : 1
                                | windows = 1 : 2
                | wrote = 1
                        | ve = 0
                                | came = 0 : 1
                                | came = 1 : 2
                        | ve = 1 : 2
| god = 1
        | use = 0 : 1
        | use = 1
                | archive = 0 : 2
                | archive = 1 : 1
writes = 1
| graphics = 0
        | image = 0
                | that = 0
                        | god = 0
                                | keith = 0 : 2
                                | keith = 1 : 1
                        | god = 1 : 1
                | that = 1
                        | program = 0
                                | comp = 0 : 1
                                | comp = 1 : 2
                        | program = 1 : 2
        | image = 1 : 2
| graphics = 1 : 2
```

Here, we can see the features that are selected by this tree. Words like ‘atheist’, ‘god’, etc are a good indication of the alt.atheism class and words like ‘image’, ‘graphics’, ‘program’, etc are good indication of comp.graphics class labels. We can also see that some feature words such as ‘that’, ‘ve’, etc are not good words as they are very common and can occur in both documents.

The full tree obtained by the decision tree classifier is:

```
| writes = 0
    | god = 0
        | that = 0
            | bible = 0
                | atheist = 0
                    | keith = 0
                        | murder = 0
                            | laughter = 0
                                | liar = 0
                                    | for = 0
                                        | name = 0
                                            | hope = 0
                                                | us = 0
                                                    | word = 0
                                                        | events = 0
                                                            | the = 0
                                                                | to = 0
                                                                    | or = 0
                                                                        | and = 0
```

```
                                                | this = 0
                                                   | on = 0
                                                      | from = 0 : 2
                                                      | from = 1 : 2
                                                   | on = 1 : 2
                                                | this = 1 : 2
                                             | and = 1 : 2
                                          | or = 1 : 2
                                       | to = 1 : 2
                                    | the = 1 : 2
                                 | events = 1 : 1
                              | word = 1 : 1
                           | us = 1 : 1
                        | hope = 1
                           | are = 0 : 1
                           | are = 1 : 2
                     | name = 1
                        | it = 0 : 1
                        | it = 1 : 2
                  | for = 1 : 2
               | liar = 1 : 1
            | laughter = 1 : 1
         | murder = 1 : 1
      | keith = 1 : 1
   | atheist = 1 : 1
| bible = 1 : 1
| that = 1
   | wrote = 0
      | people = 0
         | religious = 0
            | an = 1
               | can = 0
                  | what = 0
                     | know = 0
                        | will = 0
                           | price = 0 : 1
                           | price = 1 : 2
                        | will = 1 : 2
                     | know = 1 : 2
                  | what = 1 : 1
               | can = 1
                  | day = 1 : 1
                  | day = 0
                     | atheism = 1 : 1
                     | atheism = 0 : 2
         | an = 0
```

```
                    | he = 0
                        | face = 0
                            | organizations = 1 : 1
                            | organizations = 0
                                | biblical = 0
                                    | doubt = 1 : 1
                                    | doubt = 0 : 2
                                | biblical = 1 : 1
                        | face = 1 : 1
                    | he = 1
                        | last = 0 : 1
                        | last = 1 : 2
                | religious = 1 : 1
            | people = 1
                | windows = 0 : 1
                | windows = 1 : 2
        | wrote = 1
            | ve = 0
                | came = 0
                    | last = 0 : 1
                    | last = 1 : 2
                | came = 1 : 2
            | ve = 1 : 2
    | god = 1
        | use = 1
            | archive = 1 : 1
            | archive = 0 : 2
        | use = 0 : 1
| writes = 1
    | graphics = 0
        | image = 0
            | that = 0
                | god = 0
                    | keith = 0
                        | who = 0
                            | have = 0
                                | am = 0
                                    | time = 0
                                        | by = 0
                                            | with = 0
                                                | you = 1 : 1
                                                | you = 0
                                                    | from = 0
                                                        | article = 0
                                                            | this = 1 : 1
                                                            | this = 0 : 2
```

```
                                    | article = 1
                                        | get = 0
                                            | try = 0 : 1
                                            | try = 1 : 2
                                        | get = 1 : 2
                                    | from = 1 : 2
                                | with = 1 : 2
                            | by = 1 : 1
                        | time = 1 : 2
                    | am = 1 : 2
                | have = 1
                    | alt = 0
                        | as = 1
                            | or = 1 : 1
                            | or = 0 : 2
                        | as = 0 : 2
                    | alt = 1 : 1
            | who = 1
                | the = 1 : 1
                | the = 0 : 2
        | keith = 1 : 1
    | god = 1 : 1
| that = 1
    | program = 0
        | comp = 0
            | uchicago = 0
                | edges = 0
                    | slow = 0
                        | files = 0
                            | vnet = 0
                                | address = 0
                                    | tiff = 0
                                        | seriously = 0
                                            | in = 1
                                                | small = 0
                                                    | either = 0 : 1
                                                    | either = 1
                                                        | which = 0 : 1
                                                        | which = 1 : 2
                                                | small = 1
                                                    | would = 1
                                                        | from = 1 : 1
                                                        | from = 0 : 2
                                                    | would = 0 : 1
                                            | in = 0
                                                | you = 1 : 1
```

```
                                        | you = 0
                                            | this = 1
                                                | to = 1 : 1
                                                | to = 0 : 2
                                            | this = 0 : 2
                                    | seriously = 1
                                        | see = 1 : 1
                                        | see = 0
                                            | bible = 1 : 1
                                            | bible = 0 : 2
                                | tiff = 1 : 2
                            | address = 1
                                | make = 1 : 1
                                | make = 0 : 2
                        | vnet = 1 : 2
                    | files = 1 : 2
                | slow = 1 : 2
            | edges = 1 : 2
        | uchicago = 1 : 2
    | comp = 1 : 2
| program = 1 : 2
| image = 1 : 2
| graphics = 1 : 2
```