COMP 5630 Fall 2022 Assignment 1

Will Humphlett (wah0028)

September 2, 2022

1. Decision Tree Basics.

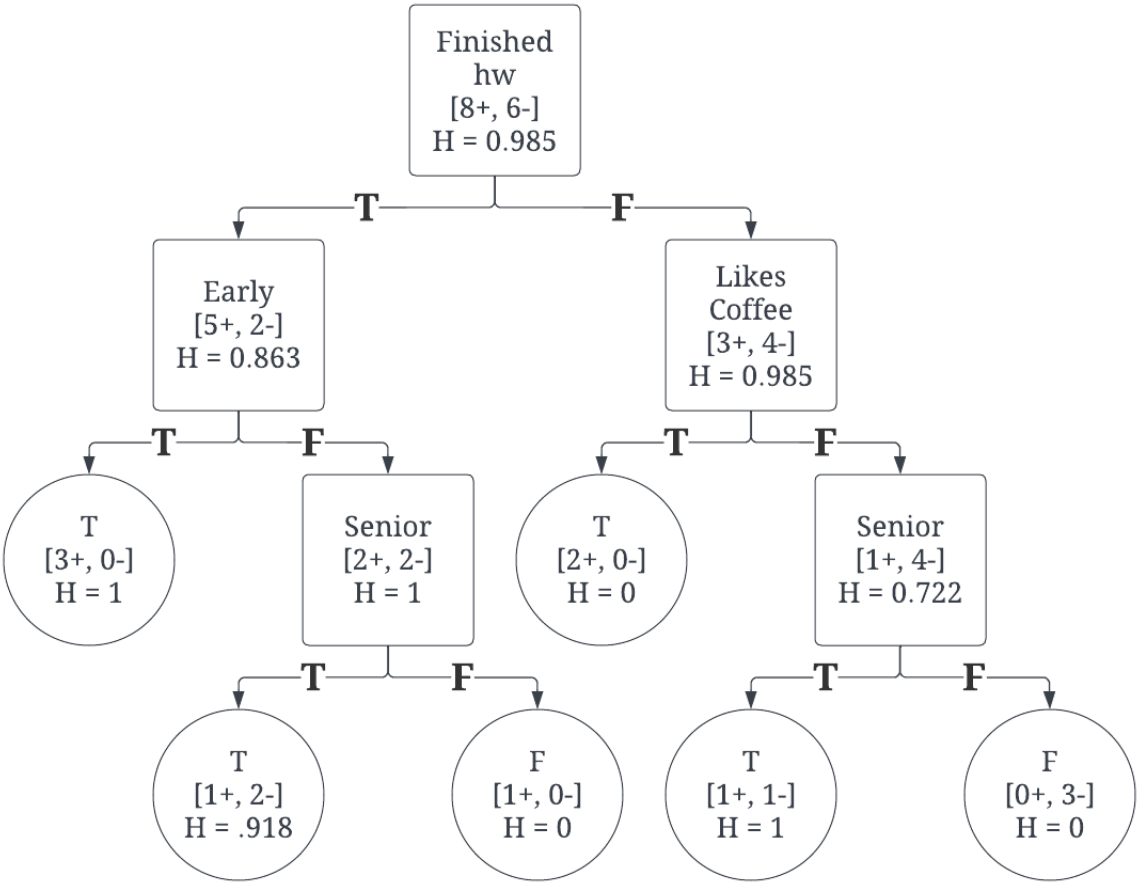
   (a) How many unique, perfect binary trees of depth 3 can be drawn if we have 5 attributes?

   i.

$$5 * 4 * 4 * 3 * 3 * 3 * 3 = \boxed{6480} \tag{1}$$

   ii.

$$\prod_{i=1}^{D} (A - i + 1)^{2^{i-1}} \tag{2}$$



   (b)

Figure 1: Tree of Depth 3 (Cut off tree at the appropriate depth for 1 & 2)

(c) Using the tree above, a good choice would be the tree of depth 3. It catches the senior split on the right side of the tree, but the split on senior on the left side of the tree may be due to overfitting. Further analysis with testing data should be completed to determine if depth=3 is overfitting when compared to depth=2.

(d) A decision tree is unrealizable if there exist 2 entries where the label values are equal but the outcomes are different. In a dataset where every combination of labels maps to one outcome, a decision tree with no depth limit could conceivably split on any label and fully realize the dataset presented. However, a split cannot be made if the labels are equal but the outcomes are different, therefore in such a case, the decision tree is unrealizable.

2. Application on a Real World Dataset (code at: github.com/wumphlett)

(a)
```
Training DecisionTreeClassifiers with variable max_depth
 2 : 93.79%
 3 : 93.79%
 4 : 94.44%
 5 : 94.74%
 6 : 94.80%
 7 : 95.03%
 8 : 95.15%
 9 : 95.24%
10 : 95.34%
k depth with highest accuracy = 10 (95.34%)
```

(b)
```
Testing DecisionTreeClassifier with optimal max_depth
10 : 94.77%
```

(c)
```
Overfitting Analysis
Training Acc: 95.34%, Total Acc: 94.76%
Overfitting?: True (0.58%)
```

There are slight overfitting issues when comparing the accuracies of the training data set to the test dataset.

```
DecisionTreeClassifiers testing accuracy with variable max_depth
 2 : 93.80%
 3 : 93.80%
 4 : 94.31%
 5 : 94.60%
 6 : 94.56%
 7 : 94.73%
 8 : 94.76%
 9 : 94.81%
10 : 94.76%
11 : 94.77%
12 : 94.70%
13 : 94.73%
```

```
14 : 94.68%
15 : 94.61%
16 : 94.52%
17 : 94.43%
18 : 94.28%
19 : 94.09%
20 : 94.01%
k depth with highest accuracy = 9 (94.81%)
```

Testing accuracy peaks at depth = 9 and roughly declines from there, implying that a decision tree of depth = 9 is best in practice instead of the depth = 10 implied by part a.