COMP 307 Assignment 1

Part 2 Report

Decision tree method applied to hepatitis-training & hepatitis-test

```
Reading data from file hepatitis-training
2 categories
16 attributes
Read 112 instances
Reading data from file hepatitis-test
2 categories
16 attributes
Read 25 instances

╔ ASCITES = True:
║  ╠ SPIDERS = True:
║  ║  ╠ VARICES = True:
║  ║  ║  ╠ STEROID = True:
║  ║  ║  ║  ╚ Class: live
║  ║  ║  ╚ STEROID = False:
║  ║  ║     ╠ SPLEENPALPABLE = True:
║  ║  ║     ║  ╠ FIRMLIVER = True:
║  ║  ║     ║  ║  ╚ Class: live
║  ║  ║     ║  ╚ FIRMLIVER = False:
║  ║  ║     ║     ╠ BIGLIVER = True:
║  ║  ║     ║     ║  ╠ SGOT = True:
║  ║  ║     ║     ║  ║  ╚ Class: live
║  ║  ║     ║     ║  ╚ SGOT = False:
║  ║  ║     ║     ║     ╠ FEMALE = True:
║  ║  ║     ║     ║     ║  ╚ Class: live
║  ║  ║     ║     ║     ╚ FEMALE = False:
║  ║  ║     ║     ║        ╠ ANOREXIA = True:
║  ║  ║     ║     ║        ║  ╚ Class: die
║  ║  ║     ║     ║        ╚ ANOREXIA = False:
║  ║  ║     ║     ║           ╚ Class: live
║  ║  ║     ║     ╚ BIGLIVER = False:
║  ║  ║     ║        ╚ Class: live
║  ║  ║     ╚ SPLEENPALPABLE = False:
║  ║  ║        ╠ HISTOLOGY = True:
║  ║  ║        ║  ╚ Class: die
║  ║  ║        ╚ HISTOLOGY = False:
║  ║  ║           ╚ Class: live
║  ║  ╚ VARICES = False:
║  ║     ╚ Class: die
║  ╚ SPIDERS = False:
```

```
|       ╠ BILIRUBIN = True:
|       |  ╠ FATIGUE = True:
|       |  |  ╠ AGE = True:
|       |  |  |  ╚ Class: live
|       |  |  ╚ AGE = False:
|       |  |     ╚ Class: die
|       |  ╚ FATIGUE = False:
|       |     ╠ ANTIVIRALS = True:
|       |     |  ╠ MALAISE = True:
|       |     |  |  ╚ Class: live
|       |     |  ╚ MALAISE = False:
|       |     |     ╚ Class: live
|       |     ╚ ANTIVIRALS = False:
|       |        ╚ Class: live
|       ╚ BILIRUBIN = False:
|          ╚ Class: live
╚ ASCITES = False:
    ╚ Class: die

Accuracy for hepatitis-test: 76.0% (19/25)
```

Testing over 10 other training and test sets

```
Accuracy for hepatitis-test/training-run-0: 76.7% (23/30)
Accuracy for hepatitis-test/training-run-1: 80.0% (24/30)
Accuracy for hepatitis-test/training-run-2: 66.7% (20/30)
Accuracy for hepatitis-test/training-run-3: 73.3% (22/30)
Accuracy for hepatitis-test/training-run-4: 80.0% (24/30)
Accuracy for hepatitis-test/training-run-5: 70.0% (21/30)
Accuracy for hepatitis-test/training-run-6: 80.0% (24/30)
Accuracy for hepatitis-test/training-run-7: 83.3% (25/30)
Accuracy for hepatitis-test/training-run-8: 63.3% (19/30)
Accuracy for hepatitis-test/training-run-9: 76.7% (23/30)

Average Accuracy: 75%
```

$$average\ accuracy = \frac{\frac{23+24+20+22+24+21+24+25+19+23}{30}}{10} = \frac{3}{4} = 0.75$$

How you could prune leaves from the tree

1. For each leaf node pair
    1. Compare the purity of the leaf node vs parent
    2. If the purity gain is under a minimal threshold
        1. Remove child leaf nodes
        2. Parent becomes leaf node

Why pruning would reduce accuracy on the training set

Pruning would reduce accuracy on the training set because you're removing some of the defined outcomes in the training set from the decision tree, the tree no longer covers every combination of outcomes defined by the training set.

Why pruning might increase accuracy on the test set

Pruning can increase accuracy on the test set as it reduces overfitting. Removing leaf nodes that have little influence on impurity helps wheedle out noise from the decision tree that may affect the classification of an instance.