Manoj Kumar Selvaraj

Professor Anjum Chida

CS 6375.001

21 February 2021

Assignment 2

Part A

```
TREE
+-- [SPLIT: x4 = 3 True]
|    +-- [SPLIT: x5 = 1 True]
|    |    +-- [SPLIT: x3 = 1 True]
|    |    |    +-- [LABEL = 0]
|    |    +-- [SPLIT: x3 = 1 False]
|    |    |    +-- [LABEL = 0]
|    +-- [SPLIT: x5 = 1 False]
|    |    +-- [SPLIT: x2 = 1 True]
|    |    |    +-- [LABEL = 1]
|    |    +-- [SPLIT: x2 = 1 False]
|    |    |    +-- [LABEL = 0]
+-- [SPLIT: x4 = 3 False]
|    +-- [SPLIT: x4 = 2 True]
|    |    +-- [SPLIT: x2 = 2 True]
|    |    |    +-- [LABEL = 0]
|    |    +-- [SPLIT: x2 = 2 False]
|    |    |    +-- [LABEL = 0]
|    +-- [SPLIT: x4 = 2 False]
|    |    +-- [SPLIT: x4 = 4 True]
|    |    |    +-- [LABEL = 0]
|    |    +-- [SPLIT: x4 = 4 False]
|    |    |    +-- [LABEL = 1]
Test Error = 27.08%.
```
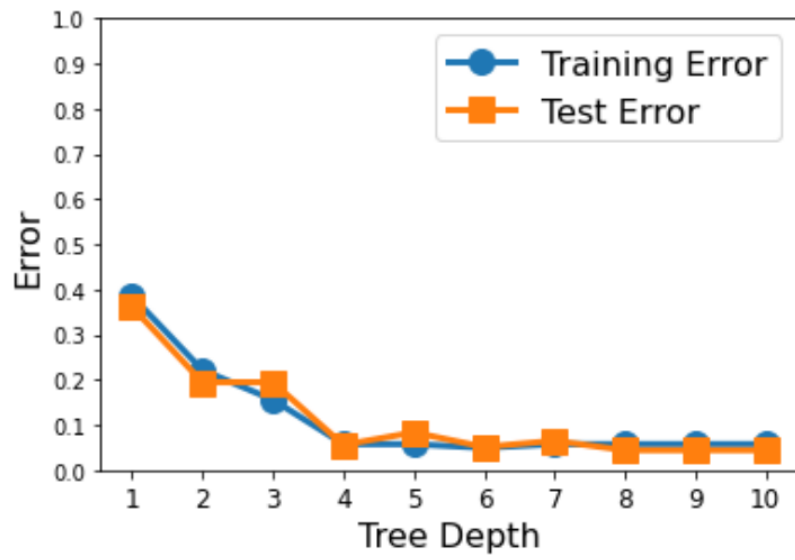
Sample output of tree on monks_data 1 for depth of 5.

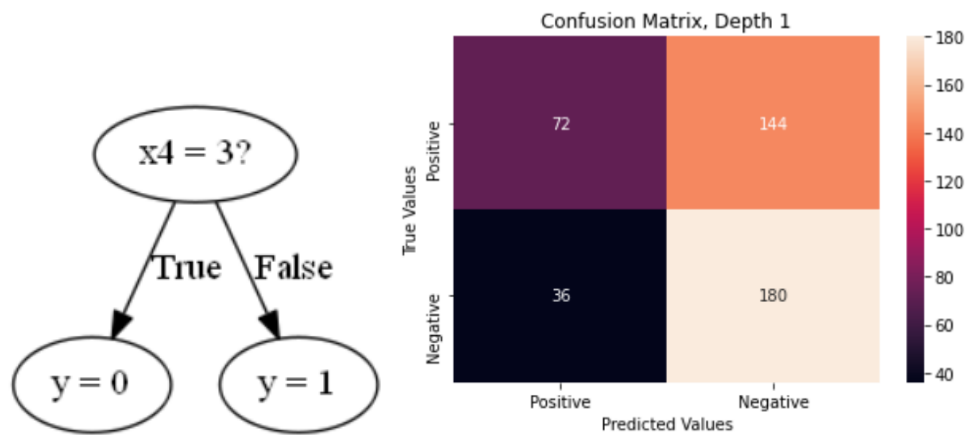**Part B**
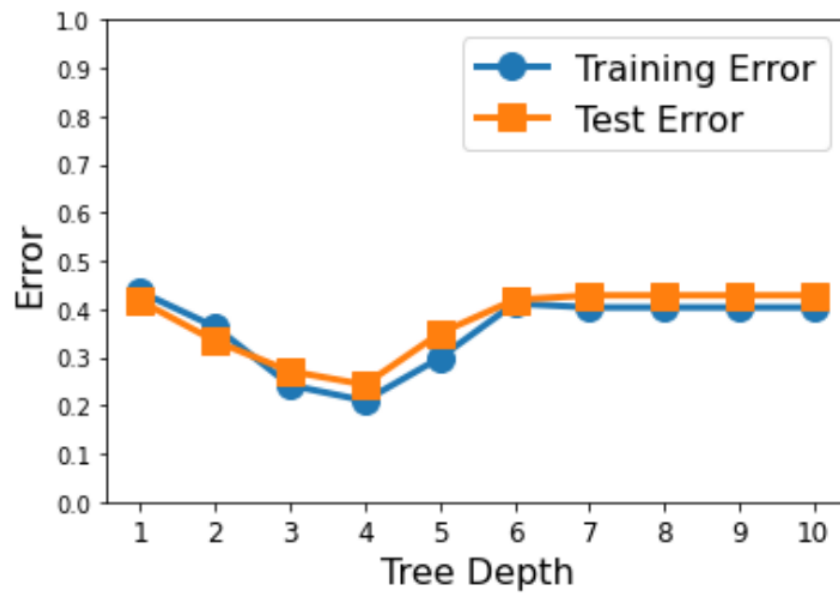
Training/Test Error plot for monks_data 1.
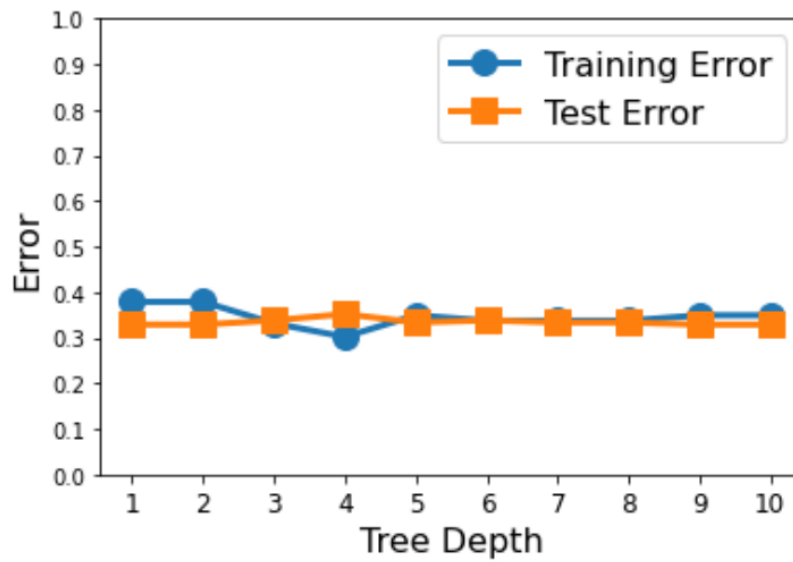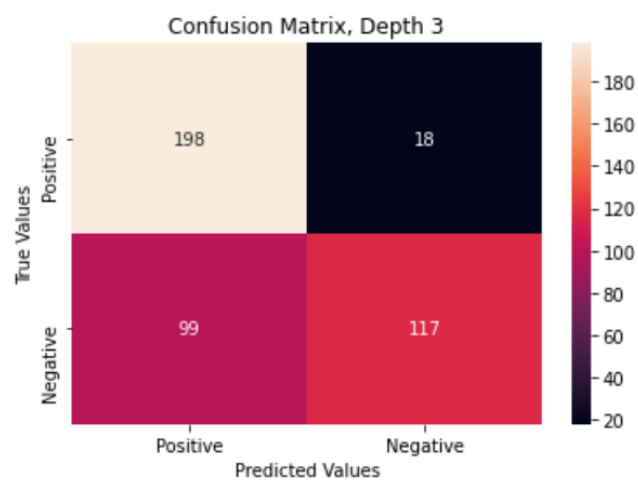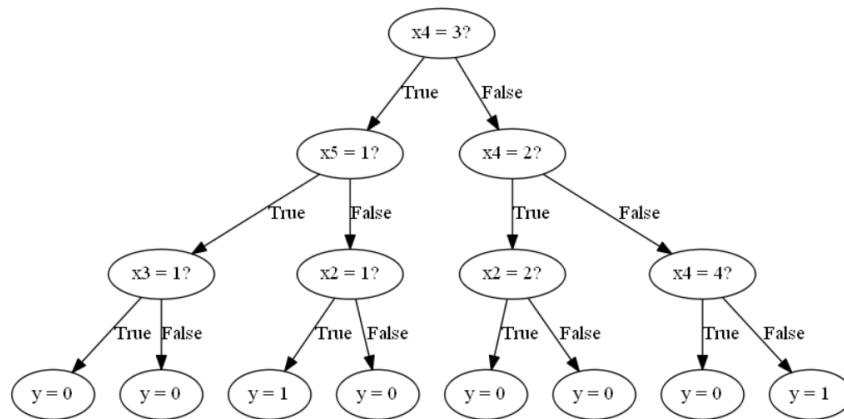

Training/Test Error plot for monks_data 2.

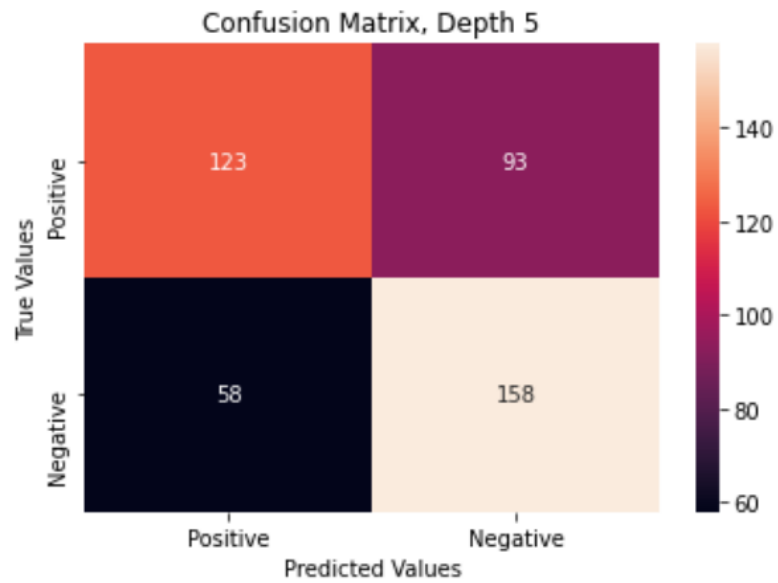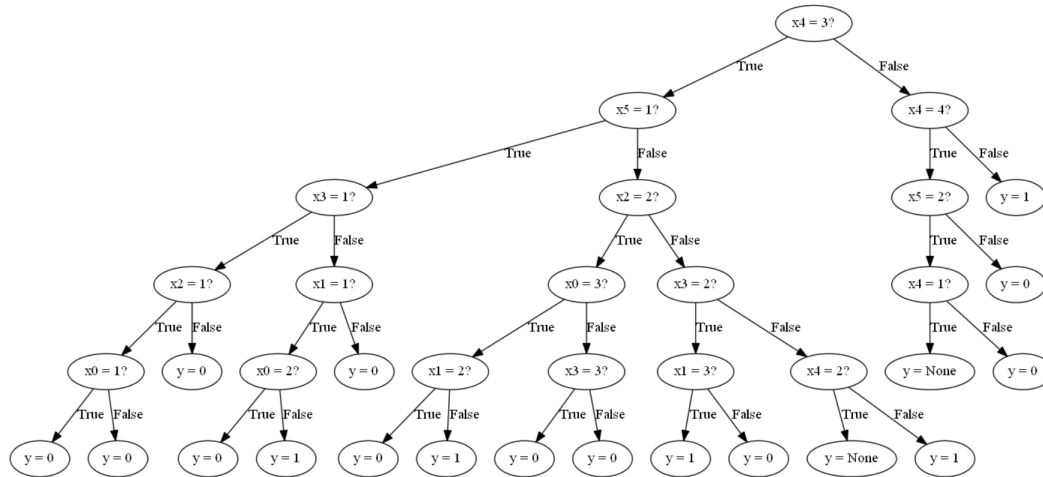Training/Test Error plot for monks_data 3.

Part C

Graphviz & Confusion Matrix output for tree on monks_data 1 for depth of 1.
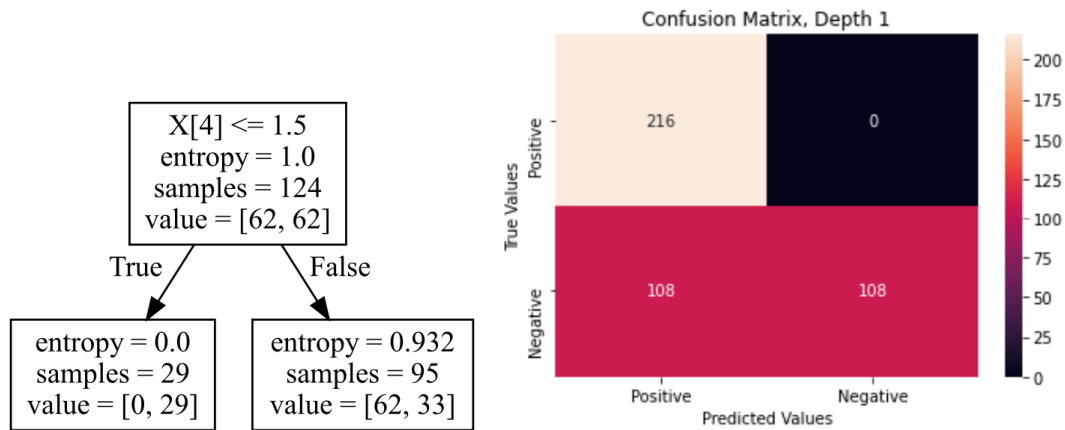

Graphviz & Confusion Matrix output for tree on monks_data 1 for depth of 3.

Graphviz & Confusion Matrix output for tree on monks_data 1 for depth of 5.
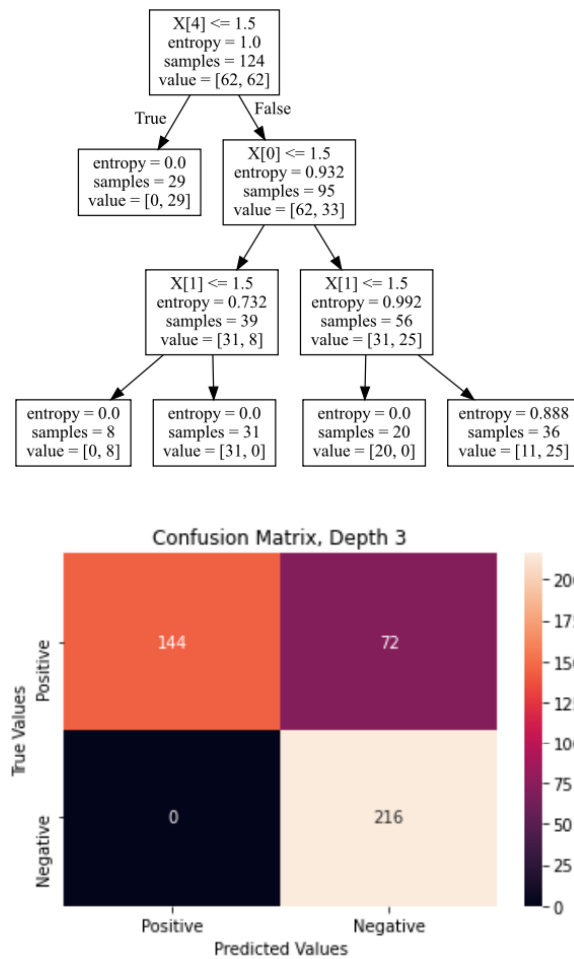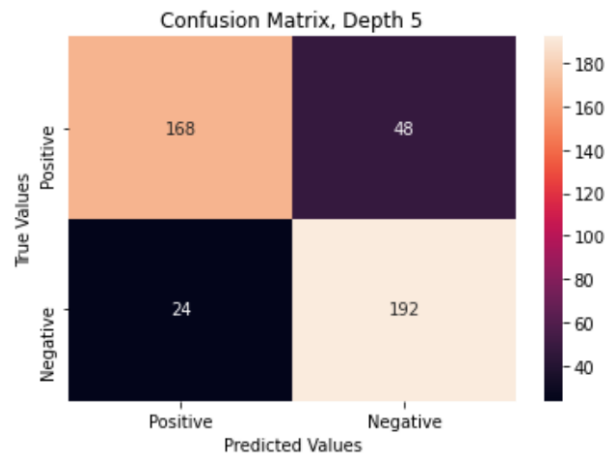
Part D

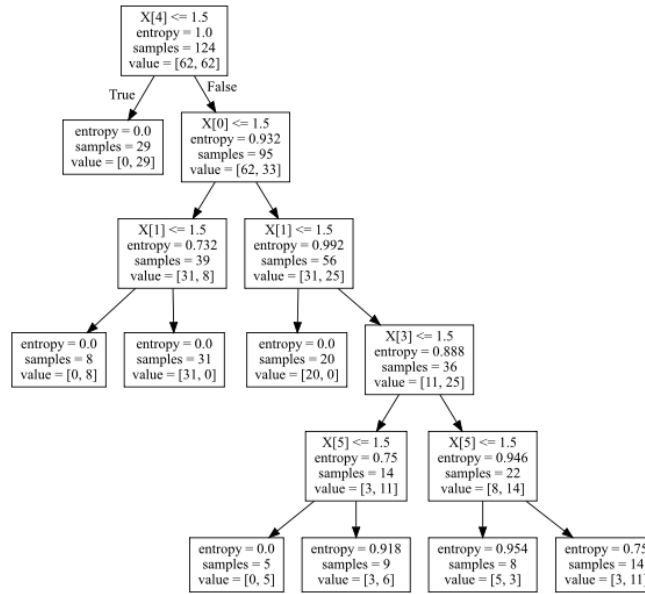Graphviz & Confusion Matrix output for SciKit tree on monks_data 1 for depth of 1.

Graphviz & Confusion Matrix output for SciKit tree on monks_data 1 for depth of 3.

Graphviz & Confusion Matrix output for SciKit tree on monks_data 1 for depth of 5.

Part E

Graphviz & Confusion Matrix output for tree on heart data for depth of 1.

Graphviz & Confusion Matrix output for tree on heart data for depth of 3.

Graphviz & Confusion Matrix output for tree on heart data for depth of 5.

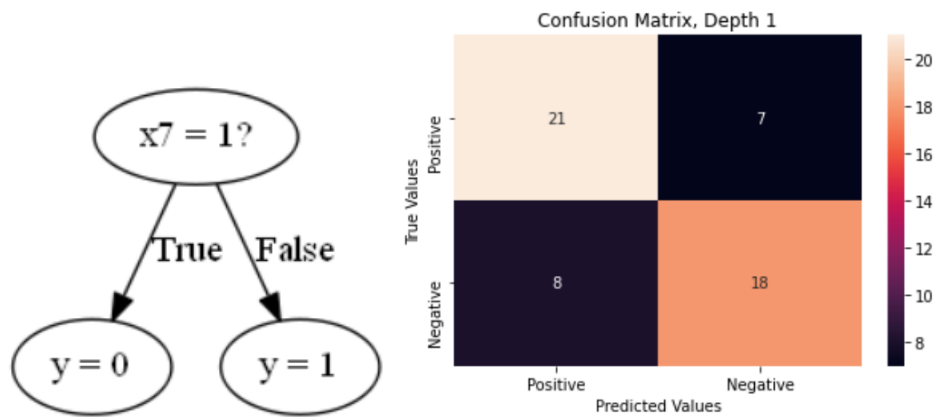Graphviz & Confusion Matrix output for SciKit tree on heart data for depth of 1.

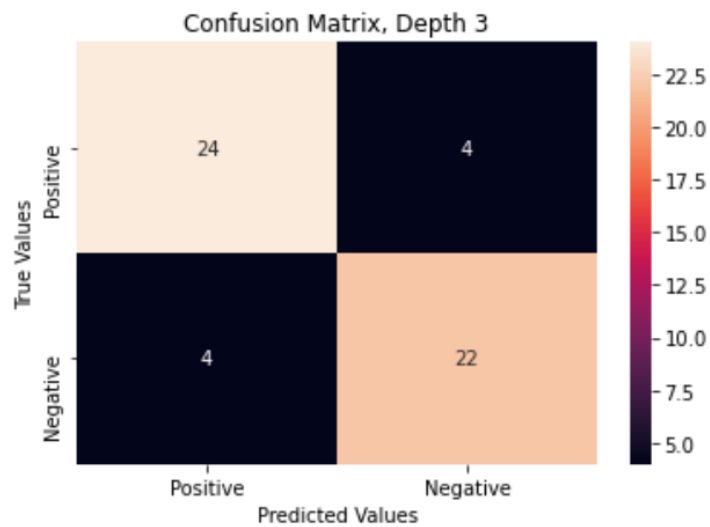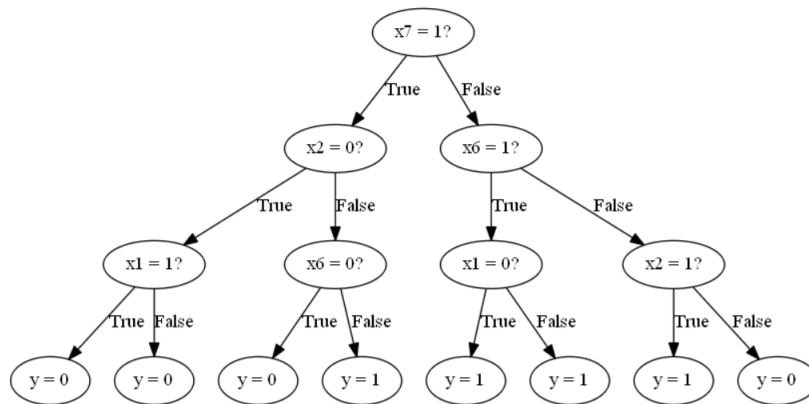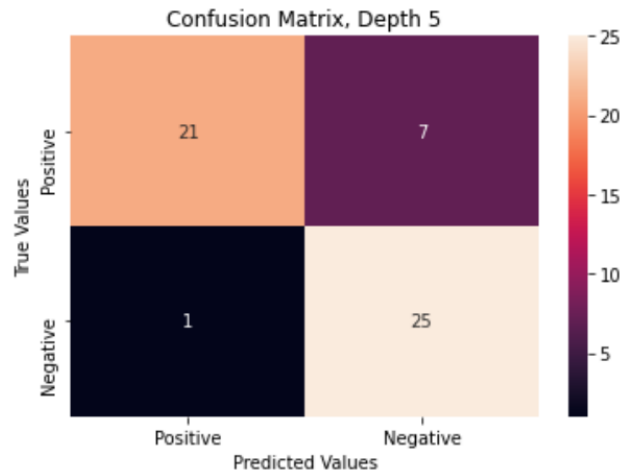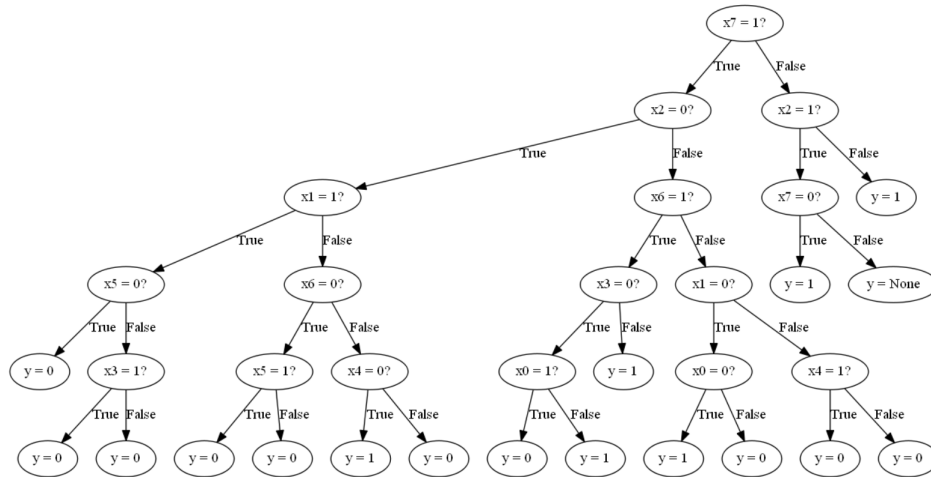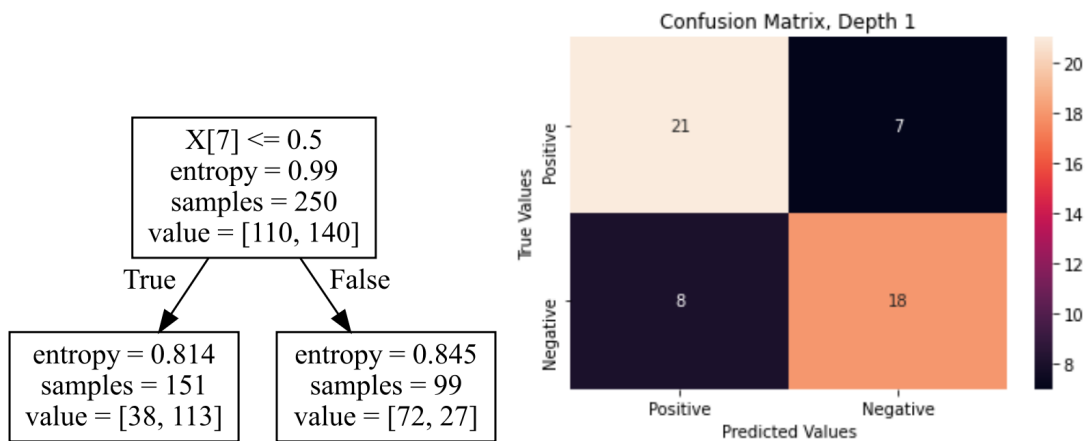Graphviz & Confusion Matrix output for SciKit tree on heart data for depth of 3.

Graphviz & Confusion Matrix output for SciKit tree on heart data for depth of 5.

Discussion

In comparison to the SciKit library, my implementation of ID3 was less accurate based on the confusion matrices. With regards to Part E, the SciKit library had a more balanced tree for the depth of 5; whereas, my implementation was skewed to the left. Tuning the tree and being able to find the correct depth for accurate estimation was challenging. With regards to Part D, the SciKit library was well rounded in terms of classification of true positives and true negatives. However, my implementation classified the true negatives with more accuracy than the true positives. As expected for both Parts D & E, both implementations classified the instances better when the tree depth was 5. However, both implementations introduced the label "None" when the depth of the tree was 5 or more. This may be due to the smaller size of the training data. In addition, it seems that the optimal depth size is unique for each data set such that different depths for the different data sets returned better classification results. Overall, this was an exciting project and was very valuable for learning the concept of decision trees.

DecisionTree_PartA.py contains the main ID3 algorithm code.

DecisionTree_PartBCD.py contains the visualization code such as Graphviz & Confusion Matrix for the ID3 and Scikit implementation.

DecisionTree_PartE.py contains the visualization code such as Graphviz & Confusion Matrix for the ID3 and Scikit implementation. The dataset used in this part is the heart disease data set. This dataset was converted into binary format by finding the average of each column and setting the individual cell value to 1 if greater than or equal to average and 0 otherwise.