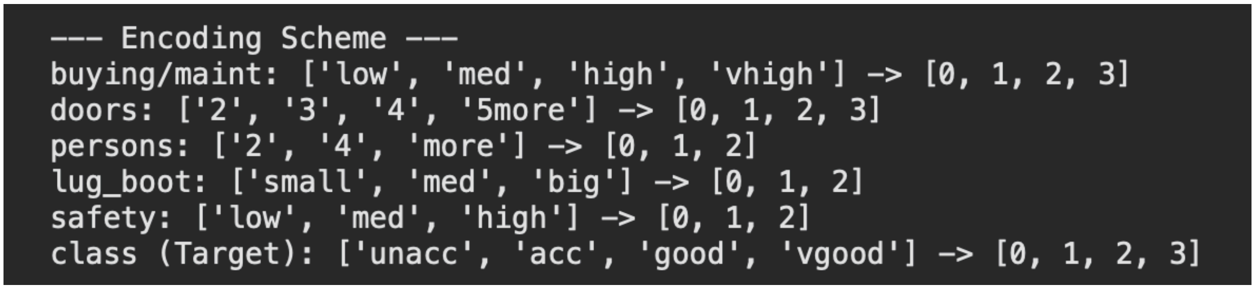
Question 2

Car_evaluation dataset

a) For the preprocessing of the dataset, we imported the dataset as df and then checked for missing values in the dataset. There were no missing values in the dataset, we then checked for the statistics and counts of the rows in our dataset and saw that everything is consistent.

b) We encoded the following categorical variables ordinally:

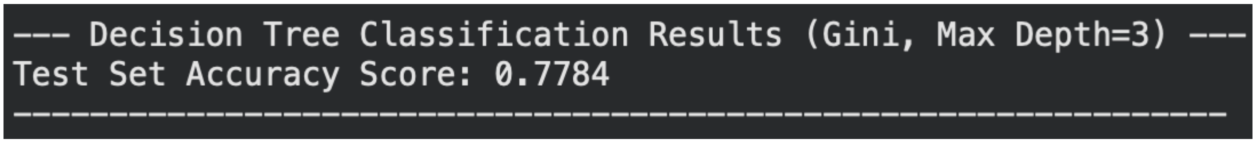


```
--- Encoding Scheme ---
buying/maint: ['low', 'med', 'high', 'vhigh'] -> [0, 1, 2, 3]
doors: ['2', '3', '4', '5more'] -> [0, 1, 2, 3]
persons: ['2', '4', 'more'] -> [0, 1, 2]
lug_boot: ['small', 'med', 'big'] -> [0, 1, 2]
safety: ['low', 'med', 'high'] -> [0, 1, 2]
class (Target): ['unacc', 'acc', 'good', 'vgood'] -> [0, 1, 2, 3]
```

c) We then built a decision tree classifier with the following conditions:

- criterion='gini'
- max_depth=3

We get the following test scores and get the following accuracy from our model:



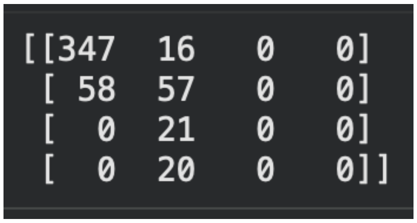
```
--- Decision Tree Classification Results (Gini, Max Depth=3) ---
Test Set Accuracy Score: 0.7784
---------------------------------------------------------------
```

d) We use our model on our training set too and get the following accuracies:

- Test accuracy: 0.7784
- Training accuracy: 0.7972

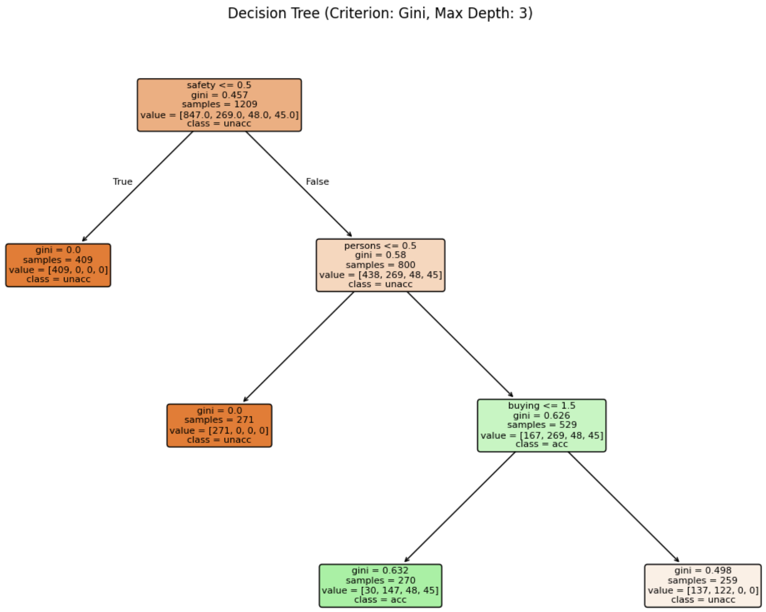
Our training and test accuracies are similar therefore we know that there is no overfitting in our model.

e) Confusion matrix:



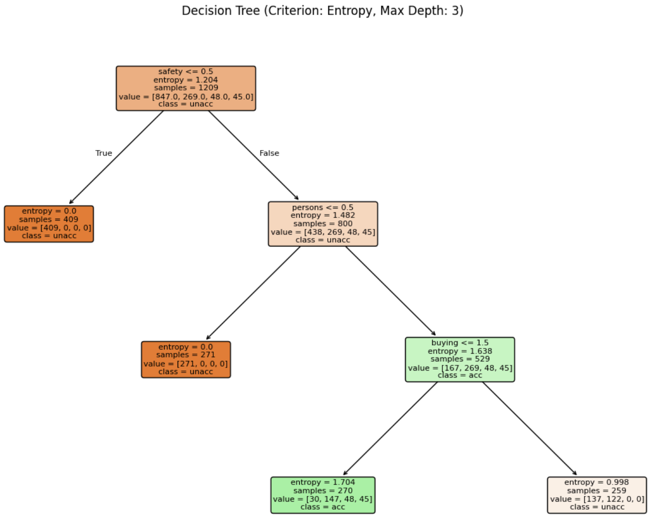
```
[[347  16   0   0]
 [ 58  57   0   0]
 [  0  21   0   0]
 [  0  20   0   0]]
```

f) We get the following decision tree from part c:

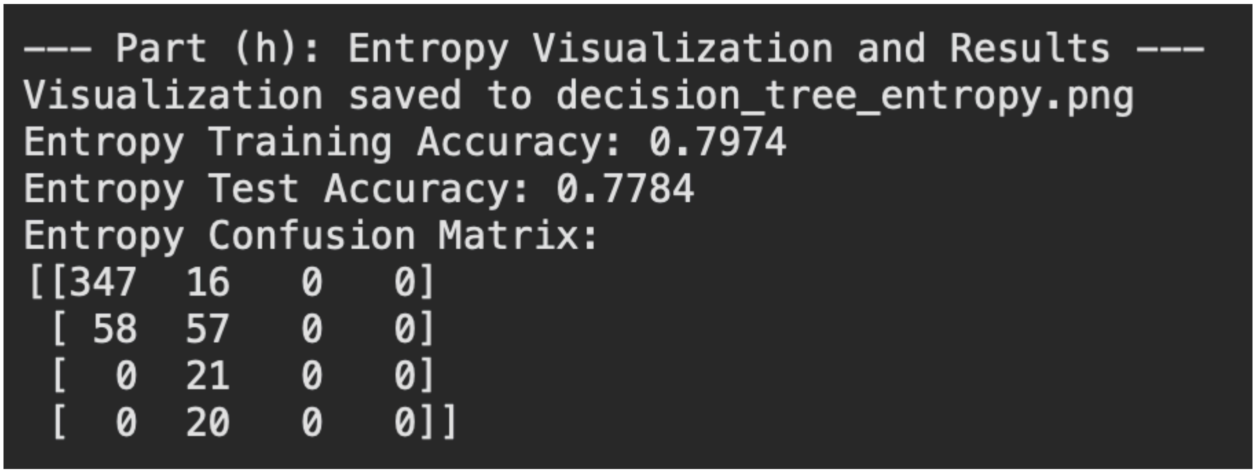


g) We couldn't import graphviz and therefore used matplotlib to print our tree instead.

h) After changing the criterion to 'entropy', we get the following output:



From this decision tree, we get the following output:



```
--- Part (h): Entropy Visualization and Results ---
Visualization saved to decision_tree_entropy.png
Entropy Training Accuracy: 0.7974
Entropy Test Accuracy: 0.7784
Entropy Confusion Matrix:
[[347  16   0   0]
 [ 58  57   0   0]
 [  0  21   0   0]
 [  0  20   0   0]]
```

i) From our accuracy and metrics, we see that the gini and entropy training and test accuracies are the same for both the decision trees. The model structures are the same and we can conclude that the model identical in structure and performance for max_depth = 3.