FoML Assignment 1 Report

Accuracy of Initial Implementation using Entropy and Information Gain

1. K-Fold(K=10) Cross Validation without Random Sampling Accuracy: **74.90%**
2. K-Fold(K=10) Cross Validation with Random Sampling Accuracy: **84.06%**

```
In [11]: #executing this cell may take 2-3 minutes
         accuracy_without_random = KFoldCross_without_random_sampling(df,10,0)#0 means using entropy
         print("Accuracy using KFold without Random Sampling: {}".format(np.mean(accuracy_without_random)))

         accuracy_with_random = KFoldCross_with_random_sampling(df,10,0) #0 means using entropy
         print("Accuracy using KFold with Random Sampling: {}".format(np.mean(accuracy_with_random)))

         Accuracy using KFold without Random Sampling: 74.90715958514554
         Accuracy using KFold with Random Sampling: 84.06122448979592
```

It can be seen by the accuracies that Cross Validation with random sampling works much better than without random sampling, as it randomly splits the data into training and testing this will ensure that the model will not be affected by any inbuilt bias in the dataset.

Accuracy of Improved Implementation

1. **Gini Index** instead of Entropy(K-Fold Cross Validation with Random Sampling) Accuracy: **83.32%**

Using Gini index instead of Entropy the accuracy doesn't seem to change that much, but the structure of the tree and the split values definitely have changed.

The KFoldCrossVal Accuracy using Gini Index

```
In [12]: accuracy_with_random = KFoldCross_with_random_sampling(df,10,1) #1 means using gini
         print("Accuracy using KFold with Random Sampling and Gini: {}".format(np.mean(accuracy_with_random)))

         Accuracy using KFold with Random Sampling and Gini: 83.3265306122449
```

In [31]:
```
#method=0 means we are constructing the decision tree using entropy
print('Decision Tree Representation using Entropy')
tree=pruned_decision_tree_algorithm(df,method=0,max_depth=3)
pprint(tree)

print('\n')

#method=1 means we are constructing the decision tree using gini
print('Decision Tree Representation using Gini')
tree=pruned_decision_tree_algorithm(df,method=1,max_depth=3)
pprint(tree)
```

```
Decision Tree Representation using Entropy
{'alcohol <= 10.6': [0.0,
                     {'alcohol <= 11.7333333333333': [0.0,
                                                      {'free_sulfur_dioxide <= 21.0': [0.0,
                                                                                       1.0]}]}]}


Decision Tree Representation using Gini
{'alcohol <= 10.8': [{'volatile_acidity <= 0.2': [{'density <= 0.99784': [0.0,
                                                                          1.0]},
                                                  0.0]},
                     {'alcohol <= 12.5': [0.0,
                                          {'free_sulfur_dioxide <= 20.0': [0.0,
                                                                           1.0]}]}]}
```

2. **Pruning the Tree** to a certain depth and using min sampling threshold(K-Fold Cross Validation with Random Sampling) Accuracy : **84.22%**

After Pruning the tree the accuracy came out to be slightly better than the previous approach but still not a huge improvement.

K-Fold CrossVal Accuracy using Pruning

In [14]:
```
#executing this cell may take 2-3 minutes
accuracy_with_pruning = KFoldCross_with_Pruning(df,10,0)
print("Accuracy using KFold with Pruning: {}".format(np.mean(accuracy_with_pruning)))
```

```
Accuracy using KFold with Pruning: 84.22448979591836
```