HW Decision Tree Creation – Classifying the Target Variable Muffins

a. What stopping criteria did you use?

The stopping criteria I used are as follows:

1. Stop and return the class when the datapoints left are less than 5.
2. Stop and return the class when the splits have datapoints such that 90% of them belong to either Class 0 or Class 1

b. Did you use any pruning or post-pruning?

Yes, I have used Pre-pruning as early stooping criteria which is mentioned above in Question 1. Part 2.

c. What splitting decision were you using?

I have used Weighted Gini Index as the Measure of Badness to find the best split i.e the best attribute, best threshold.

d. What structure did your final decision tree classifier have? What was the if-else tree you got? (Copy it into your write-up for completeness.)

The Decision tree Classifier I got is as follows:

```
If row['Sugar'] <= 19.1
        If row['Egg'] <= 12.1
                Class = 1
        Else
                Class = 0
Else:
        If row['FlourOrOats'] <= 41.31
                Class = 0
        Else:
                If row['FlourOrOats'] <= 42.42
                        Class = 1
                Else:
                        Class = 0
```

e. Run the original training data back through your classifier. What was the accuracy of your resulting classifier, on the training data?

The Accuracy is: 0.9016393442622951

f. Did your program actually create the classifier program, or did it just generate the attribute list and thresholds for you to hand-code in.

I only generated the attribute list and thresholds to hand-code in.

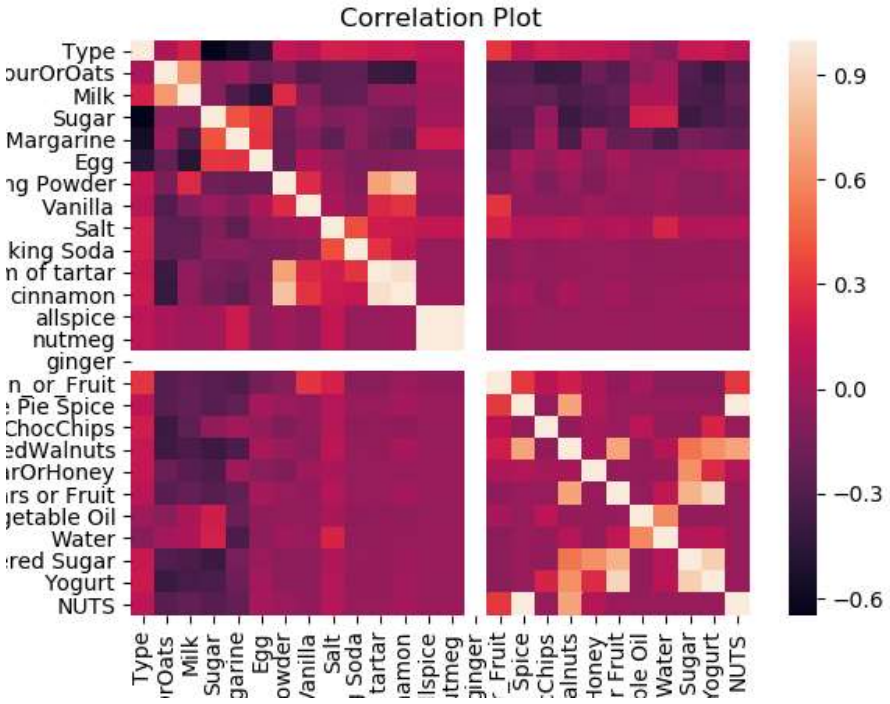
g. What else did you learn along the way here?

Even though we do not perform data cleaning and remove the variables that are insignificant (one's which have mostly 0) from the dataset, it does not affect the decision tree in any way.

Also, I understand how the pre-pruning technique (early stopping criterion) can handle overfitting. Without this stopping criterion there would be multiple splits and outcomes for the data that would result in higher error rate on the validation dataset.

The most fascinating part of this homework was to write a code that writes a code.

The correlation plot for the data is as follows:



We see in this correlation plot that the variable ginger has white blank because the data for ginger has only zero values. Also, we see a strong negative correlation within variables like FlourOrOats and cream of tartar, cinnamon, walnuts etc.

Also, we see that the correlation between the variables in my decision tree i.e FlourOrOats , Sugar and Egg is Zero or close to Zero.

h. What can you conclude?

Of all the 26 attributes, my decision tree has used only 3 variables to make a decision on whether it was a Muffin or a Cupcake.
Thus, I would conclude that the decision tree has very well made the use of significant variables based on the measure of badness to decide upon and class.
Thus, we can say that we do not need all the ingredients (attributes) used in the receipe to classify if it was a muffin or if it was a cupcake. The decision tree automatically does the feature selection within the process.