I. Your information

// Course: CS3642- Artificial Intelligence
// Student name: Jonathan Turner
// Student ID: jturn298
// Assignment #: 4
// Due Date: 7 April 2024
// Signature:

II. [65 points]. Decision tree for watermelon ripeness classification.

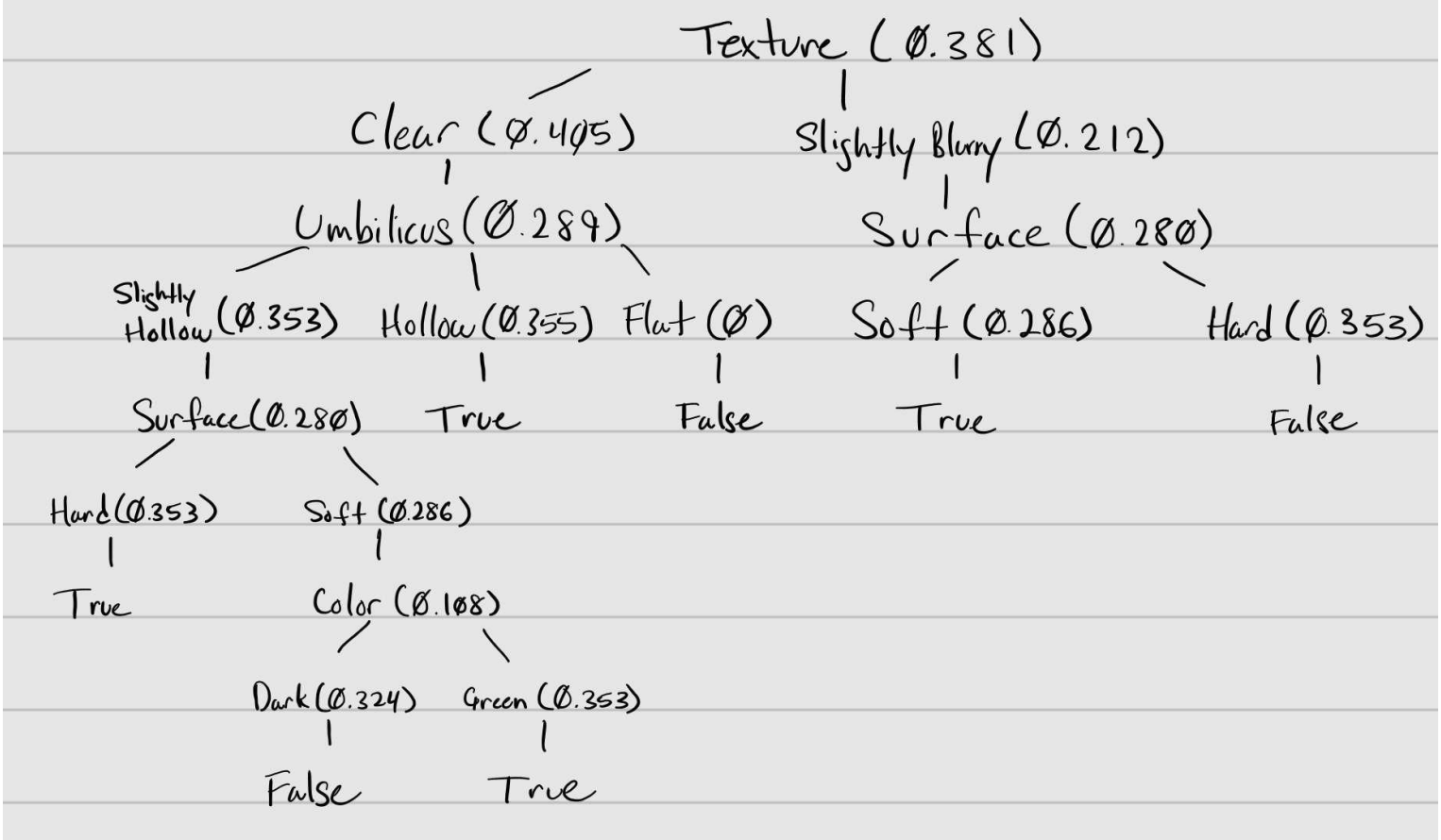
1) [5 points] What is the entropy of the root node? ($Entropy(D)$)
Entropy(D) of the root is 0.997502546

2) [5 points] Suppose that we have selected color, which has three possible values {Green, Dark, Light}. If dataset (D) is split by color, then there are three subsets: D! (Color=Green), D" (Color=Dark), and D#(Color=Light), what is the entropy of the three child nodes?
The entropy of the Color subsets are as follows:

D! = $-\frac{3}{6}log_2\frac{3}{6}-\frac{3}{6}log_2\frac{3}{6}=1$
D" = $-\frac{4}{6}log_2\frac{4}{6}—\frac{2}{6}log_2\frac{2}{6}=0.918$
D# = $-\frac{1}{5}log_2\frac{1}{5}--\frac{4}{5}log_2\frac{4}{5}=0.722$

3) [5 points] What is the information gain: $Gain(D, Color) = Entropy(D)$
The information gain for the Color set is:
The information gain for the Color set is calculated below:
$0.47058824-\left[\frac{6}{17}(1)+\frac{6}{17}(0.918)+\frac{5}{17}(0.722)\right]=0.108125165$

4) [40 points] Iterate other attributes **and draw the final decision tree**. You need to show the *entropy* calculation of each child node and the *Gain* calculation of each splitting node.



5) [10 points] Given a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$, is it a ripe watermelon? Please provide an explanation based on the decision tree built in 4).

Using the decision tree above, first we would look at the Texture value of the melon. This would be Clear, leading us to check the Umbilicus (Hollow) which leads to True.

This watermelon with the specified attributes using this tree would be classified as a Ripe melon.

III. [35 points] Naïve Bayes classifier, to predict the label of a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$

When creating the Naïve Bayes classifier, I first calculated the Prior Probability for both ripe and unripe watermelons. p(ripe)= 0.470588235 and the p(unripe)= 0.529411765. After that I created the probability of each value showing up on the ripe and unripe charts. I calculated it for all the categories/input data with each set of the following:

Ripe
{Color | Green: 0.375, Dark: 0.5, Light: 0.125}
{Root |Curly: 0.625, Slightly Curly: 0.375, Straight: 0}
{Sound | Muffled: 0.75, Dull: 0.25, Crisp: 0}
{Texture | Clear: 0.875, Slightly Blurry: 0.125, Blurry: 0}
{Umbilicus | Hollow: 0.625, Slightly Hollow: 0.375, Flat: 0}
{Surface | Soft: 0.25, Hard: 0.75}

Not Ripe
{Color | Green: 0.333, Dark: 0.222, Light: 0.444}
{Root |Curly: 0.333, Slightly Curly: 0.444, Straight: 0.222}
{Sound | Muffled: 0.444, Dull: 0.333, Crisp: 0.222}
{Texture | Clear: 0.222, Slightly Blurry: 0.444, Blurry: 0.333}
{Umbilicus | Hollow: 0.222, Slightly Hollow: 0.333, Flat: 0.444}
{Surface | Soft: 0.333, Hard: 0.667}

With these calculated, I now can calculate the Prior Probability with the probability of each of the inputs and compare the two calculations against each other.

Probability of given watermelon being ripe:

$$p(r)=p(yes)*p(C = g \mid y)*p(R = sc \mid y)*p(S = d \mid y)*p(T = c \mid y)*p(U = h \mid y)*p(S = h \mid y)$$
$$p(r)=0.470588235*0.375*0.375*0.25*0.875*0.625*0.75=0.006785673$$

Probability of given watermelon not being ripe:

$$p(nr)=p(no)*p(C = g \mid n)*p(R = sc \mid n)*p(S = d \mid n)*p(T = c \mid n)*p(U = h \mid n)*p(S = h \mid n)$$
$$p(nr)= 0.529411765*0.333*0.444*0.333*0.222*0.222*0.667=0.000860701$$

After calculating the probability, since the probability of the watermelon being ripe is greater than it not, (0.006785673 > 0.000860701), the given watermelon is ripe.

The calculation Excel files are attached as well.