Project 3

Group Name P2Eta

Group Members: Paul Rayment, Andrew Peters, Roman Formicola

Q2

2a)

Split 1: 0.545263726
Split 2: 0.509400729

2a work)

GINI Index of N = 1 - ((100 / 210) ^ 2 + (50 / 210) ^ 2 + (60 / 210) ^ 2) 0.634920635

GINI Index of N1, 1 = 1 - ((56 / 68) ^ 2 + (12 / 210) ^ 2 + (0 / 210) ^ 2) 0.318534002

GINI Index of N1, 2 = 1 - ((44 / 142) ^ 2 + (38 / 142) ^ 2 + (60 / 142) ^ 2) 0.653838524

GINI For Split 1

(68 / 210) *GINI N1, 1 + (142 / 210)* GINI N1, 2 (68 / 210) *0.318534002 + (142 / 210)* 0.653838524

0.545263726

GINI Index of N2, 1 = 1 - ((62 / 80) ^ 2 + (18 / 80) ^ 2 + (0 / 80) ^ 2) 0.34875

GINI Index of N2, 2 = 1 - ((28 / 63) ^ 2 + (11 / 63) ^ 2 + (24 / 63) ^ 2) 0.626858151

GINI Index of N2, 3 = 1 - ((10 / 67) ^ 2 + (21 / 67) ^ 2 + (36 / 67) ^ 2) 0.590777456

GINI For Split 2

(80 / 210) *GINI N2, 1 + (63 / 210)* GINI N2, 2 + (67 / 210) *GINI N2, 2 (80 / 210)* 0.34875 + (63 / 210) *0.626858151 + (67 / 210)* 0.590777456

0.509400729

2b)

Split 1 would be preferred with node N1, 2 preferred

2c)

Split 1 Information Gain: 0.247948701
Split 2 Information Gain: 0.336222303

Work for 2c)

Entropy of N = -((100 / 210) *log2((100 / 210)) + (50 / 210)* log2((50 / 210)) + (60 / 210) * log2((60 / 210))) 1.51904616

Entropy of N1, 1 = -((56 / 68) *log2((56 / 68)) + (12 / 68)* log2((12 / 68)) + (0 / 68) * log2((0 / 68))) 0.672294817

Entropy of N1, 2 = -((44 / 142) *log2((44 / 142)) + (38 / 142)* log2((38 / 142)) + (60 / 142) * log2((60 / 142))) 1.55784802

Information Gain for split 1

Entorpy of N - ((68 / 210) *Entorpy of N1, 1 + (142 / 210)* Entropy of N1, 2) (1.51904616) - ((68 / 210) *0.672294817 + (142 / 210)* 1.55784802)

Information Gain: 0.247948701

Entropy of N2, 1 = -((62 / 80) *log2((62 / 80)) + (18 / 80)* log2((18 / 80)) + (0 / 80) * log2((0 / 80))) 0.769192829

Entropy of N2, 2 = -((28 / 63) *log2((28 / 63)) + (11 / 63)* log2((11 / 63)) + (24 / 63) * log2((24 / 63))) 1.48999761

Entropy of N2, 3 = -((10 / 67) *log2((10 / 67)) + (21 / 67)* log2((21 / 63)) + (36 / 67) * log2((36 / 67))) 1.38787663

Information Gain for split 2

Entorpy of N - ((80 / 210) *Entorpy of N2, 1 + (63 / 210)* Entropy of N2, 2 + (67 / 210) * Entropy of N2, 3)

(1.51904616) - ((80 / 210) *0.769192829 + (63 / 210)* 1.48999761 + (67 / 210) * 1.38787663)

Information Gain: 0.336222303

2d)

Based on the information gain split 2 would be preferred, with node N2, 2 preferred to include next.

Q3

3a)

Using Laplace smoothing

Note: Ingnoring lemon data

Calculated by hand and manually counting values, hence no code for counting

Prior	Prob
P(apple)	19 / 38
P(orange)	19 / 38

Cond	Prob	Formula with Smoothing
P(Wt = 0 l apple)	3 / 21	(2 + 1) / (19 + 2)
P(Wt = 1 l apple)	18 / 21	(17 + 1) / (19 + 2)
P(Wt = 0 l orange)	8 / 21	(7 + 1) / (19 + 2)
P(Wt = 1 l orange)	13 / 21	(12 + 1) / (19 + 2)

Cond	Prob	Formula with Smoothing
P(Ht = 0 l apple)	7 / 22	(6 + 1) / (19 + 3)
P(Ht = 1 l apple)	14 / 22	(13 + 1) / (19 + 3)
P(Ht = 2 l apple)	1 / 22	(0 + 1) / (19 + 3)
P(Ht = 0 l orange)	12 / 22	(11 + 1) / (19 + 3)
P(Ht = 1 l orange)	6 / 22	(5 + 1) / (19 + 3)
P(Ht = 2 l orange)	4 / 22	(3 + 1) / (19 + 3)

Cond	Prob	Formula with Smoothing
P(Wid = 0 l apple)	12 / 22	(11 + 1) / (19 + 3)
P(Wid = 1 l apple)	8 / 22	(7 + 1) / (19 + 3)

Cond	Prob	Formula with Smoothing
P(Wid = 2 | apple)	2 / 22	(1 + 1) / (19 + 3)
P(Wid = 0 | orange)	5 / 22	(4 + 1) / (19 + 3)
P(Wid = 1 | orange)	8 / 22	(7 + 1) / (19 + 3)
P(Wid = 2 | orange)	9 / 22	(8 + 1) / (19 + 3)

3b)

Sample Number 1 P(Type = Apple | wt = 1, ht = 1, wid = 0) P (wt = 1 | apple) * P(ht = 1 | apple) * P(wid = 0 | apple) * P(apple) (18 / 21) * (14 / 22) * (12 / 22) * (19 / 38) = 0.14876033057 P(Type = Orange | wt = 1, ht = 1, wid = 0) P (wt = 1 | Orange) * P(ht = 1 | Orange) * P(wid = 0 | Orange) * P(Orange) (13 / 21) * (6 / 22) * (5 / 22) * (19 / 38) = 0.01918536009 Sample 1 estimated to be Apple Sample Number 2 P(Type = Apple | wt = 0, ht = 0, wid = 1) P (wt = 0 | apple) * P(ht = 0 | apple) * P(wid = 1 | apple) * P(apple) (3 / 21) * (7 / 22) * (8 / 22) * (19 / 38) = 0.0082644628 P(Type = Orange | wt = 0, ht = 0, wid = 1) P (wt = 0 | Orange) * P(ht = 0 | Orange) * P(wid = 1 | Orange) * P(Orange) (8 / 21) * (12 / 22) * (8 / 22) * (19 / 38) = 0.03778040141 Sample 2 estimated to be Orange Sample Number 3 P(Type = Apple | wt = 0, ht = 0, wid = 1) P (wt = 0 | apple) * P(ht = 0 | apple) * P(wid = 1 | apple) * P(apple) (3 / 21) * (7 / 22) * (8 / 22) * (19 / 38) = 0.0082644628 P(Type = Orange | wt = 0, ht = 0, wid = 1) P (wt = 0 | Orange) * P(ht = 0 | Orange) * P(wid = 1 | Orange) * P(Orange) (8 / 21) * (12 / 22) * (8 / 22) * (19 / 38) = 0.03778040141 Sample 3 estimated to be Orange Sample 4 P(Type = Apple | wt = 1, ht = 0, wid = 0) P (wt = 1 | apple) * P(ht = 0 | apple) * P(wid = 0 | apple) * P(apple) (18 / 21) * (7 / 22) * (12 / 22) * (19 / 38) = 0.07438016528 P(Type = Orange | wt = 1, ht = 0, wid = 0) P (wt = 1 | Orange) * P(ht = 0 | Orange) * P(wid = 0 | Orange) * P(Orange) (13 / 21) * (12 / 22) * (5 / 22) * (19 / 38) = 0.03837072018 Sample 4 estimated to be Apple

3c)

Sample 1 = TP Sample 2 = FN Sample 3 = TN Sample 4 = FP

```
In [ ]:
```