Part A:

1) Team Jarlsberg
   Siddhesh Madeshwar
   Zachary Noel
   Erin Dolson

2)
   a) The equation to calculate GINI is

      $$GINI(t) = 1 - \sum_{j=1}^{k} p_{j,t}^2$$

      For the first split, we have split A, B, and C into two separate categories. We have $A_{1,1}$, $A_{1,2}$, $B_{1,1}$, etc.

      For A

$A_{1,1}$	56
$A_{1,2}$	44

$P(A_{1,1}) = 56/100 = .56$
$P(A_{1,2}) = 44/100 = .44$
Gini = $1-P(A_{1,1})^2-P(A_{1,2})^2 = 1 - (.56)^2 - (.44)^2 = 1 - .3136 - .1936 =$ **0.4928**

For B

$B_{1,1}$	12
$B_{1,2}$	38

$P(B_{1,1}) = 12/50 = .24$
$P(B_{1,2}) = 38/50 = .76$
Gini = $1-P(B_{1,1})^2-P(B_{1,2})^2 = 1 - (.24)^2 - (.76)^2 = 1 - .0576 - .5776 =$ **0.3648**

For C

$C_{1,1}$	0
$C_{1,2}$	60

$P(B_{1,1}) = 0/60 = 0$
$P(B_{1,2}) = 60/60 = 1$
Gini = $1-P(B_{1,1})^2-P(B_{1,2})^2 = 1 - (0)^2 - (.1)^2 = 1 - 0- 1=$ **0**

The Gini Index is...
(100/210)*(0.4928) + (50/210)*(0.3648) + (60/210)*(0) = **0.3215**

We can now calculate the Gini for the second split. The calculations are as follows:

For A

$A_{2,1}$	62
$A_{2,2}$	28
$A_{2,3}$	10

$P(A_{2,1}) = 62/100 = .62$
$P(A_{2,2}) =28/100 = .28$
$P(A_{2,3}) = 10/100 = .10$
Gini $= 1\text{-}P(A_{2,1})^2\text{-}P(A_{2,2})^2 \text{-}P(A_{2,3})^2 = 1 - (.62)^2 - (.28)^2 - (.10)^2$
$= 1 - .3844 - 0.0784 -0.01 =$ **0.5272**

For B

$B_{2,1}$	18
$B_{2,2}$	11
$B_{2,3}$	21

$P(B_{2,1}) = 18/50 = .36$
$P(B_{2,2}) =11/50 = .22$
$PB_{2,3}) = 21/50 = .42$
Gini $= 1\text{-}P(B_{2,1})^2\text{-}P(B_{2,2})^2 \text{-}P(B_{2,3})^2 = 1 - (.36)^2 - (.22)^2 - (.42)^2$
$= 1 - 0.1296 - 0.0484 - 0.1764 =$ **0.6456**

For C

$C_{2,1}$	0
$C_{2,2}$	24
$C_{2,3}$	36

$P(C_{2,1}) = 0/60 = 0$
$P(C_{2,2}) =24/60 = ..4$
$P(C_{2,3}) = 36/60 = .6$
Gini $= 1\text{-}P(C_{2,1})^2\text{-}P(C_{2,2})^2 \text{-}P(C_{2,3})^2 = 1 - (0)^2 - (.4)^2 - (.6)^2$
$= 1 - 0 - 0.16\text{-}0.36 =$ **0.48**

The Gini Index is...
$(100/210)*(0.5272) + (50/210)*(0.6456) + (60/210)*(0.48) =$ **0.5419**

b)

$$Gain(t)_{GINI} = GINI(t) - \frac{\sum_{l=1}^{j} n_l GINI(t_l)}{\sum_{l=1}^{j} n_l}$$

First split: 0.3215 - (0.4928 + 0.3648 + 0)/3 = 0.0356
Second split: 0.5419 - (0.5272 + 0.6456 + 0.48)/3 = -0.0090
When splitting based on GINI the best split is the one with the largest gain. From the calculation, split 1 has the higher gain so that one should be used.

c) Information gain from entropy (difference between entropy before split and after split: $(Y,X_i) = H(Y) - H(Y|X_i)$

Entropy before split:

$$H(Y) = -\sum_y P(Y = y) \log P(Y = y)$$

Entropy after split for variable X_i:

$$H(Y \mid X_i) = \sum_x P(X_i = x) H(Y \mid X_i = x)$$

<u>For split 1:</u>
Before split, the entropy would be :
-[(100/210)$\log_2$(100/210) + (60/210)$\log_2$(60/210) + (50/210)$\log_2$(50/210)] = 1.51904

After split, the entropy would be calculated as shown below:
$E_{1,1}$ = -[(56/68)$\log_2$(56/68) + (12/68)$\log_2$(12/68)] = 0.67229
$E_{1,2}$ = -[(44/142$\log_2$(44/142) + (38/142)$\log_2$(38/142) + (60/142)$\log_2$(60/142)] = 1.55784

Proportion for $E_{1,1}$ = 68/210 = 0.3238
Proportion for $E_{1,2}$ = 142/210 = 0.6762

Calculate proportions = (0.67229)(0.3238) + (1.55784)(0.6762) = 1.27109
Information Gain = 1.51904 - 1.27109 = **0.24795**

<u>For split 2:</u>
Before split, the entropy would be:
-[$(100/210)\log_2(100/210) + (60/210)\log_2(60/210) + (50/210)\log_2(50/210)$] = 1.51904

Proportion for $E_{2,1}$ = 80/210 = 0.38095
Proportion for $E_{2,2}$ = 63/210 = 0.3
Proportion for $E_{2,3}$ = 67/210 = 0.319

After split, the entropy would be calculated as shown below:
$E_{2,1}$ = -[$(62/80)\log_2(62/80) + (18/80)\log_2(18/80)$] = 0.76919
$E_{2,2}$ = -[$(28/63)\log_2(28/63) + (11/63)\log_2(11/63) + (24/63)\log_2(24/63)$] = 1.48999
$E_{2,3}$ = -[$(10/67)\log_2(10/67) + (21/67)\log_2(21/67) + (36/67)\log_2(36/67)$] = 1.41571

Calculate proportions = (0.38095)(0.76919) + (0.3)(1.48999) + (0.319)(1.41571) = 1.19163
Information Gain = 1.51904 - 1.19163 = **0.32741**

d) Based on the entropy calculations above, we can conclude that the second split will be the "best" split amongst the 2 splits because it produced the closest to "pure distributions" based on the final information gain value. The information gain value for the second split is higher than that of the first split; this means that the second split produced more balanced results than the former.

3)

a) $\#(X_i = j, Y = y_k)$ = num of records with both those true
$\#(Y = y_k)$ = num of records with that condition met

$$P(X_i = j \mid Y = y_k) = \frac{\#(X_i = j, Y = y_k) + 1}{\#(Y = y_k) + |domain(X_i)|}$$

Prior	Prob.
P(apple)	(19 + 1) / 38 = .526
P(orange)	(19 + 1) / 38 = .526

Cond.	Prob.

$P(W_t=0 \mid apple)$	(17 + 1) / (19 + 2) = .857
$P(W_t=1 \mid apple)$	(2 + 1) / (19 + 2) = .143
$P(W_t=0 \mid orange)$	(12 + 1) / (19 + 2) = .619
$P(W_t=1 \mid orange)$	(7 + 1) / (19 + 2) = .381

Cond.	Prob.
$P(H_t=0 \mid apple)$	(6 + 1) / (19 + 3) = .318
$P(H_t=1 \mid apple)$	(13 + 1) / (19 + 3) = .636
$P(H_t=2 \mid apple)$	(0 + 1) / (19 + 3) = .045
$P(H_t=0 \mid orange)$	(11 + 1) / (19 + 3) = .545
$P(H_t=1 \mid orange)$	(5 + 1) / (19 + 3) = .273
$P(H_t=2 \mid orange)$	(3 + 1) / (19 + 3) = .182

Cond.	Prob.
$P(W_{id}=0 \mid apple)$	(11 + 1) / (19 + 3) = .545
$P(W_{id}=1 \mid apple)$	(7 + 1) / (19 + 3) = .364
$P(W_{id}=2 \mid apple)$	(7 + 1) / (19 + 3) = .364
$P(W_{id}=0 \mid orange)$	(4 + 1) / (19 + 3) = .227
$P(W_{id}=1 \mid orange)$	(7 + 1) / (19 + 3) = .364
$P(W_{id}=2 \mid orange)$	(8 + 1) / (19 + 3) = .409

b)

Sample Num.	Test Data	Predicted Class
1	[1,1,1,0]	apple
2	[1,0,0,1]	orange
3	[2,0,0,1]	orange
4	[2,1,0,0]	orange

$P(\text{Fruit} = \text{apple} \mid 1,1,0) \propto P(W_t{=}1 \mid \text{apple})P(H_t{=}1 \mid \text{apple})P(W_{id}{=}0 \mid \text{apple})P(\text{apple})$
$= .143 * .636 * .545 * .526$
$= .0261$

$P(\text{Fruit} = \text{orange} \mid 1,1,0) \propto P(W_t{=}1 \mid \text{orange})P(H_t{=}1 \mid \text{orange})P(W_{id}{=}0 \mid \text{orange})P(\text{orange})$
$= .381 * .273 * .227 * .526$
$= .0124$

$P(\text{Fruit} = \text{apple} \mid 0,0,1) \propto P(W_t{=}0 \mid \text{apple})P(H_t{=}0 \mid \text{apple})P(W_{id}{=}1 \mid \text{apple})P(\text{apple})$
$= .857 * .318 * .364 * .526$
$= .0522$

$P(\text{Fruit} = \text{orange} \mid 0,0,1) \propto P(W_t{=}0 \mid \text{orange})P(H_t{=}0 \mid \text{orange})P(W_{id}{=}1 \mid \text{orange})P(\text{orange})$
$= .619 * .545 * .364 * .526$
$= .0646$

$P(\text{Fruit} = \text{apple} \mid 1,0,0) \propto P(W_t{=}1 \mid \text{apple})P(H_t{=}0 \mid \text{apple})P(W_{id}{=}0 \mid \text{apple})P(\text{apple})$
$= .143 * .318 * .545 * .526$
$= .0130$

$P(\text{Fruit} = \text{orange} \mid 1,0,0) \propto P(W_t{=}1 \mid \text{orange})P(H_t{=}0 \mid \text{orange})P(W_{id}{=}0 \mid \text{orange})P(\text{orange})$
$= .381 * .545 * .227 * .526$
$= .0248$

c)

Sample Num.	Test Data	Predicted Class	Report
1	[1,1,1,0]	apple	TP
2	[1,0,0,1]	orange	FN
3	[2,0,0,1]	orange	TN
4	[2,1,0,0]	orange	TN