HOMEWORK 2

CS 422 – Data Mining

Rutul Mehta
A20476293

Exercise 2

A) Compute the Gini index for the overall collection of training examples.

- Gini index of Overall Collection:

$$= 1 - \left[({}^{10}/_{20})^{2} + ({}^{10}/_{20})^{2}\right]$$
$$= 1 - 0.5$$
$$= 0.5$$

B) Compute the Gini index for the "Customer ID" attribute.

- For each Customer ID, Gini value is 0. So, the overall Gini for customer Id is 0.

C) Compute the Gini index for the " Gender " attribute.

Gender	C_0	C_1
Male	6	4
Female	4	6

- $Gini_{Male}$ = $1 - [\left({}^{6}/_{10}{}^{2}\right) + ({}^{4}/_{10})^{2}]$

  $=1 - [(0.36\ + 0.16)]$

  $= 1 - 0.52$

  = 0.48

- $Gini_{Female}$ = $1 - [\left({}^{4}/_{10}{}^{2}\right) + ({}^{6}/_{10})^{2}]$

  $=1 - [(0.16\ + 0.36)]$

  $= 1 - 0.52$

  = 0.48

- $Gini_{overall} = \frac{10}{20} * (0.48) + \frac{10}{20} * (0.48)$

  $=$ 0.24 + 0.24

  $=$0.48

D) Compute the Gini index for the attribute using multiway split.

Car Type	C_0	C_1
Family	1	3
Sports	8	0
Luxury	1	7

- $Gini_{Family}$ = $1-[\left({}^{1}/_{4}{}^{2}\right)+({}^{3}/_{4})^{2}]$
  =$1-[(0.0625\ +0.5625)]$
  = $1-0.625$
  = 0.375

- $Gini_{Sports}$ = $1-[\left({}^{8}/_{8}{}^{2}\right)+({}^{0}/_{8})^{2}]$
  =$1-[(1)]$
  = 0

- $Gini_{Luxary}$ = $1-[\left({}^{1}/_{8}{}^{2}\right)+({}^{7}/_{8})^{2}]$
  =$1-[(0.01563\ +0.7656)]$
  = 0.2187

- $Gini_{Overall}$ =$\frac{4}{20}*(0.375)+\frac{8}{20}*(0)+\frac{8}{20}*(0.2187)$
  =$0.075+0+0.0875$
  = 0.1625

E) Compute the Gini index for the “Shirt Size” attribute using multiway split.

Shirt Size	C_0	C_1
Small	3	2
Medium	3	4
Large	2	2
Extra Large	2	2

- $Gini_{Small}$ = $1-[\left({}^{3}/_{5}{}^{2}\right)+({}^{2}/_{5})^{2}]$
  = $1-[(0.36\ +0.16)]$
  = 0.48

- $Gini_{Medium}$ = $1-[\left({}^{3}/_{7}{}^{2}\right)+({}^{4}/_{7})^{2}]$
  = $1-[(0.1837+0.3265)]$
  = 0.4898

- $Gini_{Large}$ = $1-[\left({}^{2}/_{4}{}^{2}\right)+({}^{2}/_{4})^{2}]$
  = $1-[(0.25+0.25)]$
  = 0.5

- $Gini_{Extra\ Large}$ = $1-[\left({}^{2}/_{4}{}^{2}\right)+({}^{2}/_{4})^{2}]$
  = $1-[(0.25+0.25)]$
  = 0.5

- $Gini_{Overall}$ = $\frac{5}{20}*(0.48)+\frac{7}{20}*(0.4898)+\frac{4}{20}*(0.5)+\frac{4}{20}*(0.5)$
  =$0.12+0.1714+0.1+0.1$
  = 0.4914

F) Which attribute is better “Gender”,” Car Type “or “Shirt Size”?

- Car Type is better attribute among all the three as it has the lowest Gini.

G) Explain why “Customer ID” should not be used as the attribute test condition even though it has the lowest Gini.

- The predictive power of “Customer ID” is very low so it should not be used as attribute test condition.

Exercise 3

A) What is the entropy of this collection of training examples with respect to the class attribute?

- According to the target class attribute,
  - P (+) = $\frac{4}{9}$
  - P (-) = $\frac{5}{9}$
- Entropy:

$$
\begin{aligned}
&= -\frac{4}{9}*log_2\left(\frac{4}{9}\right)-\frac{5}{9}*log_2\left(\frac{5}{9}\right)\\
&= -\frac{4}{9}*(-1.1699)-\frac{5}{9}*(-0.8480)\\
&= 0.5199+0.4711\\
&= 0.9910
\end{aligned}
$$

B) What are the information gains of a_1 and a_2 relative to these training examples?

a_1	+	-
T	3	1
F	1	4

- Entropy T:

$$= -\frac{3}{4} * log_2\left(\frac{3}{4}\right) - \frac{1}{4} * log_2\left(\frac{1}{4}\right)$$
$$= -\frac{3}{4} * (-0.4150) - \frac{1}{4} * (-2)$$
$$= 0.3112 + 0.5$$
$$= 0.8112$$

- Entropy F:

$$= -\frac{1}{5} * log_2\left(\frac{1}{5}\right) - \frac{4}{5} * log_2\left(\frac{4}{5}\right)$$
$$= -\frac{1}{5} * (-2.3219) - \frac{4}{5} * (-0.3219)$$
$$= 0.4644 + 0.2575$$
$$= 0.7219$$

- Information gain for a_1:

$$= \frac{4}{9} * (0.8112) + \frac{5}{9} * (0.7219)$$
$$= 0.3605 + 0.4011$$
$$= 0.7616$$

a_2	+	-
T	2	3
F	2	2

- Entropy T:

$$= -\frac{2}{5} * log_2\left(\frac{2}{5}\right) - \frac{3}{5} * log_2\left(\frac{3}{5}\right)$$
$$= 0.5288 + 0.4422$$
$$= 0.971$$

- Entropy F:

$$= -\frac{2}{4} * log_2\left(\frac{2}{4}\right) - \frac{2}{4} * log_2\left(\frac{2}{4}\right)$$
$$= -\frac{2}{4} * (-1) - \frac{2}{4} * (-1)$$
$$= 0.5 + 0.5$$
$$= 1$$

- Information gain for a_1:

$$= \frac{5}{9} * (0.971) + \frac{4}{9} * (1)$$

$= 0.5394 + 0.4444$
$= 0.9838$

- So, the information gain for a_2 is:
  $= 0.9910 - 0.9839$
  $= 0.0072$

C) For a_3, which is a continuous attribute, compute the information gain for every possible split.

A3 is continuous attribute.

Total entropy: $-\ (4/\ 9)\ log2\ (\ 4\ /9\ )\ -\ (\ 5/\ 9\ )\ log2\ (\ 5\ /9\ )\ =\ 0.9911$
Information Gain in Split $2\ =\ 0.991\ -\ 0.846\ =\ 0.415$
Information Gain in Split $3.5\ =\ 0.991\ -\ 0.9884\ =\ 0.0026$
Information Gain in Split $4.5\ =\ 0.991\ -\ 0.9782\ =0.0128$
Information Gain in Split $5.5\ =\ 0.991\ -\ 0.9838 =\ 0.0072$
Information Gain in Split $6.5\ =\ 0.991\ -\ 0.9727\ =0.0183$
Information Gain in Split $7.5\ =\ 0.991\ -\ 0.8888\ =\ 0.1022$

The highest gain of information is on split 2.

D) What is the best split (among a_1, a_2 and a_3,) according to the information gain?

- Considering the information gain a_1 produces best split.

E) What is the best split (between a_1 and a_2) according to the misclassification error rate?

- For attribute a_1 :
  - Error rate = $\frac{2}{9}$
- For attribute a_2
  - Error rate = $\frac{2}{9}$
- a_1 has lower error rate so, it is suitable for split.

F) What is the best split (between a_1 and a_2) according to the Gini index?

- For attribute a_1 :

- $Gini_T$ $= 1 - [({}^{3}/_{4})^2 + ({}^{1}/_{4})^2]$
  $= 1 - [(0.5625\ + 0.0625)]$
  $= 0.375$

- $Gini_F = 1 - [({}^{1}/_{5})^2 + ({}^{4}/_{5})^2]$
  $= 1 - [(0.04 \ + 0.64)]$
  $= 0.32$

- $Gini_{a1} = \frac{4}{9} * (0.375) + \frac{5}{9} * (0.32)$
  $= 0.1667 + 0.1778$
  $= 0.3444$

- For attribute a_2 :

- $Gini_T = 1 - [({}^{2}/_{5})^2 + ({}^{3}/_{5})^2]$
  $= 1 - [(0.16 \ + 0.36)]$
  $= 0.48$

- $Gini_F = 1 - [({}^{2}/_{4})^2 + ({}^{2}/_{4})^2]$
  $= 1 - [(0.25 \ + 0.25)]$
  $= 0.5$

- $Gini_{a2} = \frac{4}{9} * (0.48) + \frac{5}{9} * (0.5)$
  $= 0.2133 + 0.2777$
  $= 0.4911$
- Gini index a_1 has a smaller value so it can produce a better split.

Exercise 5

a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

- Splitting **A**

Class	A=T	A=F
+	4	0
-	3	3

- Splitting **B**

Class	A=T	A=F
+	3	1
-	1	5

- Entropy for Class ("+") = $-\frac{2}{5}log_2\left({}^{2}/_{5}\right)$

  =$-\frac{2}{5}*(-1.3219)$ = 0.5287
- Entropy for Class ("- ") = $-\frac{6}{10}log_2({}^{6}/_{10})$

  =$-\frac{6}{10}*(-0.7369)$ = 0.4421
- Overall entropy = Entropy for Class ("+") + Entropy for Class ("- ")

  = 0.5287 + 0.4421 = 0.9708
- Entropy when A=T: $= -\frac{4}{7}log_2({}^{4}/_{7})$ + $-\frac{3}{7}log_2({}^{3}/_{7})$

  $= -\left({}^{4}/_{7}\right) * -0.8074$ + $-\left({}^{3}/_{7}\right) * -1.2226$

  $=$ 0.4613 + 0.5239

  $= 0.9852$
- Entropy when A=F: $= -\frac{3}{3}log_2({}^{3}/_{3})$ + $-\frac{0}{3}log_2({}^{0}/_{3})$

  $= 0 + 0$

  $= 0$
- Information gain on A = Overall Entropy – [7/10*Entropy when A=T + 3/10*Entropy when A=F]

  $= 0.9710 - [\frac{7}{10} * 0.9852 +$ $\frac{3}{10}$*0]

  $= 0.9710 - 0.6896$

  $= 0.2814$
- Entropy when B=T: $= -\frac{3}{4}log_2({}^{3}/_{4})$ + $-\frac{1}{4}log_2({}^{1}/_{4})$

  $= -0.75 * (-0.4150) + (-0.25 * -2)$

  $= 0.3112 + 0.5$

  $= 0.8112$
- Entropy when B=F: $= -\frac{1}{6}log_2({}^{1}/_{6})$ + $-\frac{5}{6}log_2({}^{5}/_{6})$

  $= -0.1667 * (-2.5846) + [-0.8333 * (-0.2630)]$

  $= 0.4308 + 0.2191$

  $= 0.6499$
- Information gain on B = Overall Entropy – [4/10*Entropy when B=T + 6/10*Entropy when B=F]

  $= 0.9710 - [\frac{4}{10} * 0.8112 +$ $\frac{6}{10}$*0.6499]

  $= 0.9710 - 0.7143$

  $= 0.2567$

b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

- $Gini_{overall}$ $= 1 - [({}^{4}/_{10})^2 + ({}^{6}/_{10})^2]$

  $= 1 - [(0.4)^2 + (0.6)^2]$

  $= 1 - [0.16 + 0.36]$

  $= 1 - 0.52$

  $= 0.48$

- Gini when A=T:

  $= 1 - [({}^{4}/_{7})^2 + ({}^{3}/_{7})^2$

  $= 1 - [0.3265 + 0.1836]$

  $= 1 - 0.5101$

  $= 0.4898$

- Gini when A=F:

  $= 1 - [({}^{3}/_{3})^2 + ({}^{0}/_{3})^2$

  $= 1 - [1 - 0]$

  $= 1 - 1$

  $= 0$

- Gain in Gini on Splitting A $= Gini_{overall} - [\frac{7}{10} * \text{Gini when A} = \text{T} + \frac{3}{10} * \text{ Gini when A} = \text{F}$

  $= 0.48 - [\frac{7}{10} * 0.4898 + \frac{3}{10} * 0]$

  $= 0.48 - [0.3428]$

  $= 0.1372$

- Gini when B=F:

  $= 1 - [({}^{1}/_{6})^2 + ({}^{5}/_{6})^2$

  $= 1 - [0.0277 + 0.6944]$

  $= 1 - 0.7221 = 0.2778$

- Gini when B=T:

$$= 1 - [({}^{1}/_{4})^2 + ({}^{3}/_{4})^2$$

$$= 1 - [0.0625 + 0.5625]$$

$$= 1 - \ 0.625$$

$$= 0.3750$$

- Gain in Gini on Splitting B $= Gini_{overall} - [\frac{4}{10} * \text{Gini when B} = \text{T} + \frac{6}{10} * \ \text{Gini when B} = \text{F}$

$$= 0.48 - [\frac{4}{10} * 0.3750 + \frac{6}{10} * 0.2778]$$

$$= 0.48 - [0.3166]$$

$$= 0.1634$$

- We will be splitting B as Gini gain in B is higher than A.

c) Yes, It is possible. We can prove from the above 2 problems (a) and (b). overall gain might be different based on the different criterion (gini/entropy).

Exercise 6

A) Calculate the Gini index and misclassification error rate of the parent node P.

- $Gini_P = 1 - [\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2]$

$$= 0.42$$

Error P = $\frac{3}{10} = 0.3$

B) Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?

- Gini of C1 $= 1 - [\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2]$

$$= 1 - 1$$
$$= 0$$

- Gini of C2 $= 1 - [\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2]$

$$= 1 - [0.3262 + 0.1836$$
$$= 0.4899$$

- Weighted Gi $= \frac{3}{10} * 0 + \frac{7}{10} * 0.4899$

$= 0.34293$

- Based on the difference between Gini index of Parent and child (0.42-0.34 = 0.08) this attribute can be chosen for splitting.

C) Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?

- Error C1 = 0/3 = 0
- Error C2 = 3/7 = 0.4285
- Weighted error = $\frac{3}{10} * 0 + \frac{7}{10} * 0.4285$ = 0.3

Exercise 7

A) Compute a two-level decision tree using the greedy approach described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

- Splitting level 1

For X

X=0	X=1
C1 = 60	C1 = 40
C2 = 60	C2 = 40

The error rate of using attribute x is (60+40)/200 = 0.5

For Y

Y=0	Y=1
C1 = 40	C1 = 60
C2 = 60	C2 = 40

The error rate of using attribute x is (40+40)/200 = 0.4

For Z

Y=0	Y=1
C1 = 30	C1 = 70
C2 = 70	C2 = 30

The error rate of using attribute x is (30+30)/200 = 0.3

- For Z=0 for attribute X and Y

X=0	X =1

C1 = 15	C1 = 15
C2 = 45	C2 = 25

Y=0	Y = 1
C1 = 15	C1 = 15
C2 = 45	C2 = 25

Error rate for X and Y = 15+15/100 = 0.3

- For Z=1 for attribute X and Y

X=0	X =1
C1 = 45	C1 = 25
C2 = 15	C2 = 15

Y=0	Y = 1
C1 = 45	C1 = 25
C2 = 15	C2 = 15

Error rate for X and Y = 15+15/100 = 0.3

Overall rate is 15+15+15+15/200 = 0.3

B) Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

- For X=0 for attribute Z and Y

Z=0	Z =1
C1 = 15	C1 = 45
C2 = 45	C2 = 15

Y=0	Y = 1
C1 = 5	C1 = 55
C2 = 55	C2 = 5

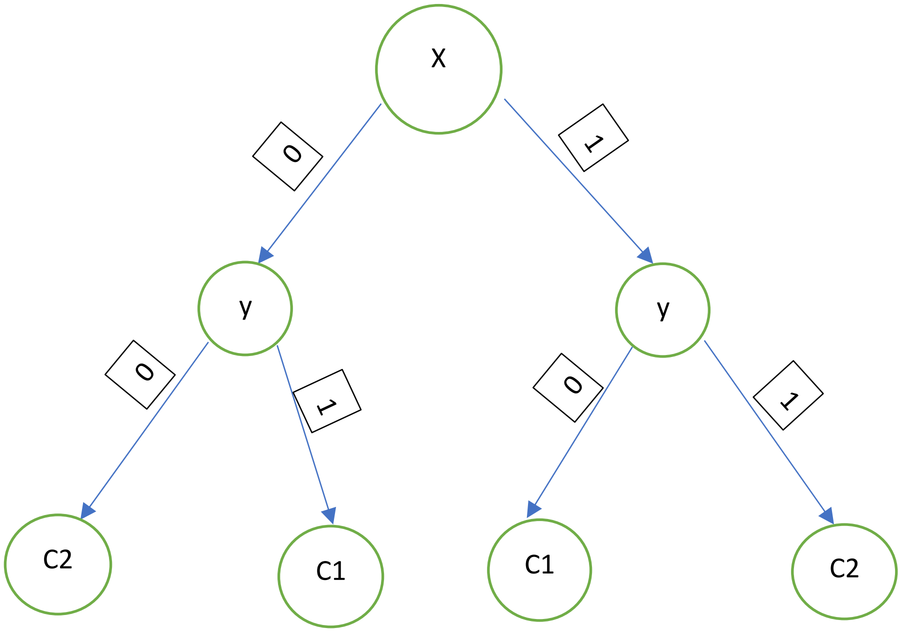
Error rate for Z and Y = 5+5/120 = 0.0833
=15+15/120 = 0.25

- For X=1 for attribute Z and Y

Z=0	Z =1
C1 = 15	C1 = 25
C2 = 25	C2 = 15

Y=0	Y = 1
C1 = 35	C1 = 5
C2 = 5	C2 = 35

Error rate for Z and Y = 5+5/80 = 0.125
= 15+15/80 = 0.375



- Error rate of induced tree is 10+10/200 = 0.1

C) According to the result of above two problems, error rate of part (a) is larger than part (b). So, we can say that greedy heuristic does not always produce an optimal solution.

Exercise 8

A) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

- Error rate of the data without pointing on any attribute is:

  = 1 − max (50/100, 50/100) = 0.50

- After splitting on attribute, A, the gain in error rate is:

Class	A=T	A=F
+	25	25
-	0	50

For A=T: = 1 − max $(\frac{25}{25}, \frac{0}{25})$

= 0

For A=F = 1-max $(\frac{25}{75}, \frac{50}{75})$

= 1 - $^{50}/_{75}$

= 0.33

Gain rate for A = 0.5 -$[\frac{25}{100} * 0 + \frac{75}{100} * 0.33]$

= 0.5 - 0.2475

= 0.2525

- After splitting on attribute, B, the gain in error rate is:

Class	B=T	B=F
+	30	30
-	20	30

For B=T: = 1 − max $(\frac{30}{50}, \frac{20}{50})$

= $1 - {}^{30}/_{50}$

= 0.4

For B=F = 1 - max $(\frac{20}{50}, \frac{30}{50})$

= 1 - $^{30}/_{50}$

= 0.4

Gain rate for B = 0.5 -$[\frac{50}{100} * 0.4 + \frac{50}{100} * 0.4]$

= 0.5 – [0.2 + 0.2]

= 0.1

- After splitting on attribute, C, the gain in error rate is:

Class	C=T	C=F
+	25	25
-	25	25

For C=T: = 1 – max $(\frac{25}{50}, \frac{25}{50})$

= $1 - {}^{25}/_{50}$

= 0.5

For C=F = 1 - max $(\frac{25}{50}, \frac{25}{50})$

= 1 - ${}^{20}/_{50}$

= 0.5

Gain rate for C = 0.5 -$[\frac{50}{100} * 0.5 + \frac{50}{100} * 0.5]$

= 0.5 – [0.25 + 0.25]

= 0

- The attribute A has highest gain.

B) Repeat for the two children of the root node.

- For A=F

B	C	+	-
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

Classification error A=F:

$= 1 - \max\,({}^{25}/_{75}, {}^{50}/_{75})$

$= 1 - ({}^{50}/_{75})$

$= 1 - {}^{2}/_{3}$

$= 1 - 0.6666$

$= 0.3333$

- Splitting B

	B=T	B=F
+	25	0
-	20	30

For B=T:

$= 1 - \max\,({}^{25}/_{45}, {}^{20}/_{45})$

$= 1 - {}^{25}/_{45}$

$= 1 - 0.5555$

$= 0.4444$

For B=F:

$= 1 - \max\,({}^{0}/_{30}, {}^{30}/_{30})$

$= 1 - {}^{30}/_{30}$

$= 1 - 1$

$= 0$

- Change in information gain after splitting B:

= Classification error (A=F) – $[{}^{45}/_{75} * Gain\ for\ (B = T) + {}^{20}/_{75} * Gain\ for\ (B = F)$

= $0.3333 - [{}^{45}/_{75} * 0.4444\ +\ {}^{20}/_{75} * 0]$

= $0.3333 - [0.2666 + 0]$

= 0.0666

- Splitting C

	C=T	C=F
+	0	25
-	25	25

For C=T:

$= 1 - \max\,({}^{0}/_{25}, {}^{25}/_{25})$

$= 1 - {}^{25}/_{25}$

$= 1 - 1$

$= 0$

For C=F:

$= 1 - \max\left({}^{25}/_{50}, {}^{25}/_{50}\right)$
$= 1 - {}^{25}/_{50}$
$= 1 - 0.5$
$= 0.5$

- Change in information gain after splitting C:

  = Classification error (A=F) – $[{}^{25}/_{75} * Gain\ for\ (C = T) + {}^{50}/_{75} * Gain\ for\ (C = F)$
  = $0.3333 - [{}^{45}/_{75} * 0\ +\ {}^{20}/_{75} * 0.5]$
  = $0.3333 - [0 + 0.3333]$
  = 0

- The max gain is when we split attribute B

C) How many instances are misclassified by the resulting decision tree?

- 20 are misclassified, therefore error rate = $({}^{20}/_{100})$.

D) Repeat parts (a), (b), and (c) using C as the splitting attribute.

- For C=T
- The error rate before splitting is:

  $= 1 - \max\left({}^{25}/_{50}, {}^{25}/_{50}\right)$
  $= 1 - {}^{25}/_{50}$
  $= 1 - 0.5$
  $= 0.5$
- Splitting attribute, A:

	A=T	A=F
+	25	0
-	0	25

For A=T:
$= 1 - \max\left({}^{25}/_{25}, {}^{0}/_{25}\right)$
$= 1 - {}^{25}/_{25}$
$= 1 - 1$
$= 0$

For A=F:

$= 1 - \max\left({}^{25}/_{25}, {}^{0}/_{25}\right)$

$= 1 - {}^{25}/_{25}$

$= 1 - 1$

$= 0$

Change in gain after splitting A:

$= \text{The error rate before splitting} - [{}^{25}/_{50} * \text{For } (\text{A} = \text{T}) + {}^{25}/_{50} * \text{For } (\text{A} = \text{F})]$

$= 0.5 - [{}^{25}/_{50} * 0 + {}^{25}/_{50} * 0]$

$= 0.5 - 0$

$= 0.5$

- Splitting attribute, B:

	B=T	B=F
+	5	20
-	20	5

For B=T:

$= 1 - \max\left({}^{5}/_{25}, {}^{20}/_{25}\right)$

$= 1 - {}^{20}/_{25}$

$= 1 - 0.8$

$= 0.2$

For B=F:

$= 1 - \max\left({}^{20}/_{25}, {}^{0}/_{25}\right)$

$= 1 - {}^{20}/_{25}$

$= 1 - 0.8$

$= 0.2$

Change in gain after splitting B :

$= \text{The error rate before splitting} - [{}^{25}/_{50} * \text{For } (\text{B} = \text{T}) + {}^{25}/_{50} * \text{For } (\text{B} = \text{F})]$

$= 0.5 - [{}^{25}/_{50} * 0.2 + {}^{25}/_{50} * 0.2]$

$= 0.5 - 0.2$

$= 0.3$

- Attribute A will be chosen to split as it has more information gain after splitting.

- For C=F

- The error rate before splitting is:

$= 1 - \max\left({}^{25}/_{50}, {}^{25}/_{50}\right)$

$= 1 - {}^{25}/_{50}$

$= 1 - 0.5$

$= 0.5$

- Splitting attribute, A:

	A=T	A=F
+	0	25
-	0	25

For A=T:

$= 0$

For A=F:

$= 1 - \max\left({}^{25}/_{50}, {}^{25}/_{50}\right)$

$= 1 - {}^{25}/_{50}$

$= 1 - 0.5$

$= 0.5$

Change in gain after splitting A for C=F:

$= \text{The error rate before splitting} - [{}^{0}/_{50} * \text{For (A = T)} + {}^{50}/_{50} * \text{For (A = F)}]$

$= 0.5 - [{}^{0}/_{50} * 0 + {}^{50}/_{50} * 0.5]$

$= 0.5 - 0.5$

$= 0$

- Splitting attribute, B:

	B=T	B=F
+	25	0
-	0	25

For B=T:

$= 1 - \max\left({}^{25}/_{25}, {}^{0}/_{25}\right)$

$= 1 - {}^{25}/_{25}$

$= 1 - 1$

$= 0$

For B=F:

$= 1 - \max\left({}^{25}/_{25}, {}^{0}/_{25}\right)$

$= 1 - {}^{25}/_{25}$

$= 1 - 1$

$= 0$

Change in gain after splitting B for C=T :

$= \text{The error rate before splitting} - [{}^{25}/_{50} * \text{For } (\text{B} = \text{T}) + {}^{25}/_{50} * \text{For } (\text{B} = \text{F})]$

$= 0.5 - [{}^{25}/_{50} * 0 + {}^{25}/_{50} * 0]$

$= 0.5$

- Attribute B will be used as information gain is more after splitting the attribute.

E) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.

- The greedy method may get stuck in local minima which may lead to wrong results.

Exercise 12

A) Based on the accuracies shown in Table 3.7, which classification model would you expect to have better performance on unseen instances?

- Training accuracy of T100 on dataset A is very high, but its test accuracy on dataset B is very low. Because of this gap between training and test accuracies T100 suffer from overfitting.
- On dataset A, Tree T10 has low training accuracy, but the test accuracy of T10 on dataset B is kind of same. That is why, T10 does not suffer from overfitting.

B) Now, you tested and on the entire data set and found that the classification accuracy of on data set is 0.85, whereas the classification accuracy of on the data set is 0.87. Based on this new information and your observations from Table 3.7, which classification model would you finally choose for classification?

- We are testing accuracy of T100 with dataset that is used to train the tree. That is why, the performance of tree T100 is not that much better compared to T10 and T10 is a better model for classification.

Reference:

https://datascience.stackexchange.com/questions/40900/whats-the-difference-between-sklearn-f1-score-micro-and-weighted-for-a-mult

https://stackoverflow.com/questions/55740220/macro-vs-micro-vs-weighted-vs-samples-f1-score

https://discuss.analyticsvidhya.com/t/decision-tree-with-continuous-variables/201/2