Assignment 1

1. You are given a transaction data shown below from a fast food restaurant. For simplicity, we assign the meal items short names [M1-M5]. For all the min_sup=2/9 and min_conf=7/9. Apply Apriori and identify all k-frequent itemsets. Find all the strong association rules and note their confidence.

Meal Item	List of Items
Order 1	{M1, M2, M5}
Order 2	{M2, M4}
Order 3	{M2, M3}
Order 4	{M1, M2, M4}
Order 5	{M1, M3}
Order 6	{M2, M3}
Order 7	{M1, M3}
Order 8	{M1, M2, M3, M5}
Order 9	{M1, M2, M3}

2. Define maximal and closed frequent itemset. Identify the above from the database:

Transaction ID	Items
T1	{A, C, T, W}
T2	{C, D, W}
T3	{A, C, T, W}
T4	{A, C, D, W}
T5	{A, C, D, T, W}
T6	{ C, D, T }

3. Consider the database d shown in the table below. Consider min_sup =60% and min_conf=80%. Apply Apriori and identify all k-frequent itemsets. Find all the strong association rules and note their confidence.

TID	Items
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

4. Consider the transaction database as follows and indicate closed and maximal frequent item sets

TID	Items
1	{A, B, C}
2	{A, B, C, D}
3	{B, C, E}
4	{A, C, D, E}
5	{D, E}

5. Draw the decision tree for the following dataset:

Color	Type	Doors	Tires	Class
Red	SUV	2	Whitewall	+
Blue	Minivan	4	Whitewall	-
Green	Car	4	Whitewall	-
Red	Minivan	4	Blackwall	-
Green	Car	2	Blackwall	+
Green	SUV	4	Blackwall	-
Blue	SUV	2	Blackwall	-
Blue	Car	2	Whitewall	+
Red	SUV	2	Blackwall	-
Blue	Car	4	Blackwall	-

6. Construct a decision tree for the following data:

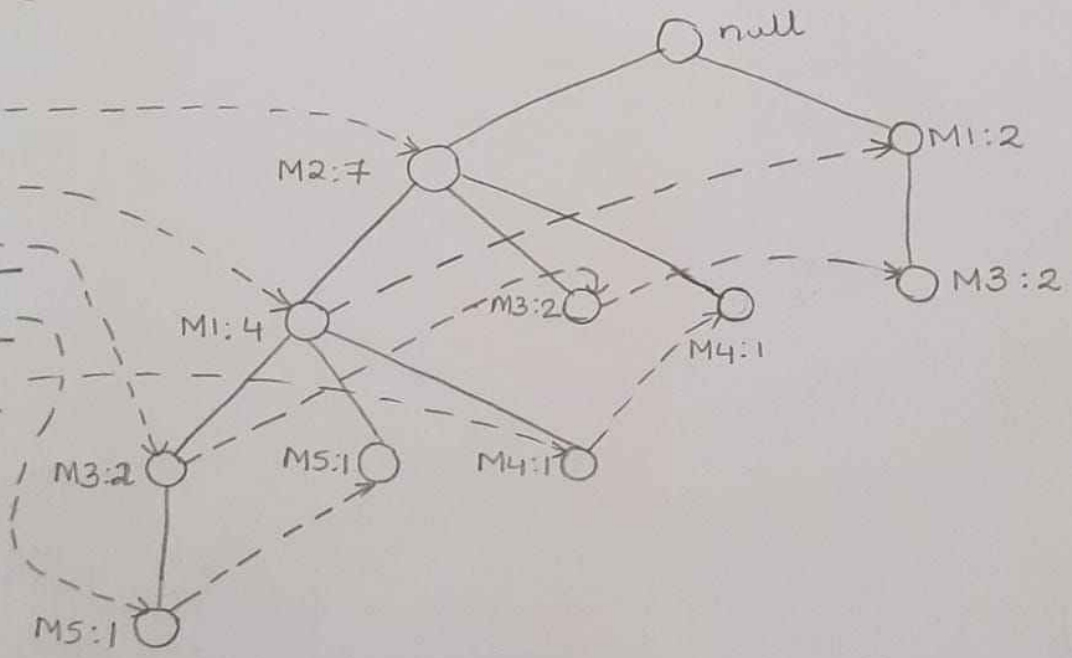
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

(2)

FP-Tree

Header Table

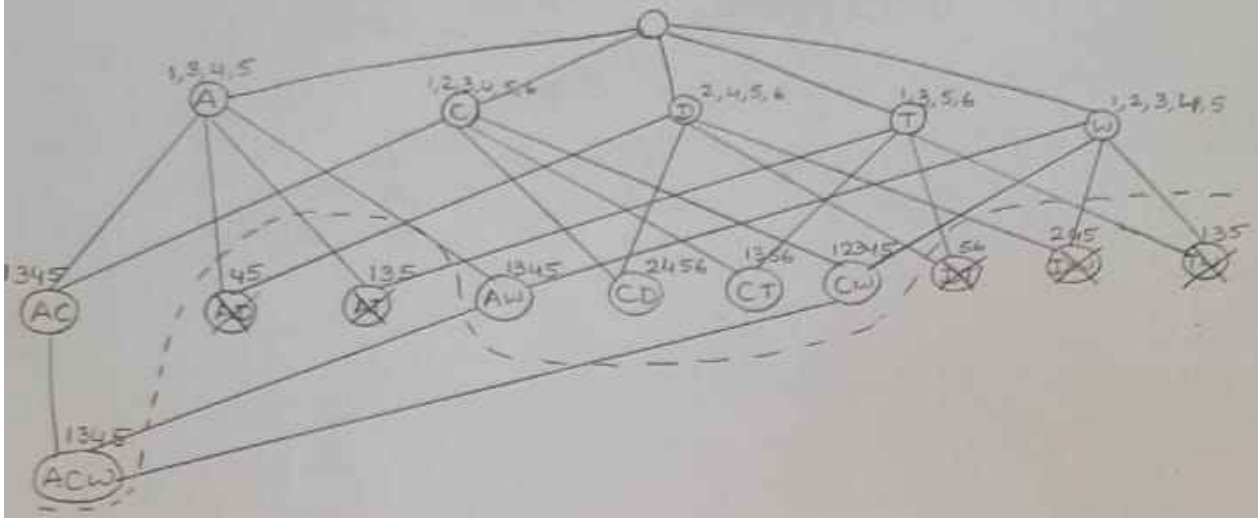
Item	Pointer
M2	
M1	
M3	
M5	
M4	



2.

Transaction ID	Items
T1	{A,C,T,W}
T2	{C,D,W}
T3	{A,C,T,W}
T4	{A,C,D,W}
T5	{A,C,D,T,W}
T6	{C,D,T}

Assume min-sup = 4



Closed frequent itemset: {A,C,W}, {C,D}, {C,T}{C,W}, {C}, {W},
Maximal frequent itemset: {A,C,W}, {C,D}, {C,T}

3.

TID	Items
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,A,K,E}
T400	{M,U,C,K,Y}
T500	{C,O,O,K,I,E}

C_1 = {{M}, {O}, {N}, {K}, {E}, {Y}, {D}, {A}, {U}, {C}, {I}}

min_sup = 60%
min_conf = 80%

Itemset	Support
{M}	3
{O}	3
{N}	2 X
{K}	5
{E}	4
{Y}	3
{D}	1 X
{A}	1 X
{U}	1 X
{C}	2 X
{I}	1 X

L_1 = {{M}, {O}, {K}, {E}, {Y}}
C_2 = {{M,O}, {M,K}, {M,E}, {M,Y}, {O,K}, {O,E}, {O,Y}, {K,E}, {K,Y}, {E,Y}}

Itemset	Support
{M,O}	1 X
{M,K}	3
{M,E}	2 X
{M,Y}	2 X
{O,K}	3
{O,E}	3
{O,Y}	2 X
{K,E}	4
{K,Y}	3
{E,Y}	2 X

∴ L_2 = {{M,K}, {O,K}, {O,E}, {K,E}, {K,Y}}
C_3 = {O,K,E}

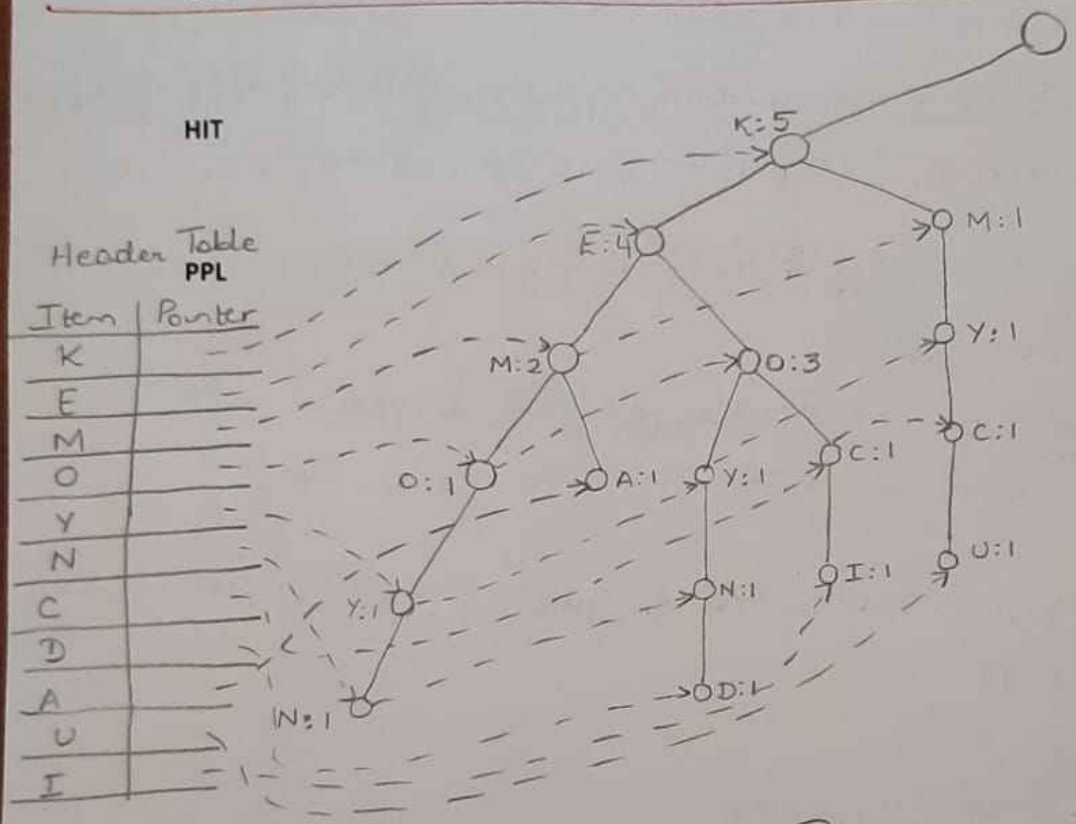
Itemset	Support
{O,K,E}	3

∴ $L_3 = C_3$

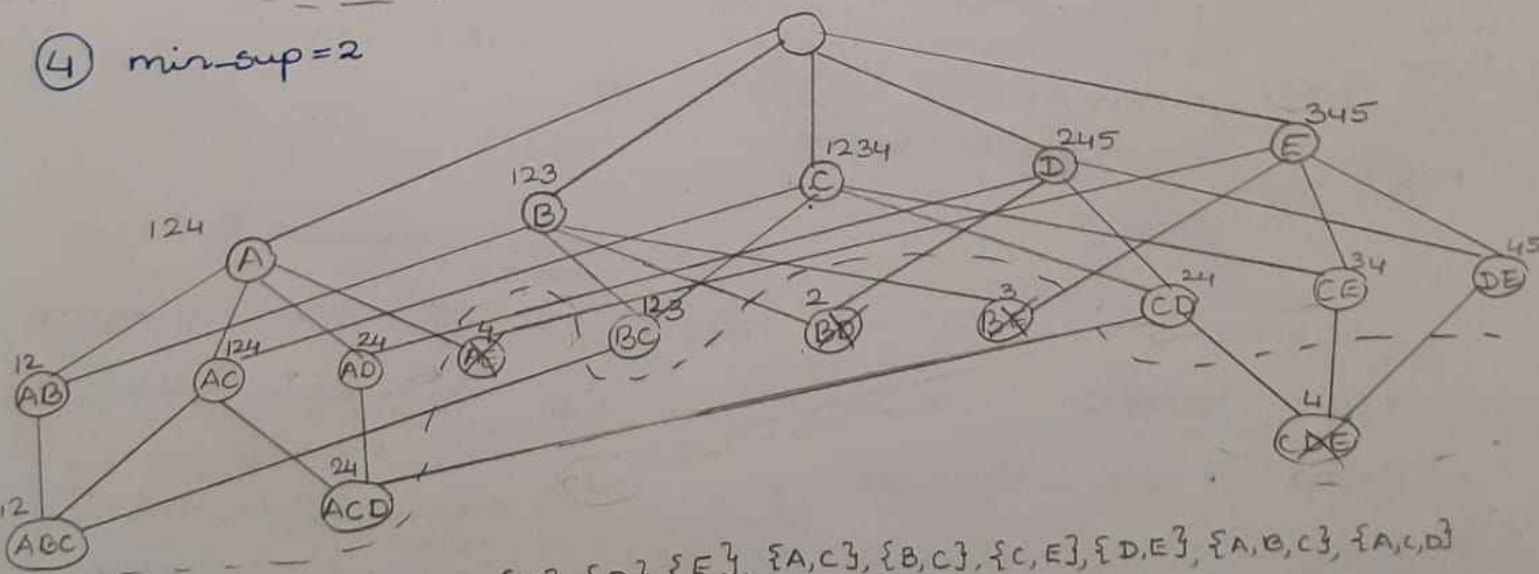
Ms. Diana Miranda
Coordinator for IBM Skill Build
GEC

(4)

HIT



(4) min_sup = 2



closed frequent itemsets: {C}, {D}, {E}, {A,C}, {B,C}, {C,E}, {D,E}, {A,B,C}, {A,C,D}

maximal frequent itemsets: {A,B,C}, {A,C,D}, {C,E}, {D,E}

5.

	COLOR	TYPE	DOORS	TYRES	CLASS
1	Red	SUV	2	WhiteWall	+
2	Blue	Minivan	4	Whitewall	—
3	Green	Car	4	Whitewall	—
4	Red	Minivan	4	Blackwall	—
5	Green	Car	2	Blackwall	+
6	Green	SUV	4	Blackwall	—
7	Blue	SUV	2	Blackwall	—
8	Blue	Car	2	Whitewall	+
9	Red	SUV	2	Blackwall	—
10	Blue	Car	4	Blackwall	—

$Info(D) = I(3,7)$

$= -\frac{3}{10}\log_2\frac{3}{10} - \frac{1}{10}\log_2\frac{1}{10}$

$= 0.8813$

$Info_{color}(D) = \frac{3}{10}I(1,2) + \frac{4}{10}I(1,3) + \frac{3}{10}I(1,2)$

$= 0.3\left\{-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right\} + 0.4\left\{-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right\} + 0.3\left\{-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right\}$

$= 0.3(0.9183) + 0.4(0.8113) + 0.3(0.9183) = 0.2755 + 0.3245 + 0.2755$

$= 0.8755$

$Info_{type}(D) = \frac{4}{10}I(1,3) + \frac{2}{10}I(0,2) + \frac{3}{10}I(2,2) = 0.4(0.8113) + 0 + 0.3(1)$

$= 0.3245 + 0.3 = 0.6245$

$Info_{Doors}(D) = \frac{5}{10}I(3,2) + \frac{5}{10}I(0,5) = 0.5\left\{-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right\} + 0$

$= 0.5(0.9709) = 0.4855$

~~Attribut~~

$Info_{Tyres}(D) = \frac{4}{10}I(2,2) + \frac{6}{10}I(1,5) = 0.4(1) + 0.6\left\{-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}\right\} = 0.4 + 0.6(0.65)$

$= 0.4 + 0.39 = 0.79$

Attribute	Gain
Color	$0.8813 - 0.8755 = 0.0058$
Type	$0.8813 - 0.6245 = 0.2568$
Doors	$0.8813 - 0.4855 = 0.3958$
Tyres	$0.8813 - 0.79 = 0.0913$

∴ Choose Doors

Let D_1 be dataset D where Doors = 2

	COLOR	TYPE	TYRES	CLASS
1	Red	SUV	Whitewall	+
5	Green	Car	Blackwall	+
7	Blue	SUV	Blackwall	—
8	Blue	Car	Whitewall	+
9	Red	SUV	Blackwall	—

$Info(D_1) = I(3,2) = 0.9709$

$Info_{color}(D_1) = \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0) + \frac{2}{5}I(1,1)$

$= 0.4(1) + 0.2(0) + 0.4(1)$

$= 0.8$

$Info_{type}(D_1) = \frac{3}{5}I(1,2) + \frac{2}{5}I(2,0)$

$= 0.6(0.9183) + 0.4(0)$

$= 0.5509$

$Info_{tyres}(D_1) = \frac{2}{5}I(2,0) + \frac{3}{5}(I(1,2))$

$= 0.4(0) + 0.6(0.9183) = 0.5509$

Attribute	Gain
Color	0.9709 - 0.8 = 0.1709
Type	0.9709 - 0.5509 = .4200
Tyres	0.9709 - 0.5509 = 0.42

→ ∵ Both Type and Tyres have a lowest Gain, we can choose either. ∴ choose Type

→ Let D_{11} be the dataset where ~~Dataset~~ Doors = 2 and Type = SUV

	Color	Tyres	Class
1	Red	Whitewall	+
7	Blue	Blackwall	−
9	Red	Blackwall	−

$Info(D_{11}) = I(1,2) = 0.9183$

$Info_{color}(D_{11}) = \frac{2}{3} I(1,1) + \frac{1}{3} I(0,1)$

$= 0.66(1) + 0.33(0) = 0.66$

$Info_{tyre}(D_{11}) = \frac{1}{3} I(1,0) + \frac{2}{3} I(0,2) = 0$

Attribute	Gain
Color	0.2583
Tyres	0.9183

∴ Split using Tyres

→ Let D_{12} be the subset of D_1 where Doors = 2 and Type = Car

	Color	Tyres	Class
5	Green	Blackwall	+
8	Blue	Whitewall	−

$Info(D_{12}) = I(1,1) = 1$

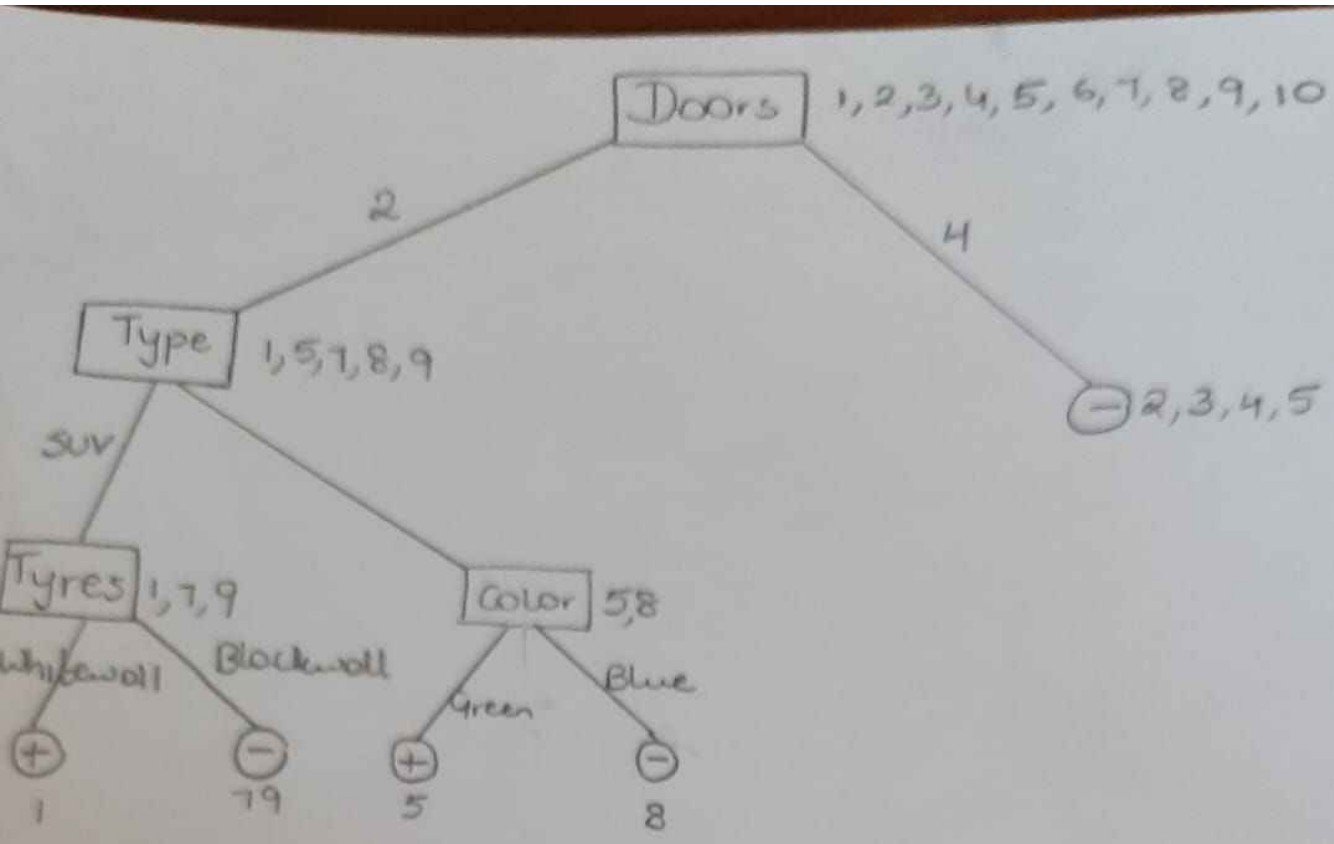
$Info_{color}(D_{12}) = \frac{1}{2} I(1,0) + \frac{1}{2} I(0,1)$

$= 0$

$Info_{tyres}(D_{12}) = \frac{1}{2} I(1,0) + \frac{1}{2} I(0,1) = 0$

∵ Color and Tyres both have Gain = 1, we can choose either ∴ choose Color

(7)


Doors
1,2,3,4,5,6,7,8,9,10
2
4
Type
1,5,7,8,9
2,3,4,5
SUV
Tyres
1,7,9
Color
5,8
Blackwall
Green
Blue
+
−
+
−
1
79
5
8

	AGE	INCOME	STUDENT	CREDIT_RATING	BUYS_COMPUTER
1	<=30	high	No	Fair	No
2	<=30	high	No	Excellent	No
3	31...40	high	No	Fair	Yes
4	>40	medium	No	Fair	Yes
5	>40	low	Yes	Fair	Yes
6	>40	low	Yes	Excellent	No
7	31...40	low	Yes	Excellent	Yes
8	<=30	medium	No	Fair	No
9	<=30	low	Yes	Fair	Yes
10	>40	medium	Yes	Fair	Yes
11	<=30	medium	Yes	Excellent	Yes
12	31...40	medium	No	Excellent	Yes
13	31...40	high	Yes	Fair	Yes
14	>40	medium	No	Excellent	No

6. $Info(D) = I(5,9) = -\frac{5}{14}\log_2\left(\frac{5}{14}\right) - \frac{9}{14}\log_2\left(\frac{9}{14}\right) = 0.9403$

$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$

$= \frac{5}{14}\left\{-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right\} + \frac{4}{14}(0) + \frac{5}{14}\left\{-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right\}$

$= \frac{5}{14}(0.9709) + 0 + \frac{5}{14}(0.9709) = 0.3468 + 0.3468 = 0.6935$

$Info_{income}(D) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1)$

$= \frac{4}{14}(1) + \frac{6}{14}\left[-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$= \frac{4}{14} + \frac{6}{14}[0.9183] + \frac{4}{14}[0.8113] = 0.2857 + 0.3936 + 0.2318$

$= 0.9111$

$Info_{student}(D) = \frac{7}{14} I(3,4) + \frac{7}{14} I(6,1) = 0.5\left[-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] + 0.5\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] = 0.5(0.9852) + 0.5(0.5917) = 0.4926 + 0.2958$

$= 0.7884$

$Info_{CR}(D) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3) = \frac{8}{14}\left[-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right] + \frac{6}{14}(1) = \frac{8}{14}(0.8113) + \frac{6}{14}$

$= 0.8922$

Attribute	Gain
Age	0.2467
Income	0.0287
Student	0.1518
CRedit_Rating	0.04813

∵ age has the highest gain, we split the dataset using age.

∴ D_1 is the subset of D with age = '<=30'

D_2 " " " " " age = '31...40'

D_3 " " " " " age = '>40'

	Income	Student	Credit Rating	Buys-Computer
1	high	No	Fair	No
2	high	No	Excellent	No
8	medium	No	Fair	No
9	low	Yes	Fair	Yes
11	medium	Yes	Excellent	Yes

$Info(D_1) = I(2,3) = 0.9709$

~~$Info_{high}$~~

$Info_{income}(D_1) = \frac{2}{5} I(2,0) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0) = 0 + 0.4(1) + 0 = 0.4$

$Info_{student}(D_1) = \frac{3}{5} I(0,3) + \frac{2}{5} I(2,0) = 0$

$Info_{credit\_rat}(D_1) = \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1) = 0.6(0.9183) + 0.4(1) = 0.5509 + 0.4 = 0.9509$

Attribute	Gain
Income	0.5709
Student	0.9709
Credit_Rating	0.02

Choose Student

	Income	Student	Credit Rating	Buys computer
4	medium	No	Fair	Yes
5	low	Yes	Fair	Yes
6	low	Yes	Excellent	No
10	medium	Yes	Fair	Yes
14	medium	No	Excellent	No

$Info(D_3) = I(3,2) = 0.9709$

$Info_{income}(D_3) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) = 0.6(0.9183) + 0.4(1)$

$= 0.5509 + 0.4 = 0.9509$

$Info_{student}(D_3) = \frac{2}{5} I(1,1) + \frac{3}{5} I(2,1) = 0.9509$

$Info_{credit\_rating}(D_3) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2) = 0$

Attribute	Gain
Income	0.02
Student	0.02
Credit_Rating	0.9709

∴ Choose credit_rating