DATA MINING
NUMERICALS

UNIT - 1

① Find frequent item sets using Apriori.
Min support = 40%. Min. confidence = 60%.

(2020-SEE)

Tid	Items
T_1	1, 3, 4
T_2	2, 3, 5
T_3	1, 2, 3, 5
T_4	2, 5
T_5	1, 3, 5

Support count = $\frac{40}{100}$ × no. of transactions

= $\frac{40}{100} \times 5 = 2$

C_1 =

items	Support
{1}	3
{2}	3
{3}	4
{4}	1
{5}	4

L_1 =

items	Support
{1}	3
{2}	3
{3}	4
{5}	4

Date:

C_2 =

items	Support
{1,2}	1
{1,3}	3
{1,5}	2
{2,3}	2
{2,5}	3
{3,5}	3

L_2 =

items	Support
{1,3}	3
{1,5}	2
{2,3}	2
{2,5}	3
{3,5}	3

C_3 =

items	Support
{1,3,5}	2
{1,2,3}	1
{1,2,5}	1
{2,3,5}	2

L_3 =

items	Support
{1,3,5}	2
{2,3,5}	2

C_4 =

items	Support
{1,2,3,5}	1

$L_4 = \phi$

$\therefore$ Frequent itemsets = L_3 = {1,3,5}, {2,3,5}.

Notes Date:

② (2019-SEE)

$L_3 = [\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}]$

i) Generate C_4 using $F_{k-1} \times F_1$ strategy.

ii) Generate C_4 using $F_{k-1} \times F_{k-1}$ strategy.

i) $\{1,2,3,4\}, \{1,2,3,5\}$
$\{1,2,4,5\}, \{1,3,4,5\}$
$\{2,3,4,5\}$

ii) $\{1,2,3,4\}, \{1,2,3,5\}$
$\{1,2,4,5\}, \{1,3,4,5\}$
$\{2,3,4,5\}$

Both result the same.

→ Another question had brute force, in that generate all possible combinations. $F_1 \times F_1 \times F_1$

Date:

Unit - 2

① (2020 - SEE)
Trace Apriori & generate rules
Min. support = 30%.
Min. confidence = 80%.

Tid	Itemset
1	1, 2, 5
2	2, 4
3	2, 3
4	1, 2, 4
5	1, 3
6	2, 3
7	1, 3
8	1, 2, 3, 5
9	1, 2, 3

Min. support count = $\frac{30}{100} \times 9 = 2.7 \Rightarrow 3$.

Notes

Date:

C_1: items	Support
{1}	6
{2}	7
{3}	6
{4}	2 ×
{5}	2 ×

L_1 = items	Support
{1}	6
{2}	7
{3}	6

C_2 = items	Support
{1, 2}	4
{1, 3}	4
{2, 3}	4

L_2 = items	Support
{1, 2}	4
{1, 3}	4
{2, 3}	4

C_3 = items	Support
{1, 2, 3}	2

$L_{\neq 3} = \phi$

∴ Frequent items = {1, 2}, {1, 3}, {2, 3}

~~* Confidence (A → B) S (~~

* Confidence $(A \rightarrow B) = \frac{S(A \cup B)}{S(A)}$

$C(1 \to 2) = \frac{S(1,2)}{S(1)} = \frac{4}{6} = 66\% < 80\%$ ×

$C(2 \to 1) = \frac{4}{7} < 80\%$ ×

$C(1 \to 3) = \frac{4}{6} < 80\%$ ×

$C(3 \to 1) = \frac{4}{6} < 80\%$ ×

$C(2 \to 3) = \frac{4}{7} < 80\%$ ×

$C(3 \to 2) = \frac{4}{6} < 80\%$ ×

∴ There are no strong rules.

② (SEE-2020) Draw FP tree & find frequent itemsets.

Min support = 30%.

Tid	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

Min. support count = $\frac{30}{100} \times 8 = \cancel{24}\ 2.4 = 3$.

Items	Support
A	5
B	6
C	3
D	6
E	4

→ Eliminate any that ~~arent~~ are below min support.

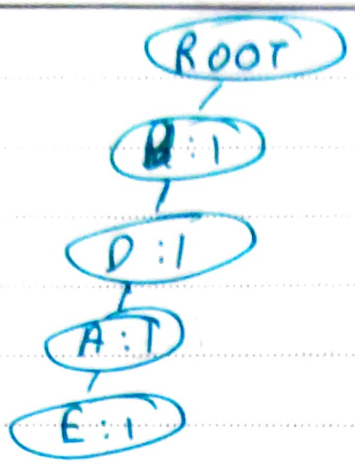
- Setting priority

Items	Support
B	6
D	6
A	5
E	4
C	3

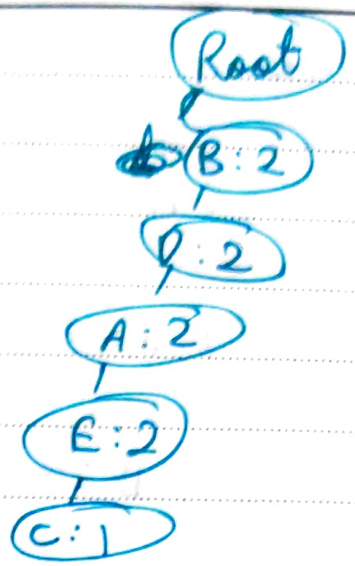
Redefine transactions, do not show items less than min support.

Tid	Items
1	B, D, A, E
2	B, D, A, E, C
3	B, A, E, C
4	B, D, A
5	D
6	B, D
7	D, A, E
8	B, C

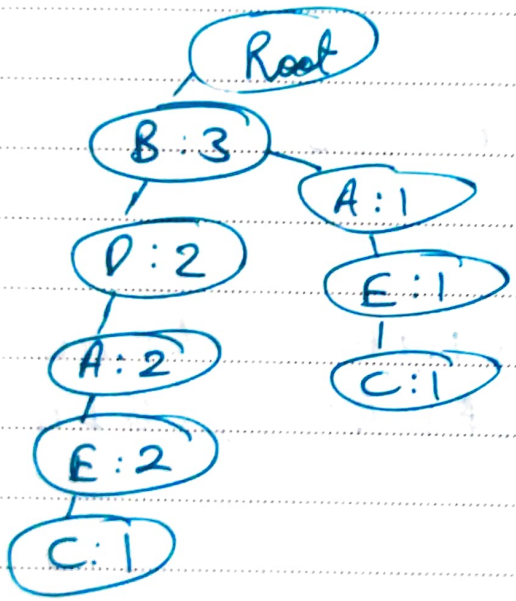
T1



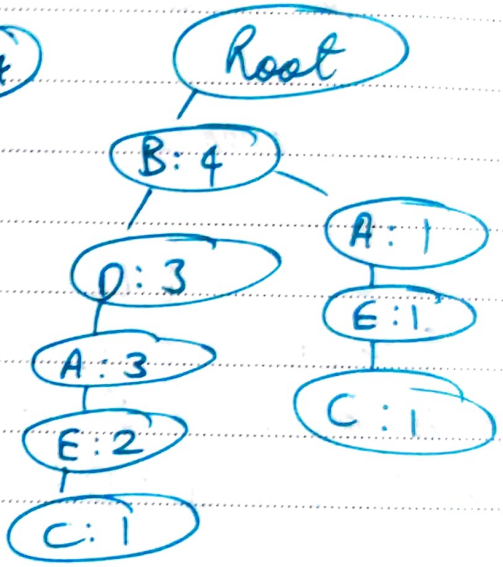
T2



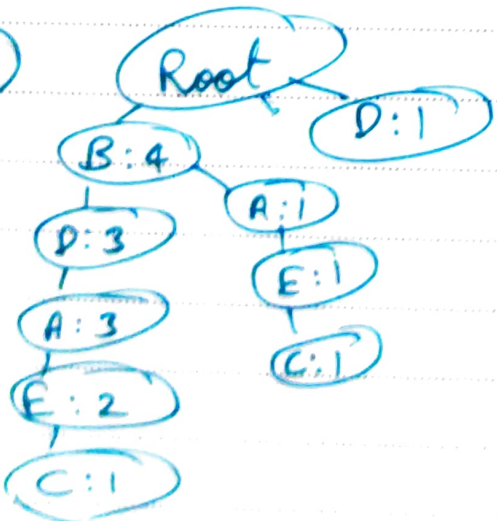
T3



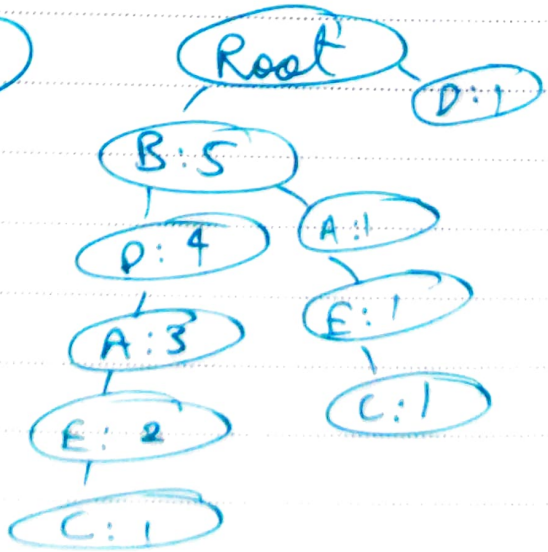
T4

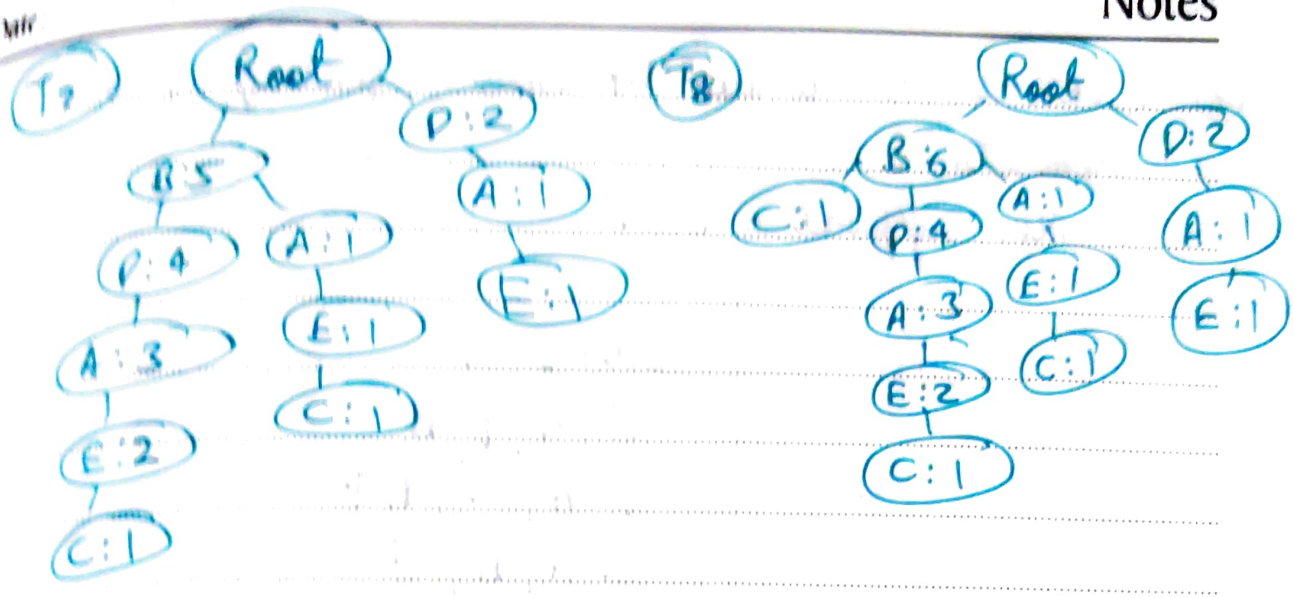


T5



T6

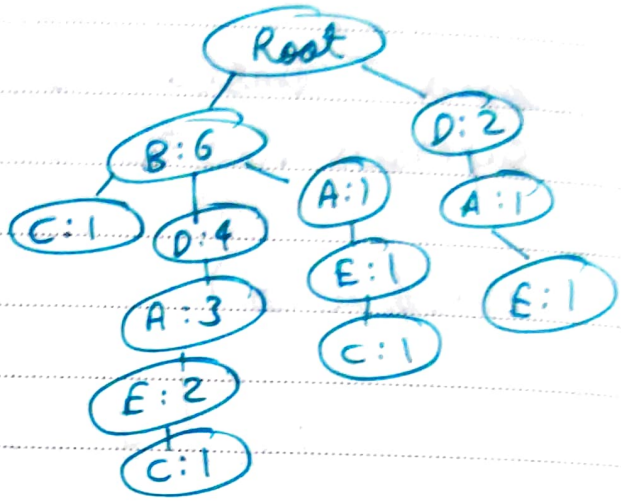




Item	Conditional Pattern base	Conditional FP tree	FP generated
C	{{B, D, A, E} : 1}, {B : 1}, {BAE : 1}}	{B : 3}, ignored others because min. support.	{B, C}
E	{BDA : 2}, {BA : 1}, {D, A : 1}	{A : 3}, {B : 3}	{A, E}
A	{BD : 3}, {		

@ Conditional Pattern base.

C : 3
E : 4
A : 5
D : 6



→ For C
{B, D, A, E : 1}
{B : 1}
{B, A, E : 1}

→ For E : {B, D, A : 2}, {B, A : 1}, {D, A : 1}

→ For A : {B, D : 3}, {B : 1}, {D : 1}

→ For D : {B : 4}

Date:

⑥ Conditional FP tree MS = 3

→ For C : {B : 3}

→ For E : {B : 3, A : 3} ~~D~~ ~~{D:~~

→ For A : {B : 4, D : 3}

→ For D : {B : 4}

⑦ Frequent ~~Itemsets~~ Patterns

→ For C : {B, C : 3}
→ For E : {B, E : 3}, {A, E : 3}, {B A E : 3}
→ For A : {B, A : 4}, {B D, A : 3}, {B D A : 3}
→ For D : {B, D : 4}

(3) (SEE 2019)

No. of Transactions = N = 100
" " items = 20

$S(a) = 25\%$ $S(b) = 90\%$.
$S(a,b) = 20\%$ $MS = 10\%$. $MC = 60\%$.

a) $C(a \to b) = ?$

b) $I(a,b) = ?$

a) $C(a \to b) = \frac{S(a,b)}{S(a)} = \frac{20}{25} = 80\% > 60\%$.

$\therefore$ It is interesting.

b) ** $I(a,b) = \frac{C(A \to B)}{S(B)} = \frac{80\%}{90\%}$

$= 88\%$

→ Interest factor compares the frequency of a pattern to the baseline frequency.

Date:

* Maximal Itemset : If none of its immediate supersets is frequent.

* Closed Itemset : If ~~not~~ none of its immediate supersets has the same support count.

④ (Makeup 2020)

Tid	Itemset
1	A B D E
2	B C E
3	A B D E
4	A B C E
5	A B C D E
6	B C D

MS = 3

compute Closed itemset & maximal itemset

C_1 =

item	support
A	4
B	6
C	4
D	4
E	5

Notes Date :

C_2 = item	Support
AB	4
~~AC~~	~~2~~ ×
AD	3
AE	4
BC	4
BD	4
BE	5
~~CD~~	~~2~~ ×
CE	3
DE	3

C_3 = item	Support
ABD	3
ABE	4
~~ABC~~	~~2~~ ×
ADE	3
~~ACE~~	~~2~~ ×
~~BCD~~	~~2~~ ×
BCE	3
BDE	3
~~CDE~~	~~1~~ ×

C_4 = item	Support
ABDE	3 ✓
~~ABCE~~	~~2~~
~~BCDE~~	~~1~~

Maximal: ABDE, BCE,

Closed: ABDE, BCE, ABE, BE, BD, BC, B

Date: (5)

	Basket Ball	Not BB	Σ
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Σ	3000	2000	5000

i) Lift = $\frac{S(A,B)}{S(A) \times S(B)} = \frac{I(A,B)}{S(A)S(B)} = \frac{S(A,B)}{S(A)S(B)} = \frac{N f_{11}}{f_{1+} f_{+1}}$

$\frac{f_{11}}{N} = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N} = \frac{f_{1+} f_{+1}}{N}$

$\therefore$ Lift $= \frac{5000 \times 2000}{3750 \times 3000} = 0.88 < 1.$

A & B are positively corelated.

Notes

Date :

Unit 3

Decision Tree

Age	Competition	Type	Profit
old	Yes	Software	Down
old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	SW	Down
Mid	Yes	HW	Down
Mid	No	HW	Up
Mid	No	SW	Up
New	Yes	SW	Up
New	No	HW	Up
New	No	SW	Up

$P = 5 \quad N = 5$

① CALCULATE ENTROPY OF CLASS ATTRIBUTE

$$\text{Entropy (class)} = \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) \frac{-N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

$$\therefore \frac{-5}{10}\log_2\left(\frac{5}{10}\right) \frac{-5}{10}\log_2\left(\frac{5}{10}\right)$$

$$-0.5 \times (-1) \quad -0.5\,(-1)$$

$$\therefore \quad = 0.5 + 0.5 = \boxed{1}$$

Date:

② FOR EACH ATTRIBUTE, calculate
  ① information gain
  ② Entropy (attribute)
  ③ Gain

① Information Gain

$$I(P_i, N_i) = \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) \frac{-N}{P+N} \log_2\left(\frac{N}{P+N}\right)$$

② Entropy of attribute

$$= \sum \frac{P_i + N_i}{P+N} \times (I(P_i, N_i))$$

③ Gain = Class entropy - Entropy of attribute.

Notes

Date:

→ For Age

	P_i	N_i	$I(P_i, N_i)$
Old	0	3	0
Mid	2	2	1
New	3	0	0

Entropy of age = $\frac{\sum P_i + N_i}{P+N} (I(P_i, N_i))$

$= \frac{0+3}{5+5} \times 0 + \frac{2+2}{5+5} \times 1 + \frac{3+0}{5+5} \times 0$

$= \frac{2}{5} = \underline{0.4}$

Gain = $1 - 0.4 = \underline{0.6}$

→ For Competition

	P_i	N_i	$I(P_i, N_i)$
Yes	1	3	0.81
No	4	2	0.92

a) Info gain = $-\frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{3}{4} \log_2\left(\frac{3}{4}\right)$

$= 0.5 + 0.31 = 0.81128$

b) $\frac{-4}{6} \log\left(\frac{4}{6}\right) - \frac{2}{6} \log\left(\frac{2}{6}\right) = $ ~~0.3899~~ 0.918

Entropy of completion = $\frac{(1+3)}{5+5} \times 0.81 + \frac{6}{10}(0.92)$

$= 0.8754.$

Gain = $1 - 0.8754 = 0.124515.$

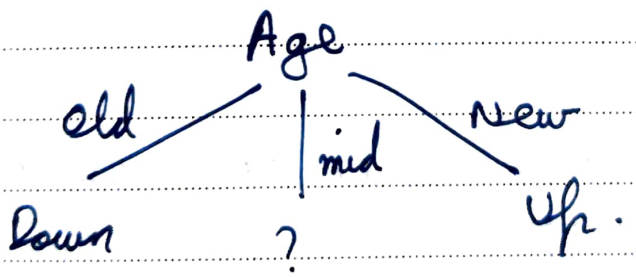
→ For Type

	P_i	N_i	$I(P_i, N_i)$
HW	2	2	1
SW	3	3	1

Entropy of ~~sw~~ type = $\frac{2+2}{10} \times 1 + \frac{6}{10} \times 1$

$= \frac{4}{10} + \frac{6}{10} = (1).$

Gain = 0



Age	competition	Type	Profit
mid	Y	SW	D
mid	Y	HW	D
mid	N	SW	U
mid	N	HW	U

For Competition:

	P_i	N_i	$I(P_i, N_i)$
Y	2	2	0
N	2	0	0

$$\text{Entropy} = \frac{2}{4} \times 0 + \frac{2}{4} \times 0 = 0$$

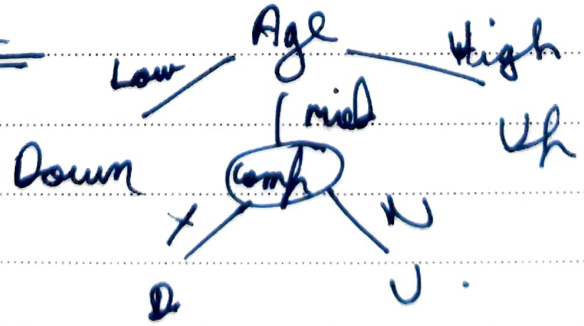
Gain = 1

For type

	P_i	N_i	$I(P_i, N_i)$
SW	1	1	1
HW	1	1	1

$$\frac{2}{4} \times 1 + \frac{2}{4} \times 1 = 1$$

Gain = 0



Notes

Date:

Decision Tree using CART

$$\text{Gini index}(\text{attribute} = \text{value}) = 1 - \sum_{i=1}^{N} (p_i)^2$$

$$\text{Gini index}(\text{attribute}) = \sum_{v = \text{values}} P_v \times GI(v)$$