PlanetOjas

Suppose that the data for analysis includes the attribute age. The age values for data tuples are (in increasing order)

{13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}

find Mean, Median, Mode, Midrange, Range, Quartiles (Q1, Q2, Q3), give five number Summary of the data. Show a box Plot of the data

$$\sum_{i=1}^{n} \frac{x_i}{n} = \frac{809}{27} = 30$$

Median $= \left(\frac{N+1}{2}\right)^{th}$ term

$= \left(\frac{27+1}{2}\right)^{th}$ term

$= \left(\frac{28}{2}\right)^{th}$ term

$= 14^{th}$ term

$= 25$

3. Mode: Since two numbers are repeating highest no. of time So the dataset is a bimodel dataset. Modes of the dataset are 25 & 35.

4. Midrange: $\frac{13+70}{2}$

$= 41.5$

5. Range: $70 - 13$

$= 57$

6. Quartile 2 (Q2) = Median

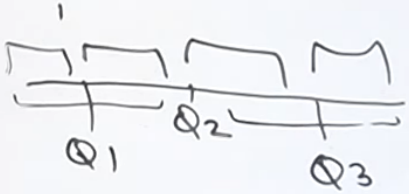
$= 25$

Quartile 1 (Q1) $= \left(\frac{N+1}{4}\right)^{th}$ term

$= 20$

Quartile 3 (Q3) $= \left((N+1) * \frac{3}{4}\right)^{th}$ term

$=$



PlanetOjas

ata

)

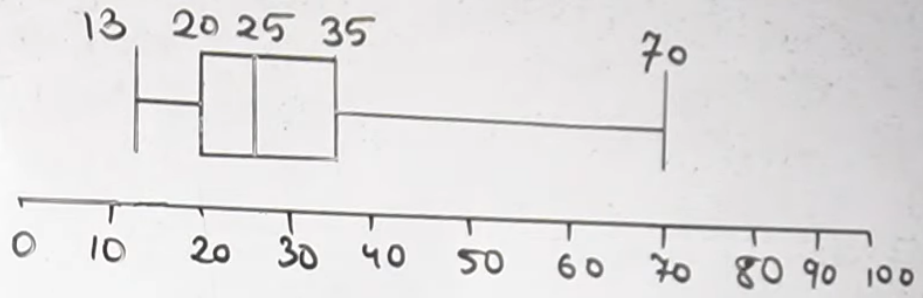
25,25,25,25,
15,46,52,70}

range, Range,
five number
a box Plot of

$= \frac{809}{27} = 30$

m

rm



4. Midrange: $\frac{13+70}{2}$

$= 41.5$

5. Range: $70 - 13$

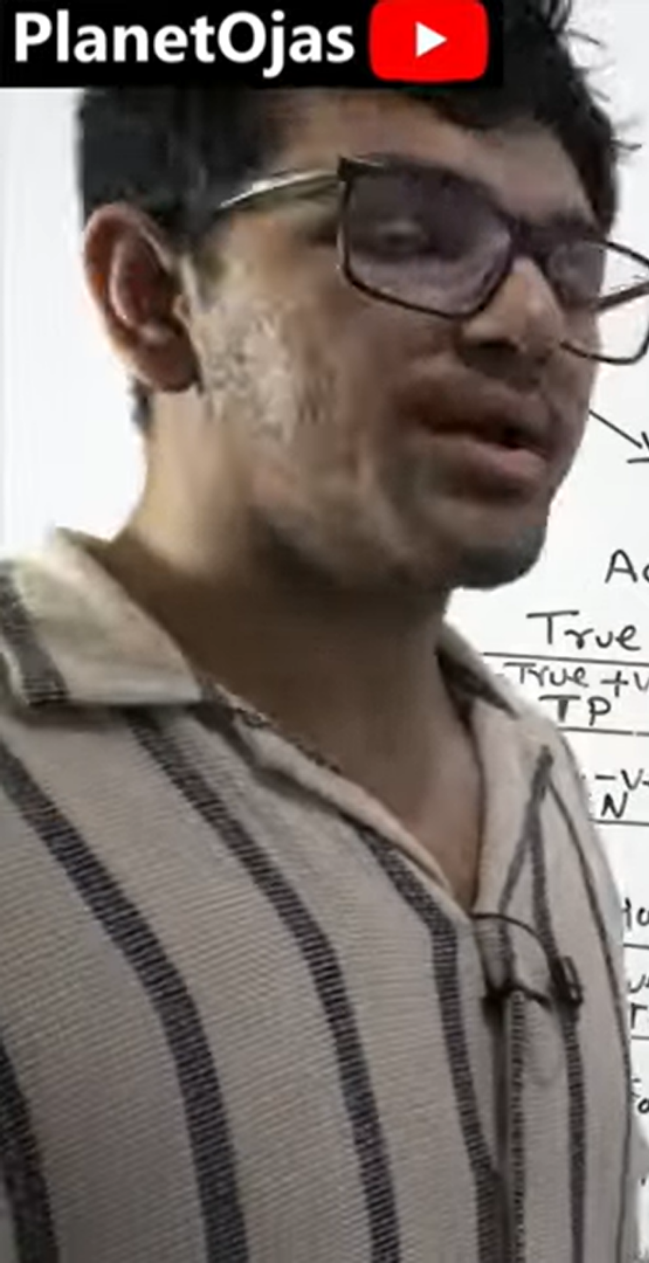
$= 57$

$Q_1 = 20$
$Q_2 = 25$
$Q_3 = 35$
Min = 13
Max = 70

6. Quartile 2 (Q2) = Median
$= 25$

Quartile 1 (Q1) = $\left(\frac{N+1}{4}\right)^{th}$ term
$= 20$

Confusion Matrix



1% logo ko TB hai

99% logo ko TB nahi hai

	Actual: True	Actual: False
True +ve	TP	False +ve FP
-ve	N	True -ve TN

	Actual: have TB	Actual: Don't have TB
ve	TP +ve	False +ve FP
False -ve	= N	True -ve TN

TP = True +ve = correctly identified +ve

TN = True -ve = correctly identified -ve

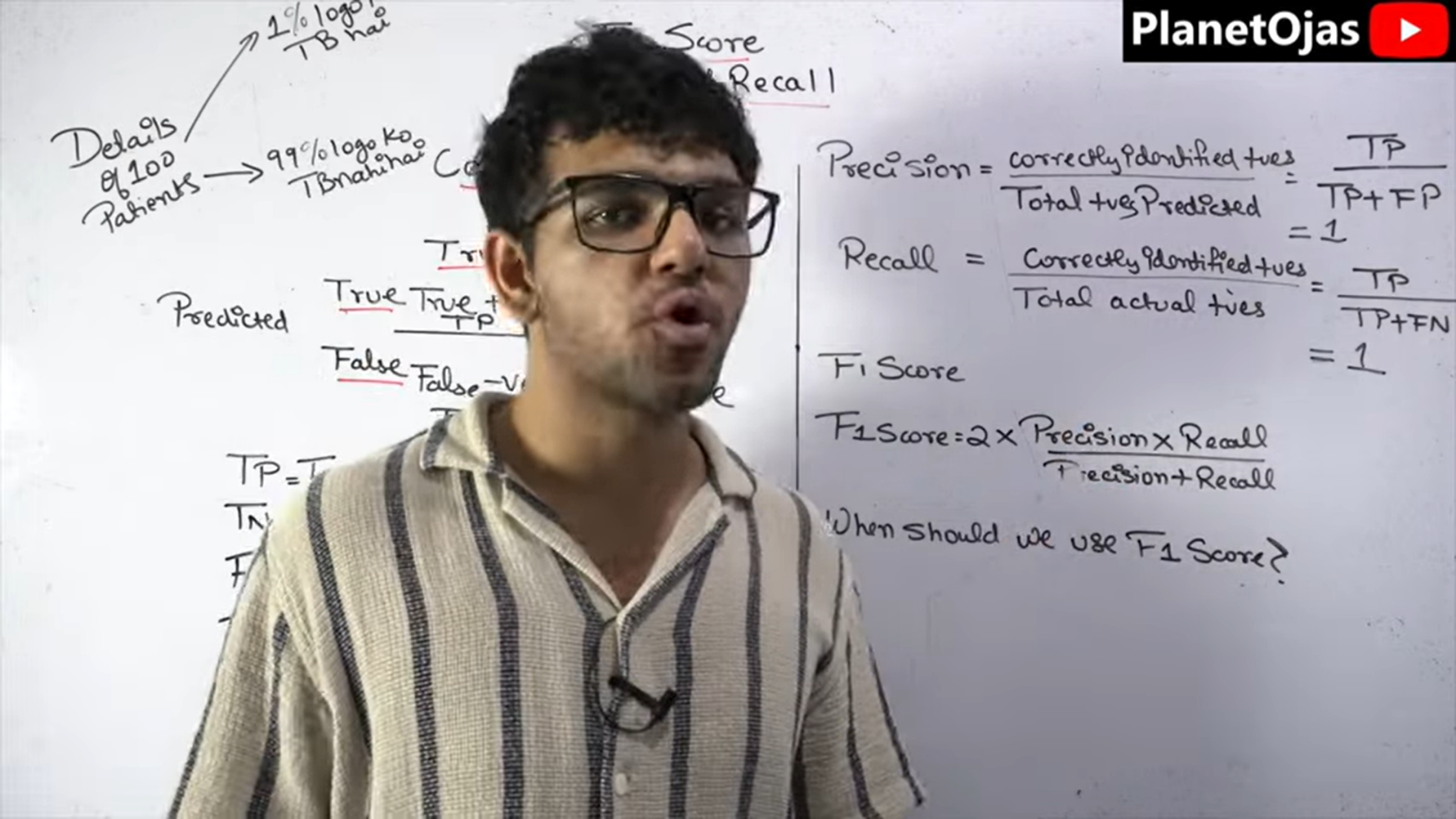
FN = False -ve = Incorrectly identified +ve

FP = False +ve = Incorrectly identified -ve

1. Accuracy = $\frac{\text{Total correct Predictions}}{\text{Total Predictions}} = \frac{TP+TN}{TP+FP+FN+TN}$

2. Recall = $\frac{\text{correctly identified +ve}}{\text{Total actual +ve}} = \frac{TP}{TP+FN}$

3. Specificity = $\frac{\text{correctly identified Negative}}{\text{Total actual -ve}} = \frac{TN}{TN+FP}$


Details of 100 Patients
1% logo ko TB hai
99% logo ko TB nahi hai
F1 Score
Recall
Predicted
True
False
TP =
TN
F
Precision = correctly identified tves / Total tves Predicted = TP / TP+FP = 1
Recall = Correctly identified tves / Total actual tives = TP / TP+FN = 1
F1 Score
F1 Score = 2 x Precision x Recall / Precision + Recall
When should we use F1 Score?

Apriori Algorithm

=> Given the following data, apply the Apriori algorithm. Given Support threshold = 50%, Confidence = 60%

Transaction	List of items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

Sol: Support threshold = 50%

min_sup = 0.5 × no. of transactions

= 0.5 × 6

= 3

Step1: Counting each itemset (C_1) we get:

Itemset	count
I_1	4
I_2	5
I_3	4
I_4	4
I_5	2

Step2: I_5 itemset does not meet min sup = 3 so after discarding it we get

Itemset	count
I_1	4
I_2	5
I_3	4
I_4	4

Step3: Join step

Itemset	count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	3
$\{I_1, I_4\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	3
$\{I_3, I_4\}$	2

Step4: $\{I_1, I_4\}$ & $\{I_3, I_4\}$ doesn't meet min sup so after discarding them we get:

Itemset	count
I_1, I_2	4
I_1, I_3	3
I_2, I_3	4
I_2, I_4	3

Step5: Jo

Itemset

$I_1, I_2,$

$I_1, I_2,$

I_1, I_3, I

$I_2, I_3,$

Step6

$I_2, I_3, I_4\}$

discarding

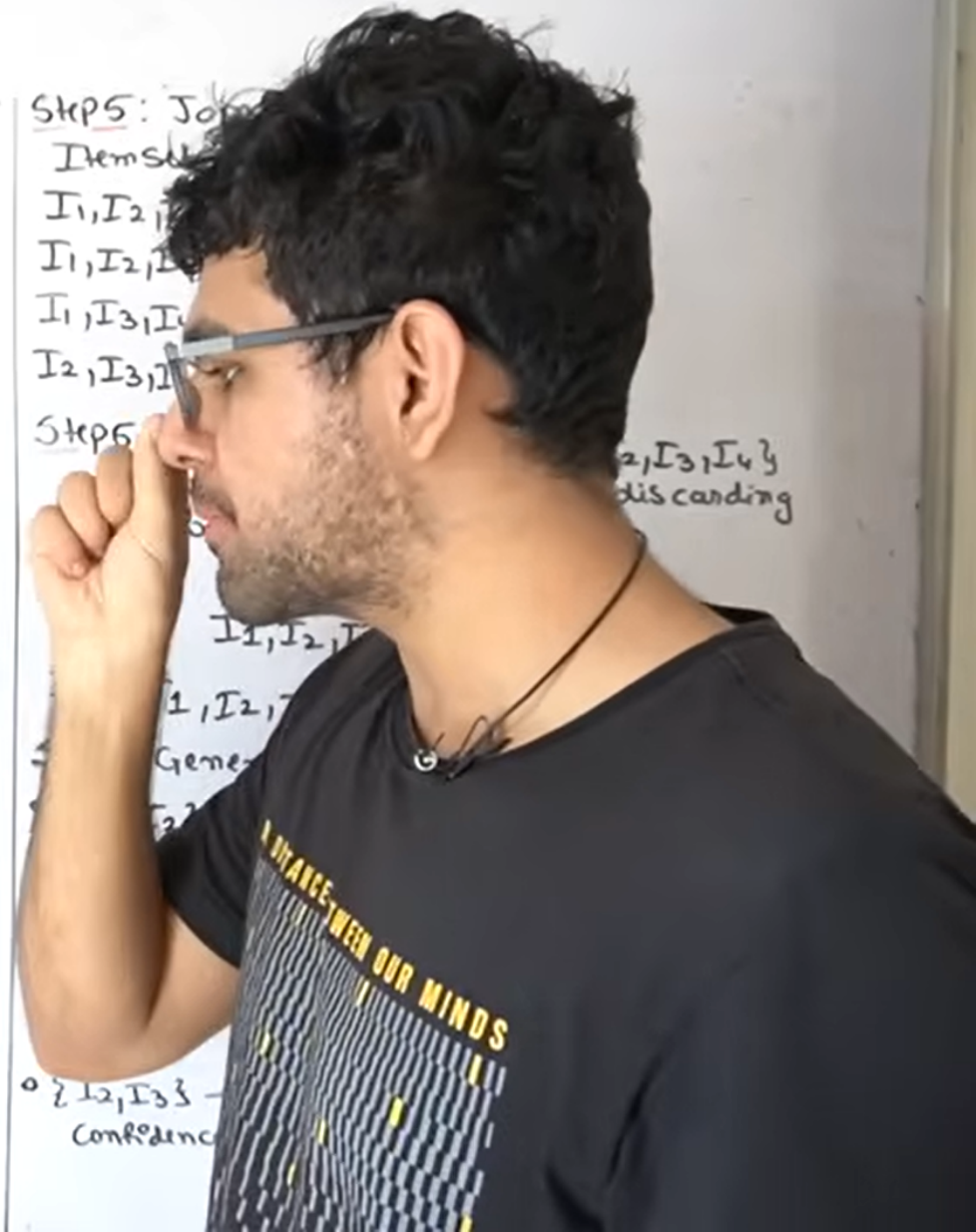
$I_1, I_2,$

$I_1, I_2,$

Gene

$\{I_2, I_3\}$ –

Confidenc



Apriori Algorithm

=> Given the following data, apply the Apriori algorithm. Given Support threshold = 50%, Confidence=

Transaction	List of item
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	$I_1, I_2, I_$

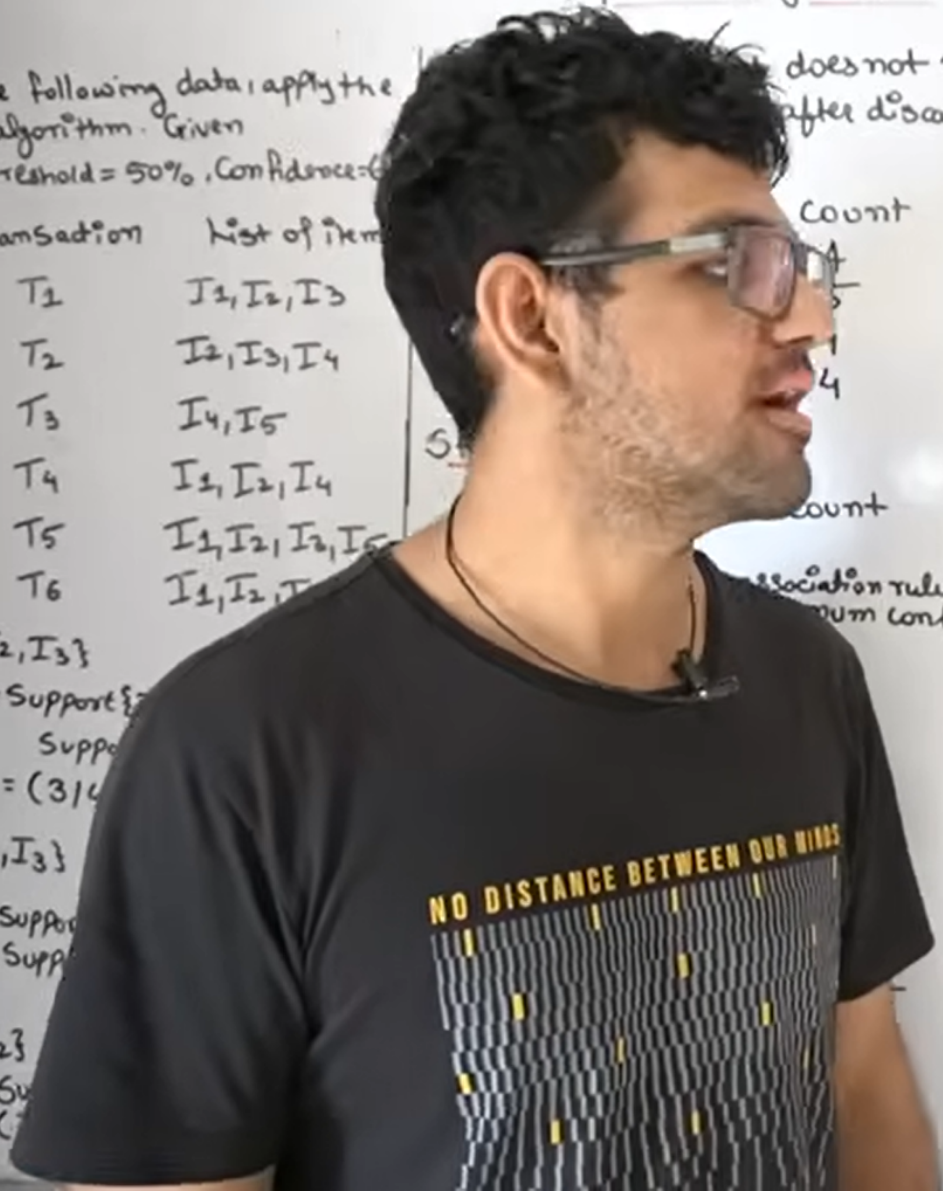
- $\{I_1\} \rightarrow \{I_2, I_3\}$
  Confidence = Support{
  Suppo
  = (3/
- $\{I_2\} \rightarrow \{I_1, I_3\}$
  Confidence = Suppo
  Supp
- $\{I_3\} \rightarrow \{I_1, I_2\}$
  Confidence = Su
  = (

does not meet
after discarding

Count
4
1
4

Count

association rules
minimum confidence



Step5: Join Step

Itemset	Count
I_1, I_2, I_3	3
I_1, I_2, I_4	2
I_1, I_3, I_4	1
I_2, I_3, I_4	2

Step6: $\{I_1, I_2, I_4\}, \{I_1, I_3, I_4\}$ & $\{I_2, I_3, I_4\}$ does not meet min_sup = 3 so after discarding them we get:

Itemset	Count
I_1, I_2, I_3	3

Thus $\{I_1, I_2, I_3\}$ is frequent.

Step7: Generate Association Rules

- $\{I_1, I_2\} \rightarrow \{I_3\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_1, I_2\}$
  = (3/4) x 100 = 75%
- $\{I_1, I_3\} \rightarrow \{I_2\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_1, I_3\}$
  = (3/3) x 100 = 100%
- $\{I_2, I_3\} \rightarrow \{I_1\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_2, I_3\}$
  = (3/4) x 100 = 75%

Apriori Algorithm

=> Given the following data, apply the Apriori algorithm. Given
Support threshold = 50%, Confidence = 60%

Transaction	List of items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

- $\{I_1\} \rightarrow \{I_2, I_3\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_1\}$
  $= (3/4) \times 100 = 75\%$
- $\{I_2\} \rightarrow \{I_1, I_3\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_2\} = (3/5) \times 100 = 60\%$
- $\{I_3\} \rightarrow \{I_1, I_2\}$
  Confidence = Support $\{I_1, I_2, I_3\}$ / Support $\{I_3\}$
  $= (3/4) \times 100 = 75\%$

Step 2: I_5 itemset does not meet minsup = 3 so after discarding it we get

Itemset	Count
I_1	4
I_2	5
I_3	4
I_4	4

Step 3: Join step

Itemset	Count

Ans: All the abo… association rules are strong … minimum confidence is 60%

…[I3, I4] doesn't … discarding

Step 5: Join Step

Itemset
I_1, I_2, I_3
I_1, I_2, I_4
I_1, I_3, I_4
I_2, I_3, I_4

Step 6: $\{I_1, I_2, I_4$ … does not meet … them we get :

Item…
I_1, I_2, I_3

Thus $\{I_1, I_2, I_3\}$ is freq…

Step 7: Generate A…

- $\{I_1, I_2$ …

PlanetOjas

FP-Growth Algorithm

⇒ A database has five tr

Let min_sup = 60% & mi

Find all the frequent item

FP-Growth

TID	Items_b
T100	{M, O,
T200	{D, O,
T300	{M, A, K,
T400	{M, U, C,
T500	{C, O,

⇒ min-sup = 60%

∴ Sup-count to be

Step 1: Coun

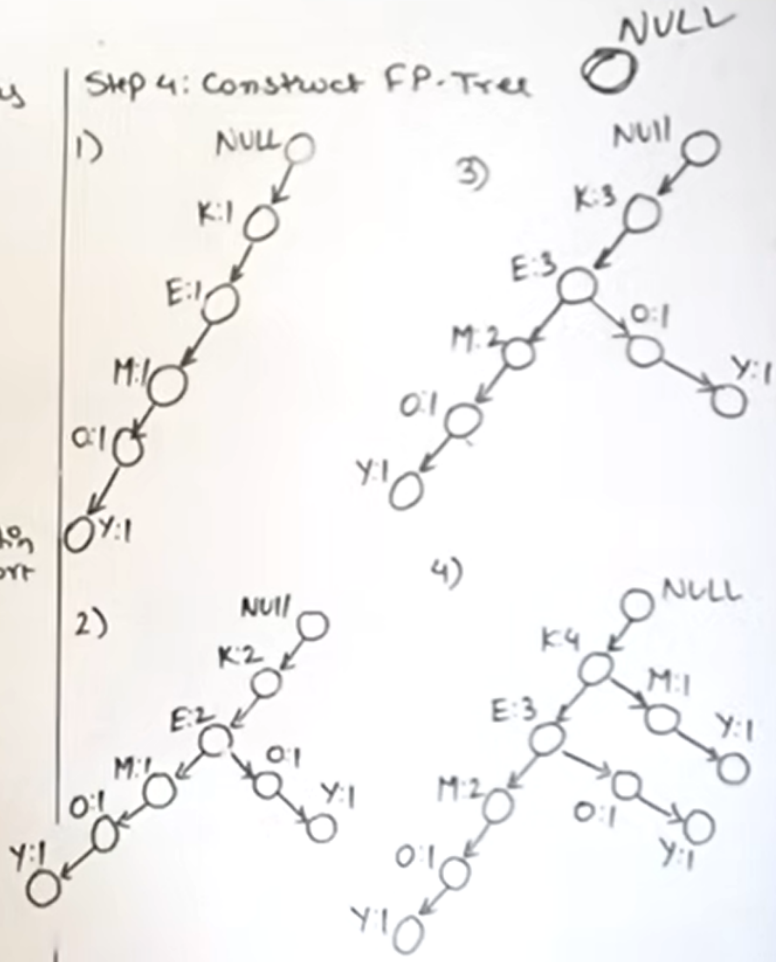
scard A, C, D, I, N, U as
satisfy sup count &
emaining items in
rder of Sup. count

set	Sup-count
	5
Y}	4
M}	3
O}	3
{Y}	3

3: Sort items in each transaction according to descending support count

TID	List of items
T100	{K, E, M, O, Y}
T200	{K, E, O, Y}
T300	{K, E, M}
T400	{K, M, Y}
T500	{K, E, O}

Step 4: Construct FP-Tree



PlanetOjas

FP-Growth Algorithm

⇒ A database h

Let min_sup

Find all the f

FP-Growth

TID
T100
T200
T300
T400
T500

⇒ min-sup = 60%

∴ Sup-count to be

Step 1: Co

Step 2 : Discard A,C,D,I,N,U

does not satisfy sup count &

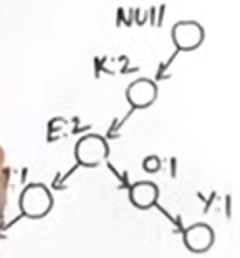
rranging remaining items in

descending order of Sup. count

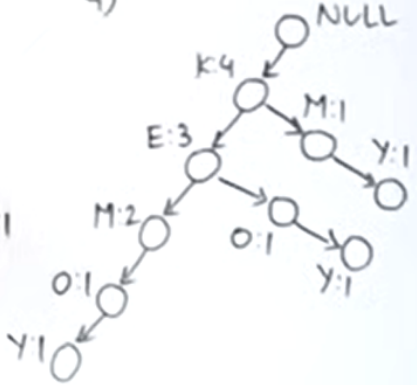
Itemset	Sup. count
{K}	5
{E}	4
{M}	3
{O}	3
{Y}	3

Itemset	Conditional Pattern base	Conditional FP Tree	Frequent Patterns Generated
{Y}	{{K,E,M,O:1},{K,E,O:1}, {K,M:1}}	(K:3)	{K,Y:3}
{O}	{{K,E,M:1},{K,E:2}}	(K:3,E:3)	{K,O:3}, {E,O:3}, {K,E,O:3}
{M}	{{K,E:2},{K:1}}	(K:3)	{K,M:3}
{E}	{{K:4}}	(K:4)	{K,E:4}
{K}	–	–	–

2)



4)



Min Max Technique [0 - 1]

House	Square Foot	Bedrooms	Price(in Lakh)
1	1200 = .25	3	50
2	1500	4	60
3	1000	2	40
4	1800	5	80

$$X' = \frac{X - min}{max - min} = \frac{1200 - 1000}{1800 - 1000} = \frac{200}{800} = .25$$

Z-score Technique

Student	Height(in inches)
1	64
2	70
3	72
4	68
5	76

$$X' = \frac{X - \text{mean } \mu}{\text{Std. dev } \sigma}$$

$$\text{Std} = \sqrt{\frac{(64-70)^2 + (70-70)^2 + (72-70)^2 + (68-70)^2 + (76-70)^2}{5}}$$

$$= \sqrt{\frac{36 + 0 + 4 + 4 + 36}{5}} = \frac{80}{5} = \sqrt{16} = 4$$

$$\frac{350}{5} = 70$$

$$\frac{64-70}{4} = \frac{-6}{4} = -1.5$$

Data Smoothing & Data Reduction

Suppose a group of age records has been sorted as follows:
{5, 12, 13, 14, 15, 18, 26, 28, 42}
Perform data smoothing & data reduction.

=> Equal Freq Bin:

Bin1: {5, 12, 13}
Bin2: {14, 15, 18}
Bin3: {26, 28, 42}

Equal Width Bin:

Bin1: {5, 12, 13, 14, 15}
Bin2: {18, 26, 28}
Bin3: {42}

* Smoothing by Mean (1):

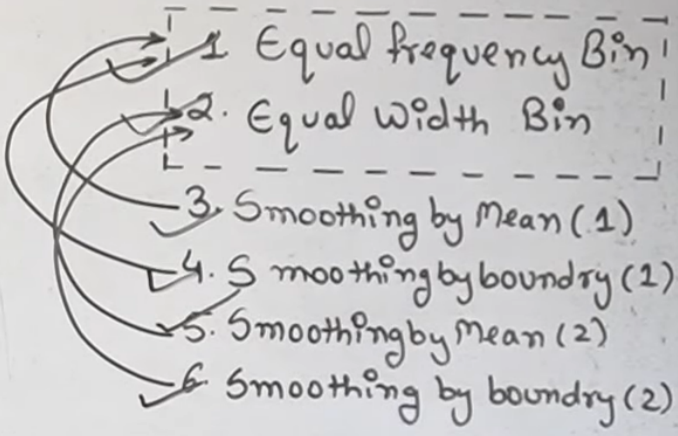
Bin1: {10, 10, 10}
Bin2: {16, 16, 16}
Bin3: {32, 32, 32}

* Smoothing by Boundry (1):

Bin1: {5, 13, 13}
Bin2: {14, 14, 18}
Bin3: {26, 26, 42}

* Smoothing by Mean (2):

Bin1: {12, 12, 12, 12, 12}
Bin2: {24, 24, 24}
Bin3: {42}



Smoothing by boundry (2)

Bin1: {5, 15, 15, 15, 15}
Bin2: {18, 28, 28}
Bin3: {42}

Chi-Square Test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Watch → / Actor ↓	Y	N	(R) Total ↓
M	140	44	184
F	178	38	216
(C) Total →	318	82	400

$$E = \frac{RT \times CT}{T} \quad / = \frac{184 \times 318}{400} = 146$$

✓ = (184 × 82)/400 = 38

✓ = (216 × 318)/400 = 172

✓ = (216 × 82)/400 = 44

✓ $\alpha = 0.05$

H0: No connection

H1: full connection

df = (Rows − 1) * (cols − 1) = 1 → 3.841 <

Actor	Watch	O	E	$(O-E)^2$	$(O-E)^2/E$
M	Y	140	146	36	0.246
M	N	44	38	36	0.947
F	Y	178	172	36	0.209
F	N	38	44	36	0.818

Cal → 2.220