Ans 1 → It refers to extracting knowledge from a given set of data.

The steps involved in data mining are:-

i) Data selection – It collects the data which are relevant enough.

ii) Data Processing – Processes the data in some patterns.

iii) Data Transformation – Transforms the data as per needed.

iv) Data Mining = extracting knowledge from data.

v) Pattern Analysis – Drawing the relativeness in a given data set and analysing the sequence and pattern of it.

v) Knowledge Extraction – Getting the desired result of the query send by the user.

Ans 2 → data set = 13, 15, 16, 17, 19, 20, 20, 21, 22, 22, 24, 24, 25, 25, 30, 33, 34, 35, 35, 35, 36, 36, 40, 45, 46, 52, 70.

i) Mean = $\frac{810}{27} = 30$

Median = 25 $\left(\frac{27+1}{2}\right) = \frac{28}{2}$ 14th element

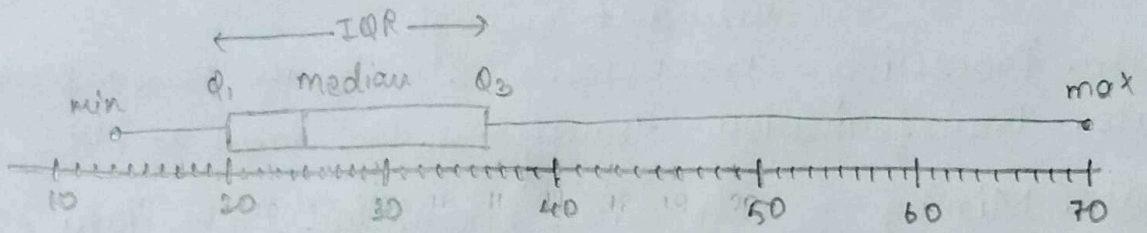
ii) Mode = 35

iii) Midrange = $\frac{\max + \min}{2} = \frac{13+70}{2} = 41.5$

iv) $Q_1 = \frac{13+1}{2} = \frac{14}{2}$ 7th = 20

Q_3 = 7th from last = 36

v/ min = 13
Q_1 = 20
med = 25
Q_3 = 36
max = 70



3/ Data = 13, 15, 16, 17, 19, 20, 20, 21, 22, 22, 24, 24, 25, 25, 30, 33, 34, 35, 35, 35, 36, 36, 40, 45, 46, 52, 70.

B1 = 13, 15, 16, 17, 19, 20, 20, 21, 22
B2 = 22, 24, 24, 25, 25, 30, 33, 34, 35
B3 = 35, 35, 36, 36, 40, 45, 46, 52, 70

i) Smoothing by bin means
B1 = 18.11, 18.11, 18.11, 18.11, 18.11, 18.11, 18.11, 18.11, 18.11
B2 = 28, 28, 28, 28, 28, 28, 28, 28, 28
B3 = 43.88, 43.88, 43.88, 43.88, 43.88, 43.88, 43.88, 43.88, 43.88

ii) Smoothing by bin medians
B1 = 19, 19, 19, 19, 19, 19, 19, 19, 19
B2 = 25, 25, 25, 25, 25, 25, 25, 25, 25
B3 = 40, 40, 40, 40, 40, 40, 40, 40, 40

iii) Smoothing by bin boundaries
B1 = 13, 13, 13, 13, 22, 22, 22, 22, 22
B2 = 22, 22, 22, 22, 22, 35, 35, 35, 35
B3 = 35, 35, 35, 35, 35, 35, 35, 35, 70

4)

Player number	No. of matches.	No. of wickets
1	11	8
2	23	2
3	25	3
4	7	9
5	6	8
6	22	3
7	25	1
8	19	4
9	10	6
10	11	7
	159	51

mean $= \frac{51}{10} = 5.1$

S.D of wickets $= \sqrt{\frac{(x_i - \bar{x})^2}{n}}$

$\Rightarrow \frac{(8-5.1)^2 + (2-5.1)^2 + (3-5.1)^2 + (9-5.1)^2 + (8-5.1)^2 + (3-5.1)^2 + (1-5.1)^2 + (4-5.1)^2 + (6-5.1)^2 + (7-5.1)^2}{10}$

$= \frac{8.41 + 9.61 + 4.41 + 15.21 + 8.41 + 4.41 + 16.81 + 1.21 + 0.81 + 3.61}{10}$

$= \frac{72.9}{10} = \sqrt{7.29} = 2.7$ Ans.

S.D of no. of matches.

mean $= \frac{159}{10} = 15.9$

$\sqrt{\frac{(11-15.9)^2 + (23-15.9)^2 + (25-15.9)^2 + \ldots + (11-15.9)^2}{10}}$

$= 7.23$ Ans.

5) Dimensional models in data warehouse is used to break data up into "facts" and "dimensions" to organize and describe entities within your data warehouse. The result is a staging layer in the data warehouse that cleans and organizes the data into the business end of the warehouse that is more accessible to data consumers.

6) Support = 60%. Confidence = 80%. Min support count = $\frac{4 \times 60}{100} = 2.4$

1-itemset	frequency
{1}	1 DISCARD < 2.4
{2}	4
{3}	3
{4}	4
{5}	2 DISCARD
{6}	2 DISCARD

Item	freq
{1}	1
{2}	4
{3}	3
{4}	4
{5}	2
{6}	2

3-itemset	freq.
{1,2,3}	1 DISCARD < 2.4
{1,2,4}	1 D
{1,2,5}	0 D
{1,2,6}	0 D
{1,3,4}	1 D
{1,3,5}	0 D
{1,3,6}	0 D
{1,4,5}	0 D
{1,4,6}	0 D
{1,5,6}	0 D
{2,3,4}	3
{2,3,5}	1 D
{2,3,6}	1 D
{2,4,5}	2 D
{2,4,6}	2 D
{2,5,6}	2 D
{3,4,5}	1 D

	freq
{3,4,6}	1 D
{3,5,6}	1 D
{4,5,6}	2 D

The new table is

{2,3,4}.

6) Support = 60%.

Minimum Support count = $4 \times \frac{60}{100} = 2.4$

Step-1.

Item set	freq	
{1}	1	Discard < 2.4
{2}	4	
{3}	3	
{4}	4	
{5}	2	Discard < 2.4
{6}	2	Discard < 2.4

New_Table.

Item	freq
{2}	4
{3}	3
{4}	4

Step-2

2-item set		freq	
{1,2}	——	1	Discard < 2.4
{1,3}	——	1	"
{1,4}	——	1	"
{1,5}	——	0	"
{1,6}	——	0	"
{2,3}	——	3	
{2,4}	——	4	
{2,5}	——	2	Discard < 2.4
{2,6}	——	2	"
{3,4}	——	3	
{3,5}	——	1	Discard < 2.4
{3,6}	——	1	"
{4,5}	——	2	"
{4,6}	——	2	"
{5,6}	——	2	"

New_Table

2-Item	freq
{2,3}	3
{2,4}	4
{3,4}	3

3-Item_Set	freq	
{1,2,3}	1	Discard
{1,2,4}	1	D
{1,2,5}	0	D
{1,2,6}	0	D
{1,3,4}	1	D
{1,3,5}	0	D
{1,3,6}	0	D
{1,4,5}	0	D
{1,4,6}	0	D
{1,5,6}	0	D
{2,3,4}	3	
{2,3,5}	1	D
{2,3,6}	1	D
{2,4,5}	2	D
{2,4,6}	2	D
{2,5,6}	2	D
{3,4,5}	1	D
{3,4,6}	1	D
{3,5,6}	1	D
{4,5,6}	2	D

The new table is.

{2,3,4}

Let's check for confidence = 80%.

The final set is

: {2, 3, 4}

→ non empty sets.

{{2}, {3}, {4}, {2,3}, {2,4}, {3,4}}

Rule 1 $\overset{A}{\{2\}} \longrightarrow \overset{B}{\{3,4\}}$

$$\text{Support} = \frac{\text{sup}(A \cup B)}{\text{Total customers}} = \frac{\{2,3,4\}}{4} \rightarrow \text{freq.}$$

$$= \frac{3}{4} = 0.75$$

$$\text{Confidence} = \frac{\text{Sup}(A \cup B)}{\text{Sup}(A)} = \frac{\text{Sup}(2,3,4)}{\text{Sup}(A)} = \frac{0.75}{1}$$

$$\text{Sup}(A) = \underset{2}{\downarrow} \frac{4}{4} = 1$$

$= 75\% < 80\%$.

invalid.

Rule 2 $\{3,4\} \rightarrow \{2\}$

$$\text{Supp} = \frac{3}{4}.$$

$$\text{Confidence} = \frac{3}{4} \times \frac{4}{3} = 100\%.$$

Valid.

Rule 3 $\{3\} \rightarrow \{2,4\}$

$$\text{Supp} = \frac{3}{4}$$

$$\text{Confidence} = \frac{3}{4} \times \frac{4}{3} = 100\%,$$

valid.

Rule 4 $\{2,4\} \rightarrow \{3\}$

Supp $= \frac{3}{4}$

Confidence $= \frac{3}{4} \times \frac{4}{4} = 0.75$

75% < 80%

Invalid.

Rule 5 $\{4\} \rightarrow \{2,3\}$

Supp $= \frac{3}{4}$ Confidence $= \frac{3}{4} \times \frac{4}{4} = 0.75 < 0.8$

invalid.

Rule 6 $\{2,3\} \rightarrow 4$

Supp $= \frac{3}{4}$ Confidence $= \frac{3}{3} \times \frac{4}{3} = 100\%$

valid.

Validity.

3-Set Item

$\{2,3,4\}$

$\{3,4\} \rightarrow \{2\}$

$\{3\} \rightarrow \{2,4\}$

$\{2,3\} \rightarrow \{4\}$

2) Assume $A(2,2)$ and $c(1,1)$ are centers of the two clusters.
C_1 C_2

Distance

$A(2,2)$ & $C_1(2,2)$

$P(A, C1)$

$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

$= \sqrt{(2-2)^2+(2-2)^2} = 0.$

$P(A, C_2)$

$= \sqrt{(1-2)^2+(1-2)^2}$

$= \sqrt{2} = 1.14$

$P(B, C_1)$

$= \sqrt{(2-3)^2+(2-2)^2}$

$= 1$

$P(B, C_2)$

$= \sqrt{(1-3)^2+(1-2)^2} = 2.24$

$P(C_1, C_1)$

$= \sqrt{(2-1)^2+(2-1)^2} = 1.401$

$P(C_1, C_2)$

$= \sqrt{(1-1)^2+(1-1)^2} = 0$

$P(D, C_1)$

$= \sqrt{(2-3)^2+(2-1)^2}$

$= \sqrt{1+1} = 1.41$

$P(D, C_2)$

$\sqrt{(1-3)^2+(1-1)^2} = 2$

$P(E, C_1)$

$= \sqrt{(2-1.5)^2+(2-0.5)^2} = 1.58$

$P(E, C_2) = 0.71$

Table.

	C_1	C_2	Point
$A(2,2)$	0	1.14	C_1
$B(3,2)$	1	2.24	C_1
$C(1,1)$	1.41	0	C_2
$D(3,1)$	1.41	2	C_1
$E(1.5,0.5)$	1.58	0.71	C_2

First cluster.

A(2, 2)
B(3, 2)
D(3, 1)

$C_1 = \frac{2+3+3}{3}, \frac{2+2+1}{3} = (2.67, 1.67)$

Second cluster

C(1, 1)
E(1.5, 0.5)

$C_2 = \frac{1+1.5}{2}, \frac{0.5+1}{2} = (1.25, 0.75)$

2nd time.

	$C_1(2.67, 1.67)$	$C_2(1.25, 0.75)$	Belongs to.
A(2, 2)	0.75	1.46	C_1
B(3, 2)	0.47	2.15	C_1
C(1, 1)	1.80	0.35	C_2
D(3, 1)	0.75	1.77	C_1
E(1.5, 0.5)	1.65	0.35	C_2

As the clusters are same So we stop here.

The final clusters are :—

C_1 = A(2, 2), B(3, 2), D(3, 1)

C_2 = { C(1, 1), E(1.5, 0.5)

9) a) Entropy$_{(t)} = -\sum_{i=0}^{i=1} P(i|t) \log_2 P(i|t) = -\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) \log_2\left(\frac{5}{9}\right)\right]$

$P(i/t) = \frac{4}{9}$ → No. of true classes

$\frac{4}{9}$ → Total classes.

$= -[-0.52 + (-0.47)]$

$= 0.99$ Ans.

For True among the true classes.

b) Entropy$(a_1) = -\sum_{i=0}^{i-1} P(i|t) \log_2 p(i|t)$

$= -\left[\left(\frac{4}{9}\right) \log_2\left(\frac{4}{9}\right)\right]$

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

$= 0.811$

Entropy$(a_2) = \left[\left(\frac{2}{5}\right) \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) \log_2\left(\frac{3}{5}\right)\right]$

$= 0.971$

For False

Entropy$(a_1) = -\left[\left(\frac{1}{5}\right) \log_2\left(\frac{1}{5}\right) + \left(\frac{4}{5}\right) \log_2\left(\frac{4}{5}\right)\right]$

$= -[(-0.46) + (-0.26)]$

$= 0.72$

Entropy$(a_2) = \left[\left(\frac{2}{4}\right) \log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) \log_2\left(\frac{2}{4}\right)\right]$

$= 1$

Information Gain

a_1 : $\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$

$= 0.991 - \left[\left(\frac{4}{9} \right) \times 0.811 + \frac{5}{9} \times 0.721 \right] = 0.229$

a_2 : $\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$

$= 0.991 - \left[\left(\frac{5}{9} \right) * 0.971 + \left(\frac{4}{9} \right) * 1 \right] = 0.0007$

10% i) Accuracy $= \frac{TP + TN}{(TP + FP + FN + TN)}$

$= \frac{105 + 10}{105 + 10 + 55 + 15} = 0.6216$

ii) Error rate = 1 − accuracy. or $\frac{FP + FN}{\text{Total}}$

10%

	Class	Predicted No	Predicted Yes
Actual	No	TN(55)	FP(15)
	Yes	FN(10)	TP(105)

i) Accuracy $= \frac{TP + TN}{\text{Total}} = \frac{105 + 55}{185} = 0.86.$

ii) Error rate = 1 − accuracy or $\frac{FP + FN}{\text{Total}}$

$= 1 - 0.86$ or $\frac{15 + 10}{185}$

$= 0.14$

iii) Precision $= \frac{TP}{TP+fP} = \frac{105}{120} = \underline{0.875}$

iv) Recall $= \frac{TP}{TP+fN} = \frac{105}{115} = \underline{0.913}$.