NAME:- SHIVPRASAD · C· PREMARAJAN
ROLL NO:- 9696
SUB : DWM

Marvel 26/10/23

Assignment 1

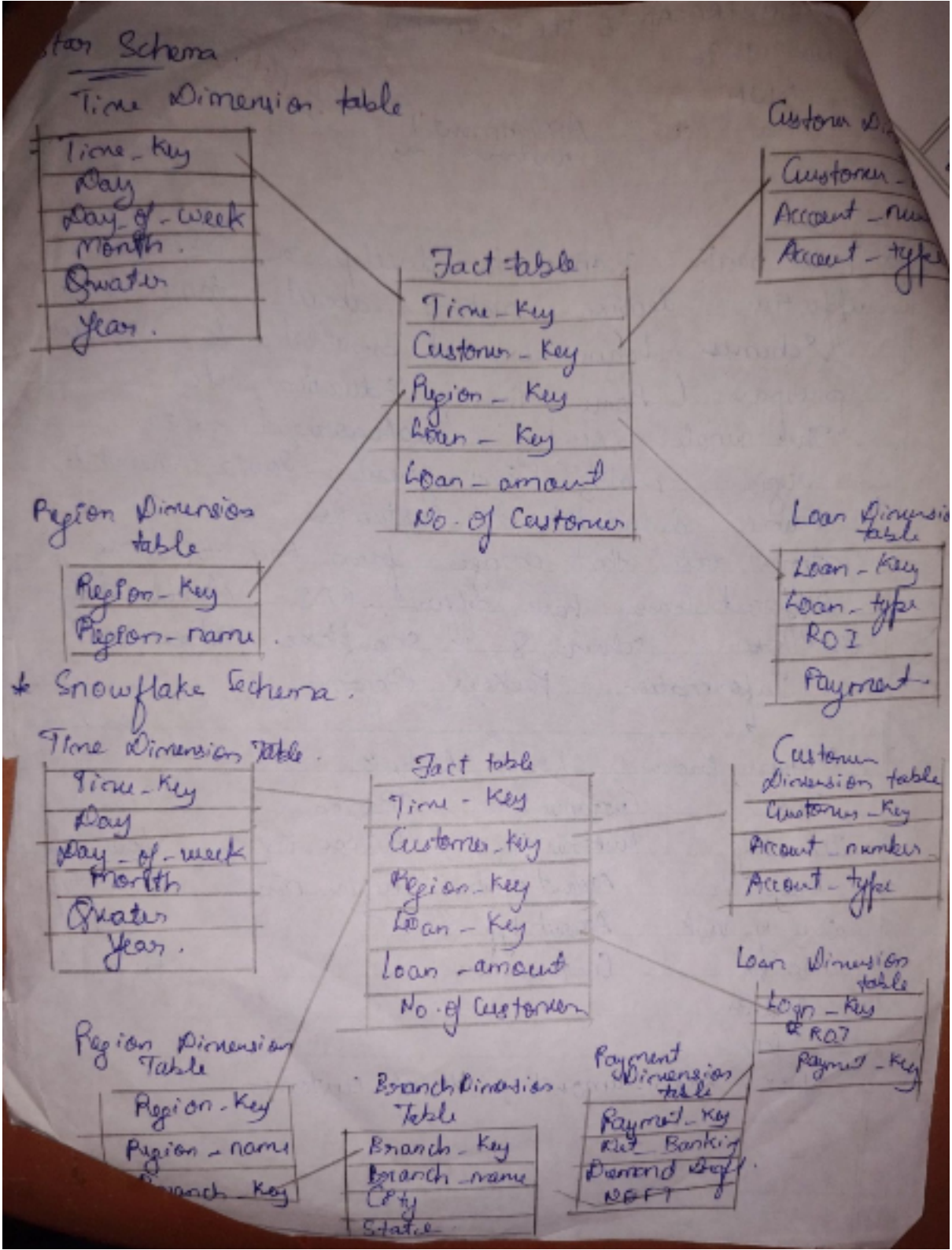
Q.1] A bank wants to develop a DW for effective decision making about their loan Schemes. Loans are provided for various purpose (Home Loan, Education etc). The whole country is categorized into regions namely (East, west, South, North). Loan is distributed to Customers at the interest rate that changes from time to time. Different loans have different ROI. Clearly design the Star schema & snowflake schema

⇒ Information Package Diagram.

Business Process: Loan Analysis.

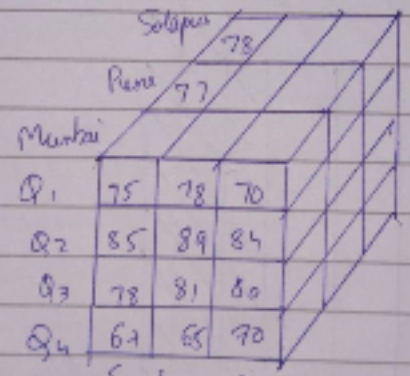
Time	Customer	Region	Loan
Time_key	Customer_Key	Region_Key	loanKey
Day	Account_number	Region_name	Loan_type
day_of_week	Account_type		ROI
month	Customer_name		Payment
Quarter			
Year			

Fact: loan_amount, No. of customers.


Star Schema
Time Dimension table
Time_Key
Day
Day_of_week
Month
Quater
Year
Fact table
Time_Key
Customer_Key
Region_Key
Loan_Key
Loan_amount
No_of Customer
Region Dimension table
Region_Key
Region_name
Loan Dimension table
Loan_Key
Loan_type
ROI
Payment
* Snowflake Schema
Time Dimension Table
Time_Key
Day
day_of_week
Month
Quater
Year
Fact table
Time_Key
Customer_Key
Region_Key
Loan_Key
Loan_amount
No_of Customer
Customer Dimension table
Customer_Key
Account_number
Account_type
Loan Dimension table
Loan_Key
ROI
Region Dimension Table
Region_Key
Region_name
Branch Dimension Table
Branch_Key
Branch_name
City
State
Payment Dimension table

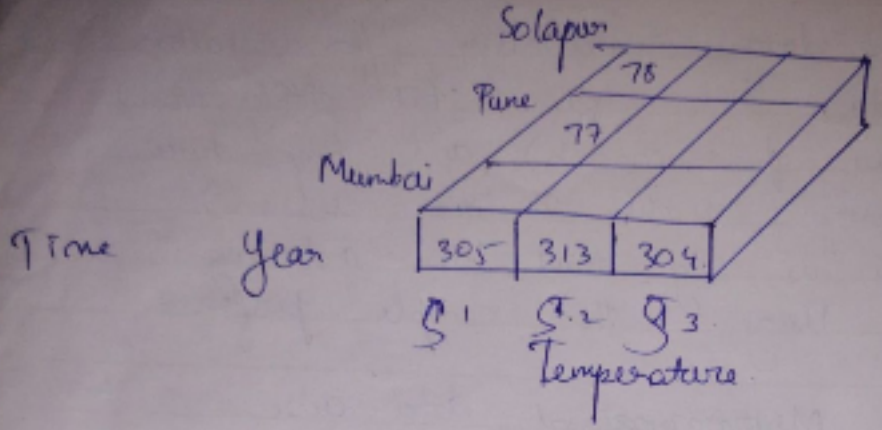
2] Consider a data warehouse for weather related data like region (state, city, area), time (year, qtr, month) and temperature (temp_category, weather_condition, season) with facts as avg_unit_int and avg humidity. Using this example perform following.

1) Draw a Multidimensional data cube.
2) perform following OLAP operations.
- Roll-up.
- Drill-down.
- Slice
- Dice
- Pivot.

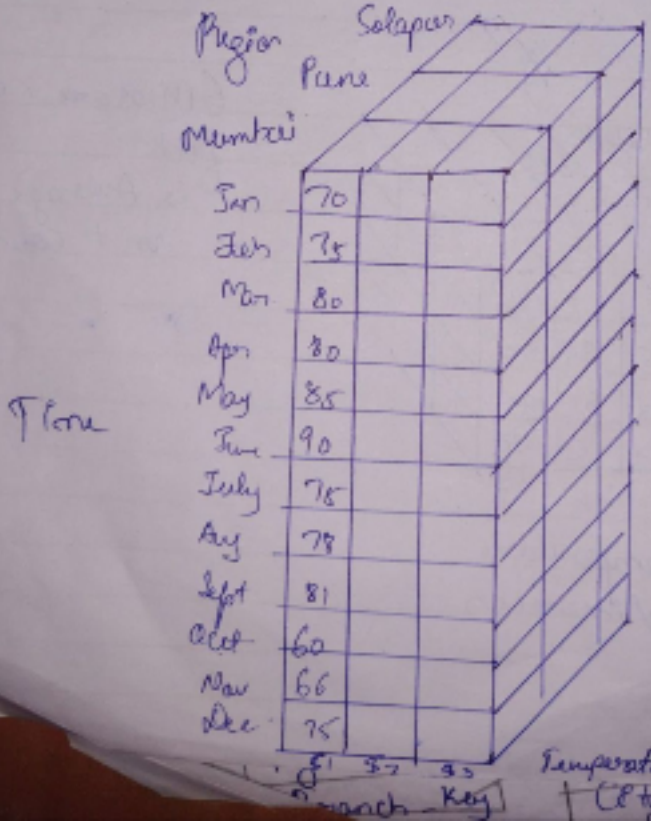


(Measure i.e fact is Average unit count).

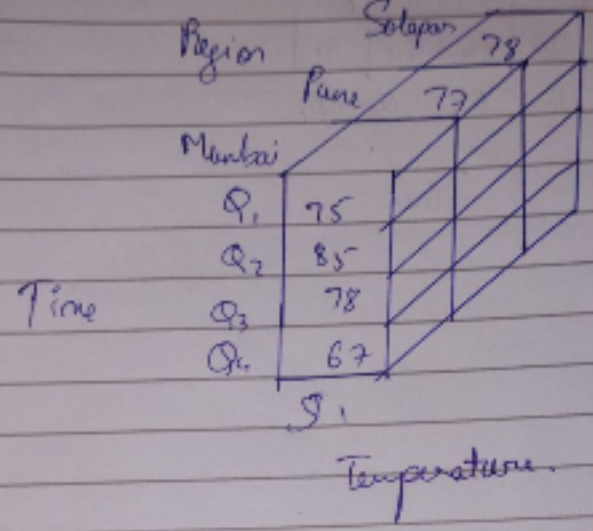
1) Roll-up (Time)



2) Drill-down (Time)

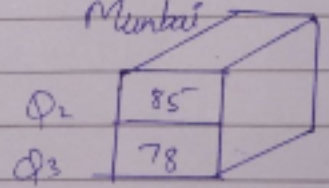


Slice (Temperature = T1)



Dice
=

Time = Q2 & Q3
Temperature = T1
Region = Mumbai



Pivot

Temperature	Q1	Q2	Q3	Q4
S1	75	85	78	67
S2	78	89	81	65
S3	70	84	80	70

Time

DWM Assignment no - 2

Q1] The following data (in increasing order) for the attribute age : 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.

⇒

Partition into (equi-depth) bins.

Bin1 :- 13, 15, 16, 16, 19, 20, 20, 21, 22.

Bin2 :- 22, 25, 25, 25, 25, 30, 33, 33, 35.

Bin3 :- 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Smoothing by bin mean :-

Bin1 :- 18, 15, 18, 18, 18, 18, 18, 18, 18.

Bin2 :- 28, 28, 28, 28, 28, 28, 28, 28, 28.

Bin3 :- 44, 44, 44, 44, 44, 44, 44, 44, 44.

b) Smoothing by bin median :-

Bin1 :- 19, 19, 19, 19, 19, 19, 19, 19, 19.

Bin2 :- 25, 25, 25, 25, 25, 25, 25, 25, 25

Bin3 :- 40, 40, 40, 40, 40, 40, 40, 40, 40.

Q.2] Explain with examples Sampling methods:-

a) Simple random sample without replacement

→ Simple random sampling without replacement (SRSWOR) is a sampling method in which each unit in the population has an equal chance of being selected and each unit can only be selected once.

Eg:- Suppose in class of 20 students and you want to select a sample of 5 students to participate in debate. We can use SRS WOR to select the sample.

Step1: Write each student's name on slip of paper.
Step2: put the slips of paper in a hat or bowl & mix them.
Step 3:- Draw 5 slips of paper without replacement.
Step4:- The students whose name are drawn will be the sample.

b) Simple random sample with replacement.

→ Simple random Sampling with replacement (SRSWR) is a sampling method in which each unit in the population has an equal chance of being selected and each unit can be selected multiple times.

Eg:- Suppose we have a bag of 10 marbles, 5 red & 5 blue. You want to draw a sample of 3 marbles with replacement.

Step1:- Randomly select one marble from bag.
Step2:- Record the color of marble & put it back in bag.
Step3:- Repeat steps 1 & 2 until you have a sample of 3 marbles P/O :- Red, Blue, Red.

c) Cluster sample.

→ Cluster sampling is a sampling method in which population is divided into groups or clusters. and a random sample of cluster is selected. All units within the selected clusters are then included in the sample.

Eg.

→ Suppose we want to survey the opinion of students at a university. We could divide the university into the clusters by college. & then randomly select a sample of college & survey all the students within these college.

d) Stratified Sample.

→ A Stratified Sample is a sample that is divided into groups or strata based on specific characteristics The characteristics used to stratify the sample is called the stratification variable. Once the sample is stratified, a random sample is taken from each stratum.

Eg.

Suppose you want to survey the opinions of student at a university. Stratify by gender (M/F). Then randomly select the sample of Male & female from each college at university.

Q.3] Normalise the following group of data.

200, 300, 400, 500, 1000.

a) Min-max normalization by setting min 0 & max 1.

min = 200, max = 1000, new-min = 0, new-max = 1.

$$X = \frac{x - min}{max - min}(new\_max - new\_min) + (new\_min)$$

$$x_1 = \frac{200 - 200}{1000 - 200}(1-0) + 0 = 0$$

$$x_2 = \frac{300 - 200}{1000 - 200}(1-0) + 0 = 0.125$$

$$x_3 = \frac{400 - 200}{1000 - 200}(1-0) + 0 = 0.25$$

$$x_4 = \frac{500 - 200}{1000 - 200}(1-0) + 0 = 0.5$$

$$x_5 = \frac{1000 - 200}{1000 - 200}(1-0) + 0 = 1$$

Normalised values are 0, 0.125, 0.25, 0.5, 1.

b) Z - Score Normalization.

→

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{200 + 300 + 400 + 600 + 1000}{5} = 500$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$$\sigma = 282.84$$

$$Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{200 - 500}{282.84} = -1.0606$$

$$Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{300 - 500}{282.84} = -0.7071$$

$$Z_3 = \frac{x_3 - \mu}{\sigma} = \frac{400 - 500}{282.84} = -0.3535$$

$$Z_4 = \frac{x_4 - \mu}{\sigma} = \frac{600 - 500}{282.84} = 0.3535$$

$$Z_5 = \frac{x_5 - \mu}{\sigma} = \frac{1000 - 500}{282.84} = 1.7677$$

c) Normalization by decimal scaling.

$x' = \frac{x}{10^j}$ where j is the smallest integer such that $\max(|v'|) < 1$

10^j is 10000 ($j = 4$ as we have maximum 4 digit number in the range)

$x_1' = \frac{200}{10^4} = \frac{200}{10000} = 0.02$

$x_2' = \frac{300}{10^4} = \frac{300}{10000} = 0.03$

$x_3' = \frac{400}{10^4} = \frac{400}{10000} = 0.04$

$x_4' = \frac{600}{10^4} = \frac{600}{10000} = 0.06$

$x_5' = \frac{1000}{10^4} = \frac{1000}{10000} = 0.1$

Normalized values are 0.02, 0.03, 0.04, 0.06, 0.1

0.4.

a) Min-Max normalization to transform the value 35 for age onto the range [0.0, 1.0]

$x' = \frac{x - min}{max - min} [new\_max - new\_min] + [new\_min]$

Min = 13, Max = 70, new_Min = 0.0, new_Max = 1.0

$$x' = \frac{35 - 13}{70 - 13} [1.0 - 0.0] + [0.0]$$

$$x' = 0.3859.$$

b]. Z- Use Z-Score normalization to transform the value 35 for age, where Standard deviation of age is 12.94 yrs.

→

$$Z = \frac{x - \mu}{\sigma}, \quad \mu = 29.96, \quad \sigma = 12.94.$$

$$\therefore Z = \frac{35 - 29.96}{12.94} = 0.3894.$$

c]. Use normalization by decimal scaling to transform the value 35 for age.

→

$$x' = \frac{x}{10^j}$$

10^j is 100 (j = 2 We have maximum 2 digit number in range)

$$\therefore x' = \frac{35}{10^2} = 0.35$$

d] Min-Max normalization reduce the dataset in the value range of 0 to 1 reducing the outliers. Hence its the better technique.

Q.5]

a] Stepwise forward selection.

→ i] Start with empty set of attributes.

ii] Determine the best of original attributes and add it to set.

iii] At each step, find the best of remaining original attributes and add it to set.

Ex: Initial attribute set

{A1, A2, A3, A4, A5, A6}.

Initial reduced set {}

⇒ {A1} - (most imp)

⇒ {A1, A4}

⇒ Reduced attribute set {A1, A4, A6}

b] Stepwise backward selection.

→ i) Starts with full set of attributes.

ii] At each step, it removes the worst attributes remaining in set.

Ex: Initial attribute set

{A1, A2, A3, A4, A5, A6}

Initial reduced set {A1, A2, A3, A4, A5, A6} - {drop last imp A2 & A3 attr}

⇒ Reduced attribute set {A1, A4, A6}.

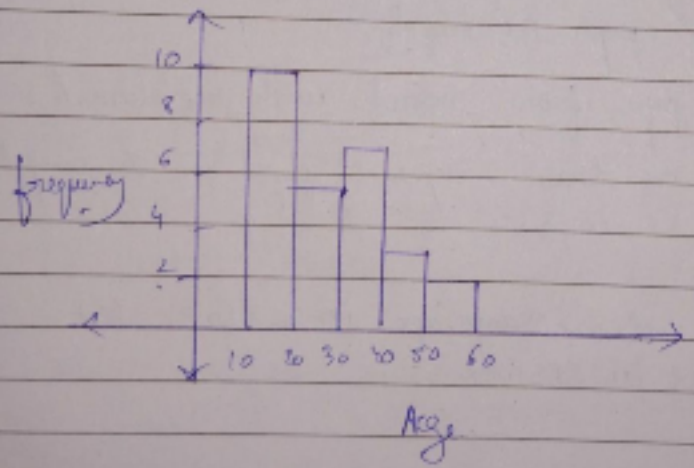
c) A combination of forward selection & backward selection.

→ i] The procedure combines & selects the best attributes & removes the worst among the remaining attributes.

ii] For all above method stopping criteria is different and it requires a threshold on the measure used to stop the attribute selection process.

Q.6]

a) Plot an equal width histogram of width 10



6) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR.

→ a) Simple Random Sampling without Replacement (SRSWOR)

1) Write down of the age attribute on a separate slip of paper.
2) Put all the slips of paper in bowl.
3) Mix the slips of paper
4) Draw 5 slips of paper from bowl.
5) The age attribute on drawn slips of paper
   P/O :- 20, 22, 25, 35, 46.

b) Simple Random Sampling with Replacement (SRSWR)

1) Write each of the age attribute on a separate slip of paper.
2) Put all the slips of paper in bowl.
3) Mix the slips paper thoroughly.
4) Draw 5 slips of paper from bowl, with replacement.
5) The age attribute on drawn slips of paper are the sample
   P/O :- 25, 25, 35, 20, 35.

Note:- possible to draw same age in SRSWOR but not in SRSWOR.

c) Cluster Sampling & Stratified Sampling. Use same of size 5 and the strata "youth", "middle aged" and "senior"

→ a) Cluster Sampling.

1) Cluster 1: Youth (age < 25)
2) Cluster 2: Middle-age (age >= 25 & age < 65)
3) Cluster 3: Senior (age >= 65)

Next we can randomly select one cluster from each of the three strata. Finally we can select all the age attributes from selected cluster to form our sample.

e.g.: Cluster 1: 19
Cluster 2: 30, 35
Cluster 3: 70.

b) Stratified sampling

1) Cluster 1: Youth (age < 25)
2) Cluster 2: middle-age (age >= 25 and age < 65)
3) cluster 3: Senior (age >= 65)

Next, we can randomly select 5 age attributes from each stratum.

Cluster 1: 15, 19
Cluster 2: 25, 30, 35
Cluster 3: 70

Note :-

Cluster sampling & stratified sampling are two different sampling methods cluster sampling is a method of selecting groups of individual while stratified sampling is a method of selecting individual from different groups

Name:- Chavarattil Shivprasad Bramarajan
Roll no:- 9696
FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

Marvellous

Assignment 3.

Q.1] Apply Naive Bayesian Classification Algorithm on following data

→ Tuple (Homeowner = Yes, Status = Employed, Income = Avg)

$P(\text{Defaulted} = \text{'yes'}) = 3/10$
$P(\text{Defaulted} = \text{'No'}) = 7/10$

To calculate probability

$$P(c_i/x) = \frac{p(x/c_i)\,p(c_i)}{p(x)}$$

$\because p(x)$ is constant for all classes
$\therefore p(c_i/x) = p(x/c_i)\,p(c_i)$

$p(x/c_i) = p(x/\text{Defaulted} = \text{'yes'})$
$= p(\text{Homeowner} = \text{'yes'} / \text{Defaulted} = \text{'yes'}) \times$
$= p(\text{Status} = \text{'Employed'} / \text{Defaulted} = \text{'yes'}) \times$
$= p(\text{Income} = \text{'Average'} / \text{Defaulted} = \text{'yes'})$
$= 0 \times \frac{2}{3} \times \frac{0}{3}$
$= 0$

$P(\text{Defaulted} = \text{'yes'} / x) =$
$p(x/\text{Defaulted} = \text{'yes'}) \times P(\text{Defaulted} = \text{'yes'})$
$= 0 \times 3/10 = 0.$

$P(X|C_i) = P(X|\text{Defaulted}=\text{'No'})$

$= P(\text{Homeowner}=\text{'Yes'}|\text{Defaulted}=\text{'No'}) \times$

$= P(\text{Status}=\text{'Employed'}|\text{Defaulted}=\text{'No'}) \times$

$= P(\text{Income}=\text{'Average'}|\text{Defaulted}=\text{'No'})$

$= \frac{2}{7} \times \frac{2}{7} \times \frac{1}{7} = \frac{6}{7^3}$

$\therefore P(\text{Defaulted}=\text{'No'}|X)$

$= P(X|\text{Defaulted}=\text{'No'}) \times P(\text{Defaulted}=\text{'No'})$

$= \frac{6}{7^3} \times \frac{7}{10} = \frac{6}{490}$

$= 0.012$

$\because P(\text{Defaulted}=\text{'No'}|X) > P(\text{Defaulted}=\text{'Yes'}|X)$

$\therefore$ Therefore the naive bayes classifier predicts Defaulted = 'No' for sample X.

Q.2. Apply naive bayesian Classification Algorithm on following data.

tuple (District = 'Rural', House Type = 'Semi detached', Income = 'low', Previous customer = 'No').

→

P(Outcome = 'Nothing') = $\frac{5}{14}$.

P(Outcome = 'Responded') = $\frac{9}{14}$.

To Calculate probability

$$P(c_i/x) = \frac{P(x|c_i)\,P(c_i)}{P(x)}$$

∵ P(x) is constant for all classes

$$\therefore P(c_i/x) = P(x|c_i)\,P(c_i)$$

∴ P(X|Ci) = P(X| Outcome = 'Nothing') =
= P(District = 'Rural' | Outcome = 'Nothing') ×
= P(House Type = 'Semi detached' | Outcome = 'Nothing') ×
= P(Income = 'low' | Outcome = 'Nothing') ×
= P(Previous Customer = 'No' | Outcome = 'Nothing')
= $0 \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5}$

$\therefore P(\text{Outcome} = \text{'Nothing'} \mid x) = P(x \mid \text{Outcome} = \text{'Nothing'})$

$= x\, P(\text{Outcome} = \text{'Nothing'})$

$= 0 \times \frac{5}{14} = 0$

$\therefore P(x \mid C_i) = P(x \mid \text{Outcome} = \text{'Responded'})$.

$= P(\text{District} = \text{'Rural'} \mid \text{Outcome} = \text{'Responded'}) \times$

$= P(\text{House Type} = \text{'Semi detached'} \mid \text{Outcome} = \text{'Responded'}$

$= P(\text{Income} = \text{'low'} \mid \text{Outcome} = \text{'Responded'}) \times$

$= P(\text{Previous Customer} = \text{'No'} \mid \text{Outcome} = \text{'Responded'})$

$= \frac{4}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9}$

$= \frac{64}{729}$

$P(\text{Outcome} = \text{'Responded'} \mid x) = P(x \mid \text{Outcome} = \text{'Responded'})$

$\times P(\text{Outcome} = \text{'Responded'})$

$= \frac{64}{729} \times \frac{9}{14} = 0.056$.

$\therefore P(\text{Outcome} = \text{'Responded'} \mid x) > P(\text{Outcome} = \text{'Nothing'} \mid x)$

Therefore the naive bayes classifier predicts Outcome = 'Responded' for given sample x.

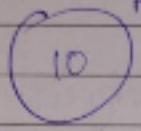
Q.3] Cluster the following data (10, 16, 6, 29, 16, 38, 42, 8, 11).

Where $K=2$

→ Let centroids be $K_1 = 10$ & $K_2 = 29$.

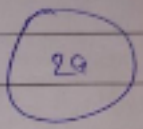
Iteration 1

K_1 $m_1 = 10$



$K_1 = \{10, 15, 6, 16, 8, 11\}$

K_2 $m_2 = 29$



$K_2 = \{29, 38, 42\}$

Update the centroids.

$\therefore C_1 = \frac{10+15+6+16+8+11}{6}$

$= 11$

$C_2 = \frac{29+38+42}{3}$

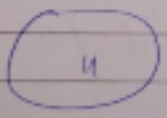
$= 36.33.$

∴ New centroids & assumed centroids are not equal, therefore repeat the steps

Iteration 2

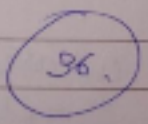
K_1

$m_1 = 11$



K_2

$m_2 = 36$



update the centroids

$\therefore C_1 = \frac{10+15+6+16+8+11}{6}$

$= 11.$

$C_2 = \frac{24+38+42}{3}$

$= 36.$

∵ The centroids did not change in this iteration

∴ The clusters are

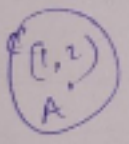
$K_1 = \{10, 15, 6, 16, 8, 11\}$

$K_2 = \{29, 38, 42\}$

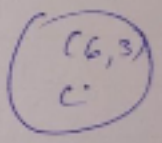
Q.4] Cluster the following data objects A(1,2) B(3,4) C(6,3) D(2,3) E(4,4) F(1,5) where K=2.

→ Let centroids be $C_1 = A(1,2)$, $C_2 = C(6,3)$.

K_1



K_2



$K_1 = \{A, B, D, F\}$

update the centroid

$C_1 = \left\{\frac{1+3+2+1}{4}, \frac{2+4+3+5}{4}\right\}$

$= (1.73, 3.5)$

$K_2 = \{C, E\}$

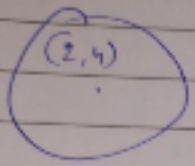
$C_2 = \left(\frac{6+4}{2}, \frac{3+4}{2}\right)$

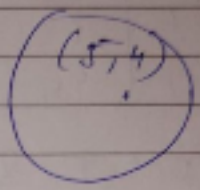
$= (5, 3.5).$

∵ New centroids & assumed centroids are not same, therefore repeat the step

Iteration 2:

K_1 K_2





$K_1 = \{A, B, D, F\}$

$K_2 = \{C, E\}$

update the centroid

$C_1 = \left(\frac{1+3+2+1}{4}, \frac{2+4+3+5}{4}\right)$

$C_1 = (1.75, 3.5) \approx (2,4)$

$C_2 = \left(\frac{6+4}{2}, \frac{3+4}{2}\right)$

$= (5, 3.5) \approx (5,4)$

∵ The centroids did not change in the iteration

∴ The clusters are

$K_1 = \{A, B, D, F\}$

$K_2 = \{C, E\}$.

Q.5] Cluster the following data using agglomerative hierarchical clustering technique.

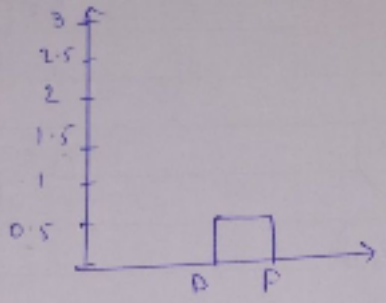
	x_1	x_2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

1) Single Linkage

→ Distance matrix

	A	B	C	D	E	F
A	0					
B	1	0				
C	8	7	0			
D	5	4	3	0		
E	6	5	2	1	0	
F	4.5	3.5	3.5	0.5	1.5	0

Point F & D has minimum distance 0.5.



Re-computing the distance matrix

	A	B	C	(D,F)	E
A	0				
B	1	0			
C	8	7	0		
(D,F)	4.5	3.5	3	0	
E	6	5	2	1	0

dist ((D,F), A)

= min (dist (D,A), dist (F,A))

= min (5, 4.5)

= 4.5

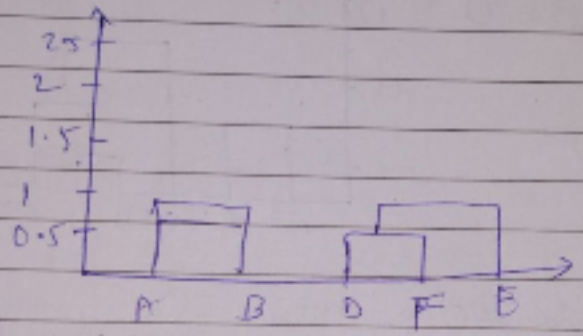
dist ((D,F), B) = min (dist (D,B), dist (F,B))

= min (4, 3.5)

= 3.5

dist {(D,F), C} = min (dist (D,C), dist (F,C))
= min (3, 3.5)
= 3

dist ((D,F), E) = min (dist (D,E), dist (F,E))
= min (1, 1.5)
= 1



Recomputing the distance matrix

dist ((A,B), C)
= min (dist (A,C), dist (B,C))
= min (8, 7)
= 7.

((D,F),E) C (A,B)

(A,B)	0		
C	7	0	
((D,F),E)	3.5	2	0

dist [(A,B), ((D,F),E)]

= min [dist (A,E), dist (B,E), dist (A,D), dist (A,F),
dist (B,D), dist (BF)]

= min (6, 5, 5, 4.5, 4, 3.5)

= 3.5

dist [((D,F),E), C]

= min [dist (D,C), dist (F,C), dist (E,C)]

= min [3, 3.5, 2]

= 2

Re computing the distance matrix

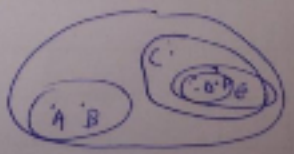
	(A,B)	((D,F),E,C)
(A,B)	0	
((D,F),E,C)	3.5	0

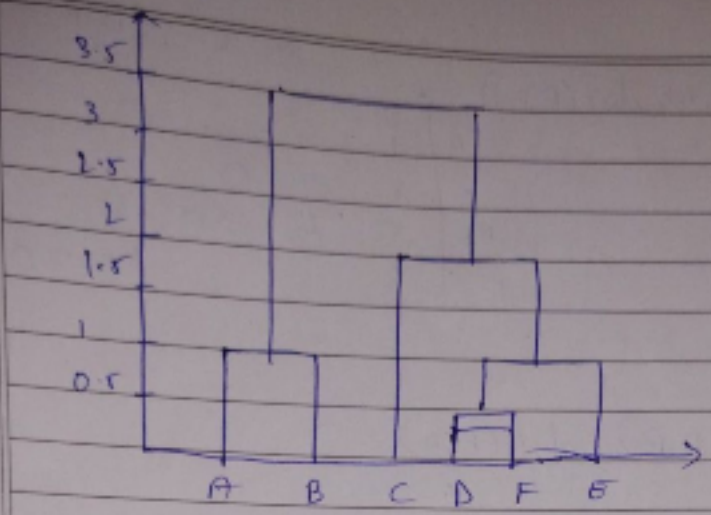
dist [(A,B), (D,F), E, C]

= min (dist (A,E), dist (A,C), dist (B,E), dist (B,C),
dist (A,D), dist (A,F), dist (B,D), dist (B,F))

= min (6, 5, 5, 7, 5, 4.5, 4, 3.5)

= 3.5

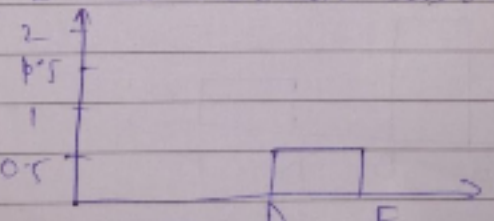




11) Complete linkage distance matrix

	A	B	C	D	E	F
A	0					
B	1	0				
C	8	7	0			
D	5	4	3	0		
E	6	5	2	1	0	
F	4.5	3.5	3.5	0.5	1.5	0

Pair F & D has minimum distance 0.5



Re Computing the distance matrix

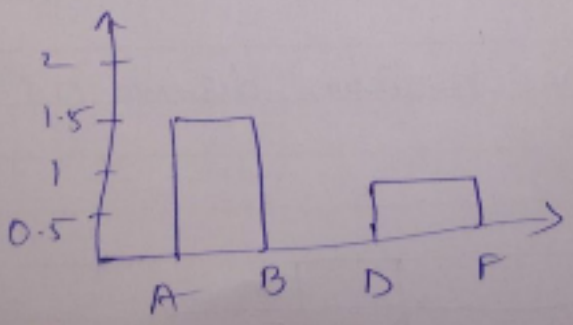
dist((D,F),A)
= max(dist(D,A), dist(F,A))
= max(5, 4.5)
= 5

	A	B	C	(D,F)	E
A	0				
B	1	0			
C	8	7	0		
D,F	5	4	3.5	0	
E	6	5	2	1.5	0

dist((D,F),B)
= max(dist(D,B), dist(F,B))
= max(4, 3.5)
= 4

dist((D,F),C) = max(dist(D,C), dist(F,C))
= max(3, 3.5)
= 3.5

dist((D,F),E) = max(dist(D,E), dist(F,E))
= max(1, 1.5)
= 1.5



Re computing the distance matrix

dist [(A,B), C) = max (dist (A,C), dist (B,C))
= max (8, 7)
= 7.

	(A,B)	C	(D,F)	E
(A,B)	0			
C	7	0		
(D,F)	5	3.5	0	
E	6	2	1.5	0

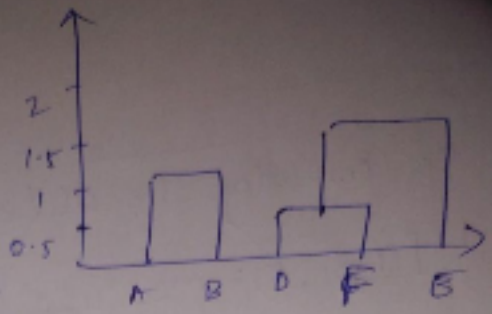
dist ((A,B), (D,F))
= max (dist (A,D), dist (A,F), dist (B,D) dist (B,F))
= max (5, 4.5, 4, 3.5) ∴
= 5

dist ((D,F), C) = max (dist (D,C), dist (E,C))
= max (3, 3.5)
= 3.5

dist ((A,B), E) = max [dist (A,E), dist (B,E))
= max (6, 2)
= 6

dist ((D,F), E)) = max (dist (D,E), dist (F,E))
= max (1, 1.5)
= 1.5

Re Computing the distance matrix.



dist((A,B), (D,F), E))

= max(dist(A,E), dist(B,E), dist(A,D), dist(A,F), dist(B,D), dist(B,F))

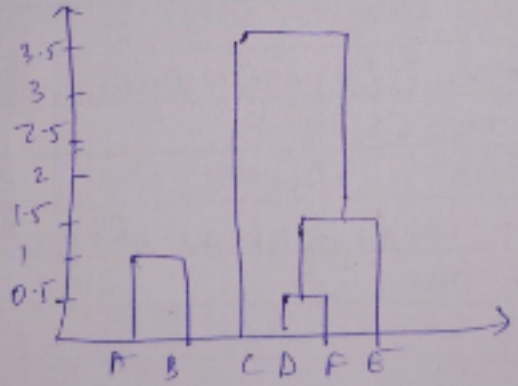
= max(6, 5, 5, 4.5, 4, 3.5)

= 6.

dist((EDF), E), C) = max (dist(D,C), dist(F,C), dist(E,C))

= max(3, 3.5, 2)

= 3.5



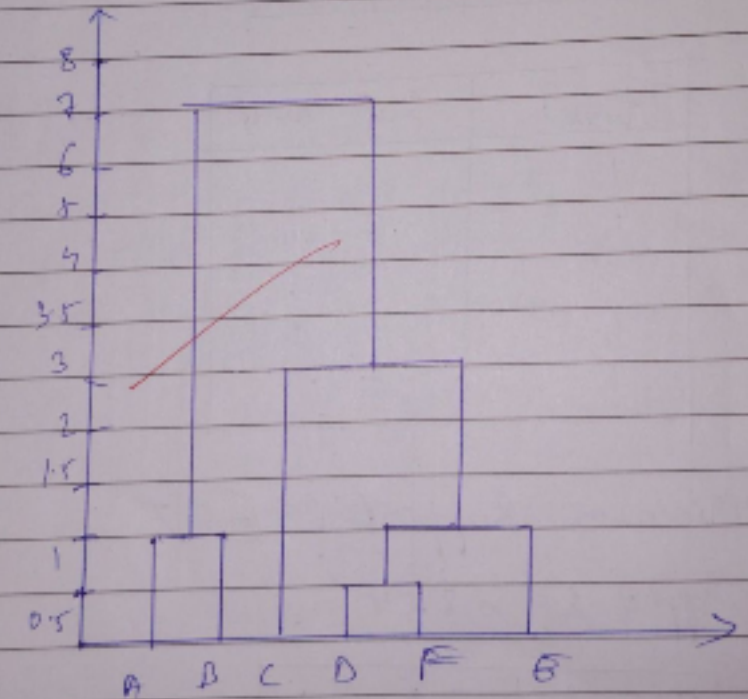
Recomputing the distance matrix

dist((A,B), (D,F), (E,C))

$$\begin{array}{c} \\ (A,B) \\ (D,F),C \end{array} \begin{array}{c} \begin{array}{cc} (A,B) & (C,D,F)\,6,C \end{array} \\ \begin{bmatrix} 0 & \\ 8 & 0 \end{bmatrix} \end{array}$$

= max (dist(A,C), dist(B,C), dist(B,E), dist(B,C), dist(A,D), dist(D,F), dist(B,D), dist(B,F))

= max (6, 8, 5+3, 5+4.5, 4, 3.5)



Q.6. Apply Apriory algorithm of following data to find frequent itemset & strong association rules minimum Support = 50 % & confidence = 80 %

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5
500	2, 3, 5
600	2, 5
700	2, 4
800	2, 3, 5

→ Step 1 :-

C_1

Itemset	Sup. count
1	2
2	6
3	5
4	2
5	6

Step 2 :-

Minimum Support is 50 %.

∴ Support = $\frac{50}{100} \times 8 = 4$.

L_1 =

Itemset	Sup. count
2	6
3	5
5	6

Step 3:-

Generate Candidate C_2 from L_1 & find the Support of 2-itemset

C_2 =

Itemset	Sup-count
2, 3	4
2, 5	6
3, 5	4.

Step 4:-

Compare C_2 generated in step 3 with support count & prune those itemsets which do not satisfy the minimum support count.

L_2 =

Itemset	Sup-count
2, 3	4
2, 5	6
3, 5	4.

Step 5:-

Generate candidate C_3 from L_2 & find the Support.

C_3 =

Itemset	Sup-count
2, 3, 5	4

Step 6 :-

Compare L_3 with minimum support count.

L_3 =

Itemset	Sup-count
2, 3, 5	4.

∴ The database contain the frequent itemsets (2, 3, 5)

Association rule	Support	Confidence	Confidence %
$2 \wedge 3 \Rightarrow 5$	4	$4/4 = 1$	100
$3 \wedge 5 \Rightarrow 2$	4	$4/4 = 1$	100
$2 \wedge 5 \Rightarrow 3$	4	$4/6 = 0.667$	66.7
$5 \Rightarrow 2 \wedge 3$	4	$4/6 = 0.067$	66.7
$2 \Rightarrow 3 \wedge 5$	4	$4/6 = 0.667$	66.7
$3 \Rightarrow 2 \wedge 5$	4	$4/6 = 0.667$	66.7

Given minimum confidence = 80 %, so only the first & second rules above are output

∴ Strong rules are

Rule 1 : $2 \wedge 3 \Rightarrow 5$

Rule 2 : $3 \wedge 5 \Rightarrow 2$

These are strong rules because Pts above confidence %