COMPUTER ENGINEERING DEPARTMENT

SUBJECT: DATA WAREHOUSING & MINING

COURSE: T.E. **YEAR: 2020-2021** **SEMESTER: VI**
DEPT: COMPUTER ENGINEERING
SUBJECT CODE: CSC603 **EXAMINATION DATE: 07/06/2021**

==========

DATA WAREHOUSING & MINING
ANSWER SHEET

NAME : AMEY MAHENDRA THAKUR

SEAT NO. : 61021145

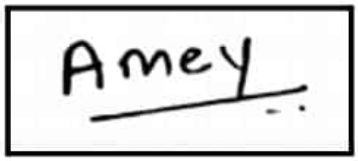
EXAM : SEMESTER VI

SUBJECT : DATA WAREHOUSING & MINING

DATE : 07-06-2021

DAY : MONDAY

STUDENT SIGNATURE:

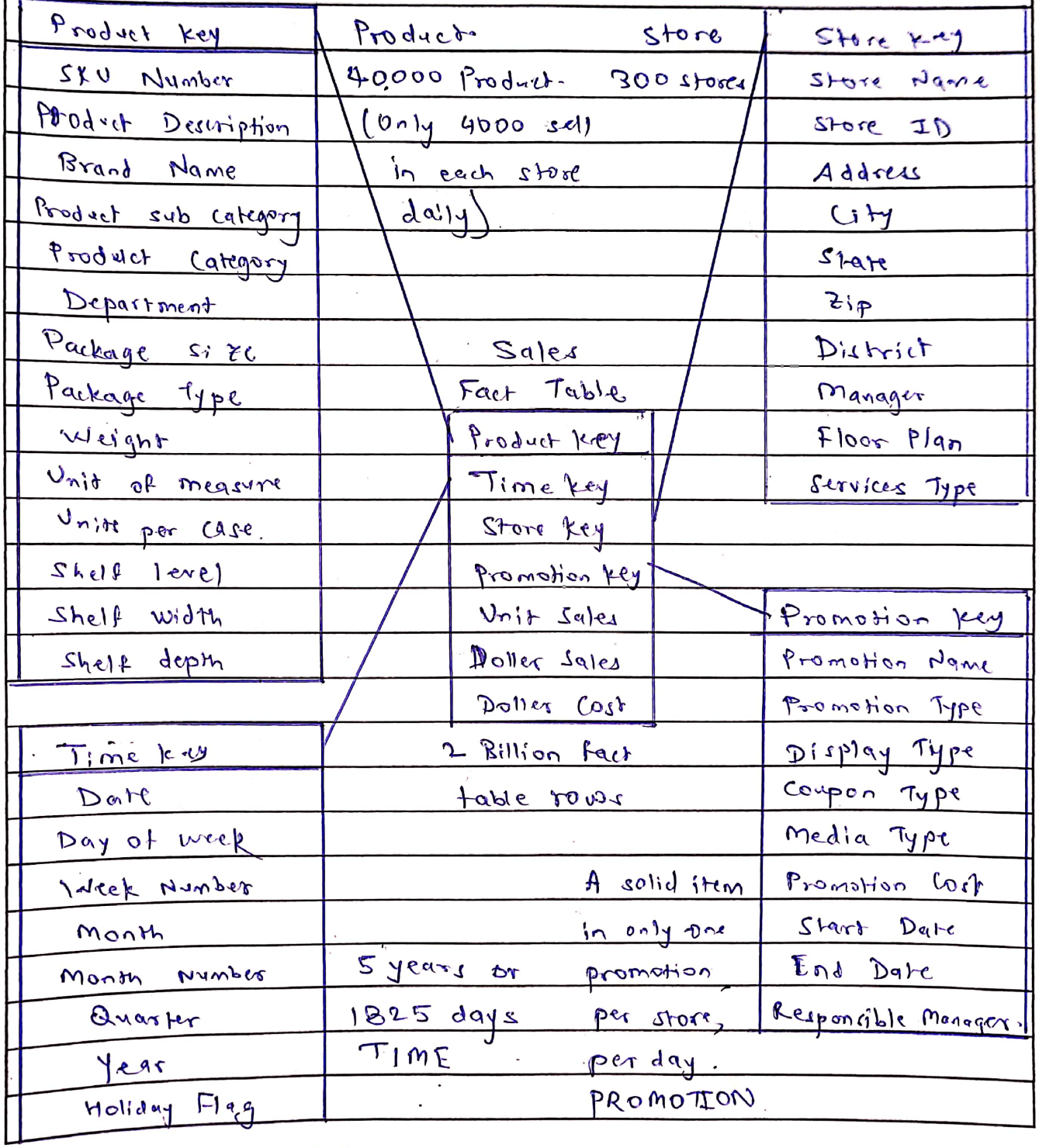


Q3

A

Soln:

(a) Star Schema



07-06-2021 STUDENT SIGNATURE: Amey

(b) Time period = 5 years x 365 days = 1825

There are 300 stores.

Each stores daily sale = 4000

Promotion = 1

Maximum numbers of fact table records
= 1825 x 300 x 4000 x 1

= 2 billion.

Q3

B]

Soln:

Adjacent matrix

$$d(i,j) = \sqrt{|x_i - x_j|^2 - |y_2 - y_1|^2}$$

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

∴ Adjacen Matrix

	A	B	C	D	E
A	0				
B	1	0			
C	1.41	2.24	0		
D	1.41	1	2	0	
E	1.58	2.12	0.71	1.58	0

Complete linkage : maximum Distance.

07-06-2021

STUDENT SIGNATURE: Amey

1st pairing (C & E)

	A	B	C, E	D
A	0			
B	1	0		
C, E	1.38	2.24	0	
D	1.41	1	2	0

Dist ((C,E), A) = max(dist(C,A), distance(E,A))
= max{1.41, 1.58}
= 1.58

Dist ((C,E), B) = max(2.24, 2.12)
= 2.24

Dist ((C,E), C) = max(2, 1.58)
= 2

Now,
2nd Pairing
A & B are having minimum closest measure value.

	A, B	C, E	D
A, B	0		
C, E	2.24	0	
D	1	2	0

07-06-2021 STUDENT SIGNATURE: Amey

Dist ((A,B), D) = max(dist (A, D), dist (B, D))
= max { 1.41, 1 }
= 1.41

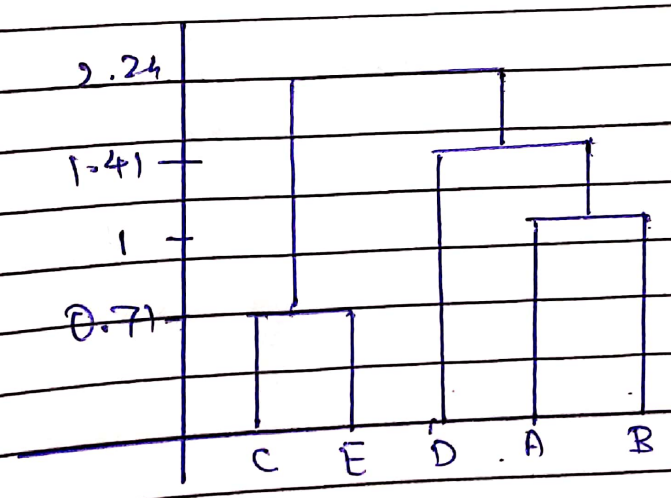
Dist ((A,B), (C, E)) = max (1.88, 1.58, 2.24, 2.12)
= ~~max (2.24)~~
= 2.24

3rd pairing
cluster (A, B) & D can be merged together

	A, B, D	C, E
A, B, D	0	
C, E	2.24	0

Dist ((A, B, D), (C, E)) = max (1.41, 1.58, 2.24, 2.12, 2, 1.58)
= 2.24

Dendogram



07-06-2021

STUDENT SIGNATURE: Amey

Q3

G]

Soln:

Class P: Own House = "yes"

Class N: Own House = "rented"

Total number of records = 12

Count the number of records with "yes" class and "rented" class.

Number of records with "yes" class = 7

Number of records with "Rented" class = 5

Information gain = $I(P, n)$

$$= \frac{P}{P+n} \log_2 \frac{P}{P+n} - \frac{n}{P+n} \log_2 \frac{n}{P+n}$$

$$I(P, n) = I(7, 5) = (7/12) \log_2 (7/12) - (5/12) \log_2 (5/12)$$

$$= 0.979$$

07-06-2021

STUDENT SIGNATURE: Amey

Step 1: Compute the entropy for income (very high, high, medium, low)

For income = very high

p_i = with "yes" class = 2 and

n_i = with "no" class = 0

$I(p_i, n_i) = I(2,0) = 0$

Income	p_i	n_i	$I(p_i, n_i)$
very high	2	0	0
High	4	0	0
Medium	1	2	0.918
Low	0	3	0

Entropy:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$E(\text{income}) = (2/12) * I(2,0) + 4/12 * I(4,0) + 3/12 * I(0,3)$

$= 0.229.$

07-06-2021 STUDENT SIGNATURE: Amey

Hence

Gain (S, income) = I(P, n) − E(Income)

= 0.979 − 0.229

= 0.75

Step 2: Compute the entropy for age:
(Young, Medium, Old)

Age	Pi	Ni	I(Pi, ni)
Young	3	1	0.811
Medium	3	2	0.971
Old	1	2	0.918

E(Age) = 4/12 * I(3,1) + 5/12 * I(3,2) + 3/12 * I(1,2)

= 0.904

Hence

Gain (S, age) = I(P, n) − E(age)

= 0.979 − 0.904

= 0.075

07-06-2021 STUDENT SIGNATURE: Amey

Income has highest gain

∴ It is decision attribute of root node.

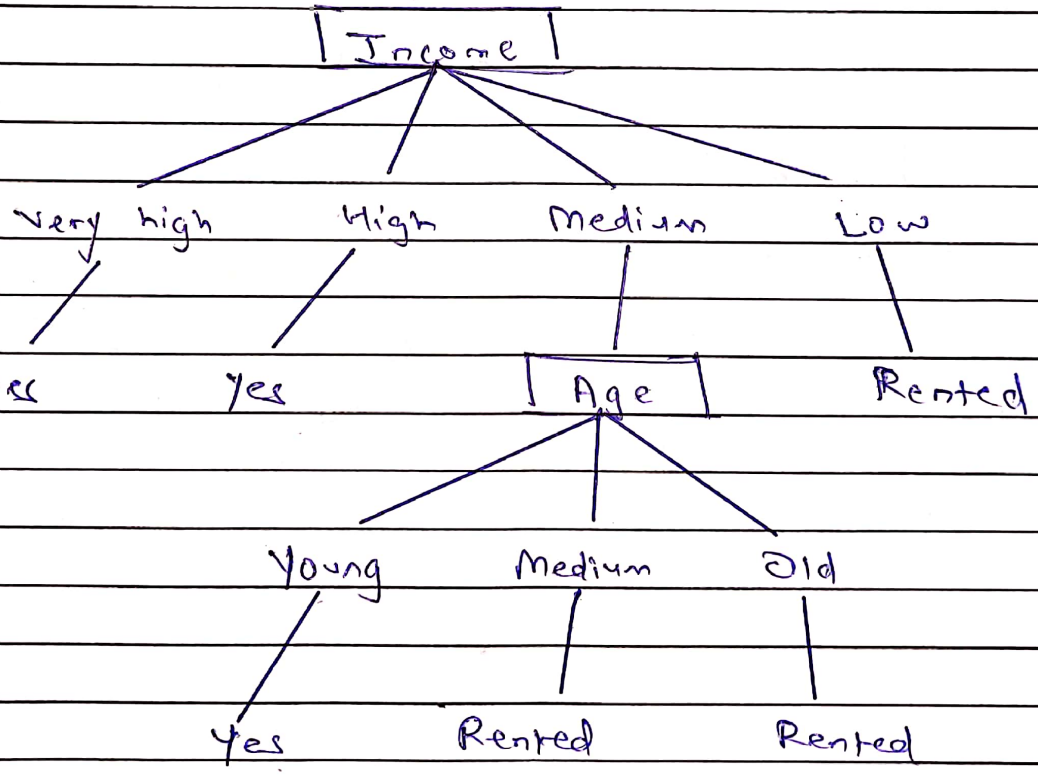
Step 3:

Consider income = "very high" and count the number of tuples

$S_{very\ high} = 2$

So both labels = "yes"

∴ Final Decision Tree



Decision Tree

07-06-2021 STUDENT SIGNATURE: Amey