

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

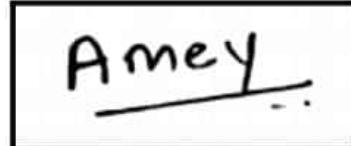
EXAM : SEMESTER VI

SUBJECT : DATA WAREHOUSING & MINING

DATE : 07-06-2021

DAY : MONDAY

STUDENT SIGNATURE:

A handwritten signature in black ink, reading "Amey", enclosed in a rectangular border.

Q3

A]

Sol:

(a) Star Schema

Product Key	Product	Store	Store Key
SKU Number	4,000 Product	300 stores	store name
Product Description	(Only 4000 sell)		store ID
Brand Name	in each store		Address
Product sub category	daily)		City
Product Category			State
Department			Zip
Package size			District
Package type			Manager
Weight			Floor Plan
Unit of measure			Services Type
Units per case.			
Shelf level			
Shelf width			
Shelf depth			
			Promotion key
			Promotion Name
			Promotion Type
Time key	2 Billion fact		Display Type
Date	table rows		Coupon Type
Day of week			Media Type
Week Number	A solid item		Promotion Cost
Month	in only one		Start Date
Month Number	5 years or		End Date
Quarter	promotion		Responsible Manager.
Year	1825 days	per store,	
Holiday Flag	TIME	per day.	
		PROMOTION	

NAME: AMEY THAKUR

BRANCH: COMPUTER

SEAT NO.: 61021145

SUBJECT: DWM

EXAM: SEMESTER VI

PAGE NO.: 2 / 9

(b) Time period = 5 years \times 265 days = 1825

There are 300 stores,

Each store daily sale = 4000

Promotion = 1

Maximum numbers of fact table records
= $1825 \times 300 \times 4000 \times 1$
= 2 billion.

NAME: AMEY THAKUR

BRANCH: COMPUTER

SEAT NO.: 61021145

SUBJECT: DWM

EXAM: SEMESTER VI

PAGE NO.: 3 / 9

Q 2

B]

SOL:

Adjacent matrix

$$d(i,j) = \sqrt{x_i - x_j)^2 + (y_i - y_j)^2}$$

$$d(A,B) = \sqrt{(2-1)^2 + (2-2)^2} = 1$$

Adjacent Matrix

j	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

1st pairing (A, B)

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

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

$$\text{Dist}((C, E), B) = \max(2.24, 2.12) \\ = 2.24$$

$$\text{Dist}((C, E), C) = \max(2, 1.58) \\ = 2$$

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

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

$$\text{Dist}((A, B), (C, E)) = \max(1.41, 1.58, 2.24, 2.12) \\ = \cancel{\max(1.41, 1.58)} \\ = 2.24$$

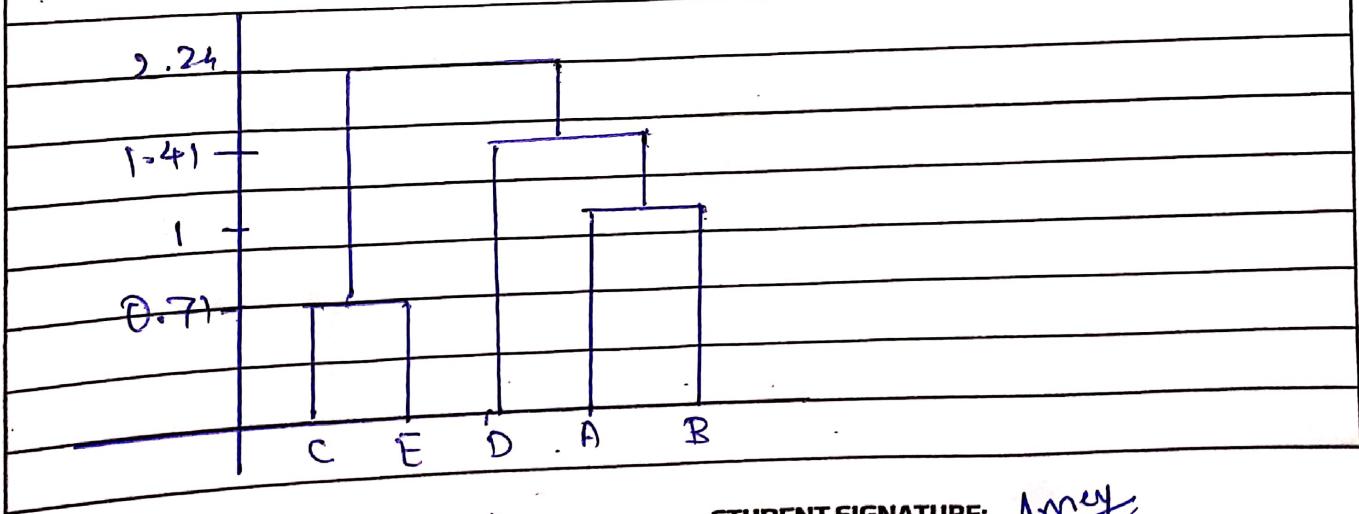
3rd pairing

Cluster (A, B) + D can be merged together

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

$$\text{Dist}((A, B, D), (C, E)) = \max(1.41, 1.58, 2.24, 2.12, 2, 1.41) \\ = 2.24$$

Dendrogram



Q3

G]

Sol:

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 = $I(P, n)$

$$= P \frac{\log_2 \frac{P}{P+n} - 0}{P+n} \Rightarrow \log_2 \frac{n}{P+n}$$

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

$$= \underline{0.979}$$

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 :

$$E(A) = \sum_{i=1}^v P_i + n_i \cdot I(P_i, n_i)$$

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

$$= 0.229.$$

Hence

$$\begin{aligned}\text{Gain } (S, \text{ income}) &= I(p, n) - E(\text{income}) \\ &= 0.979 - 0.229 \\ &= \underline{\underline{0.75}}\end{aligned}$$

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

Age	P _i	n _i	I(P _i , n _i)
Young	3	1	0.811
Medium	3	2	0.971
Old	1	2	0.948

$$\begin{aligned}E(\text{Age}) &= 4/12^{\underline{\underline{I(3,1)}}} + 5/12^{\underline{\underline{I(3,2)}}} + 3/12^{\underline{\underline{I(1,2)}}} \\ &= \underline{\underline{0.904}}\end{aligned}$$

Hence

$$\begin{aligned}\text{Gain } (S, \text{ age}) &= I(p, n) - E(\text{age}) \\ &= 0.979 - 0.904 \\ &= \underline{\underline{0.075}}\end{aligned}$$

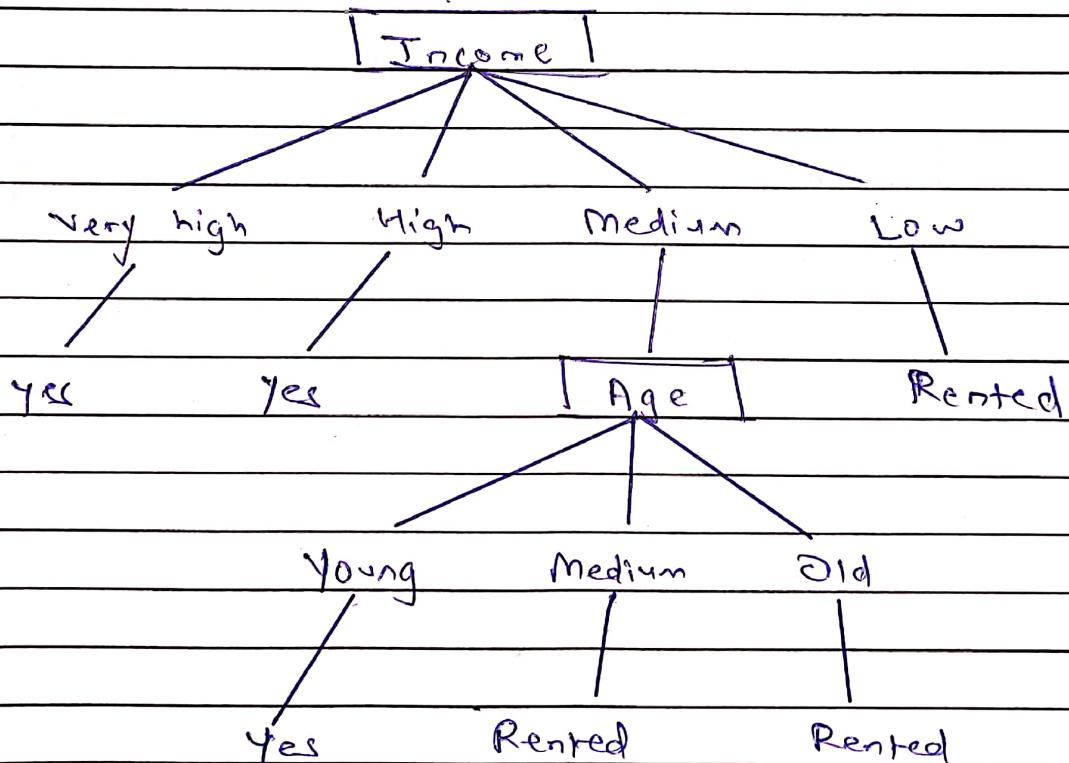
Income has highest gain
 \therefore It is decision attribute of root node.

Step 3:

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

$$\begin{aligned} S_{\text{very high}} &= 2 \\ \text{So both labels} &= \text{"yes"} \end{aligned}$$

\therefore Final Decision Tree



Decision Tree