Name of the Subject: DATA ANALYSIS & INFO. EXT. Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q2 d

Expected Info

$$I(S_1, S_2) = -\sum_{i=1}^{i} \frac{S_i}{S} \log\left(\frac{S_i}{S}\right)$$

$$= -\frac{S_1}{S} \log\left(\frac{S_1}{S}\right) - \frac{S_2}{2} \log\left(\frac{S_2}{S}\right)$$

$$= -\frac{120}{250} \log\left(\frac{120}{250}\right) - \frac{130}{250} \log\left(\frac{120}{250}\right)$$

$$= 0.99088$$

- To check whether "major" is a relevant attribute or not.

Here, V = 3 (Sinc, Engineering, Busines)

Expected Info. for Science

$S_{11} = 84$ (For grad.)
$S_{21} = 42$ (For Undergrad.)

$$I(S_{11}, S_{21}) = -\frac{84}{126} \log\left(\frac{84}{126}\right) - \frac{42}{126} \log\left(\frac{42}{126}\right)$$

$$= 0.9183$$

Section – II Date: 25/11/2021 Signature of Student: Page No: 01

For Engg.

$S_{12} = 36$ (grad.)

$S_{22} = 46$ (Undergrad.)

$$I(S_{12}, S_{22}) = -\frac{36}{82}\log\left(\frac{36}{82}\right) - \frac{46}{82}\log\left(\frac{46}{82}\right)$$

$$= 0.9892.$$

For Bussines

$S_{13} = 0$ (grad.)

$S_{23} = 42$ (Undergrad.)

$$I(S_{13}, S_{23}) = \frac{-42}{42}\log(1) = 0$$

$$E(\text{major}) = \frac{126}{250} I(S_{11}, S_{21}) + \frac{82}{250} I(S_{12}, S_{22}) + \frac{42}{250} I(S_{13}, S_{23})$$

$$\text{Gain(major)} = 0.9988 - 0.7873$$

$$= \boxed{0.2115}$$

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q 2 e

Class

C_1 = buys_Computer = 'Yes'

C_2 = buys_Computer = 'No'

Data to be classified

X = (age <= 30, Income = medium, Student = yes, Credit_rating = Fair.

$P(C_1) = P(\text{buys\_Comp} = \text{'Yes'}) = \frac{9}{14} = 0.643.$

$P(\text{buys\_Comp} = \text{'No'}) = \frac{5}{14} = 0.357$

— Compute $P(X/C_i)$ for each class

$P(\text{age} = \text{'<=30'} \mid \text{buys\_Comp} = \text{'Yes'}) = \frac{2}{9} = 0.22.$

$P(\text{age} = \text{'<=30'} \mid \text{buys\_Comp} = \text{'No'}) = \frac{3}{5} = 0.6.$

$P(\text{Income} = \text{'medium'} \mid \text{buys\_Comp} = \text{'Yes'}) = \frac{6}{9} = 0.667$

$P(\text{Incom} = \text{'medium'} \mid \text{buys\_Comp} = \text{'No'}) = \frac{1}{5} = 0.2.$

$P(\text{Credit\_rating} = \text{"fair'} \mid \text{buys\_Comp} = \text{'Yes'}) = \frac{6}{9} = 0.667$

$P(\text{Credit\_rating} = \text{'fair'} \mid \text{buys\_Comp} = \text{'No'}) = \frac{2}{5} = 0.4.$

Section – II Date: 25/11/2021 Signature of Student: Page No: 03.

X = (age <=30, income = medium, Student=yes, Credit_rate = fair).

$P(X|C_i) = P(X|buys\_comp = 'Yes') = 0.222 \times 0.44 \times 0.667 \times 0.667$

$= 0.044$

$P(X|buys\_Comp = 'No') = 0.6 \times 0.4 \times 0.2 \times 0.4$

$= 0.019$

$P(X|C_i) \times P(C_i) = P(X|buys\_comp = Yes) \times P(buys\_comp = Yes)$

$= 0.028$

$P(X|buys\_Comp = 'No') * P(buys\_Comp = 'No') = 0.007.$

→ Therefore, X belongs to class ("Buys_Computer = 'Yes'")

Q2 C.

T_{ID}	Item.
T_{100}	K, A, D, B.
T_{200}	D, A, C, E, B
T_{300}	C, A, B, E
T_{400}	B, A, D.

↓

Scan D.

↓

PTO.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

C_1

Item set	Sup.
{A}	4
{B}	4
{C}	2
{D}	3
{E}	2
{K}	1.

As, Min. Sup = 2 ↓

L_1

Item set	Sup
{A}	4
{B}	4
{C}	2
{D}	3
{E}	2.

↓

C_2.

Item set
{A, B}
{A, C}
{A, D}
{A, E}
{B, C}
{B, D}
{B, E}
{C, D}
{C, E}
{D, E}.

→ turn over.

Section – II Date: 25/11/2021 Signature of Student: Page No: 05

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Candidate Set C_2	Itemset	Sup.
	{A, B}	4
	{A, C}	2
	{A, D}	3
	{A, E}	2
	{B, C}	2
	{B, D}	3
	{B, E}	2
	{C, D}	1
	{C, E}	2
	{D, E}.	1

↓

L_2	Itemset	Sup
	{A, B}	4
	{A, C}	2
	{A, D}	3
	{A, E}	2
	{B, C}	2
	{B, D}	3
	{B, E}	2
	{C, E}.	2

↓

Q3.

Itemset	Sup.
{A, B, D}	3
{A, B, C}	2
{A, B, E}	2
{A, C, D}	1
{A, C, E}	2
{B, C, D}	1
{B, C, E}	1
{B, D, E}.	2

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

C_3

Itemset	Sup
{A, B, C}	2
{A, B, D}	3
{A, B, E}	2
{A, C, D}	1
{A C E}	2
{A D E}	1
{B C D}	1
{B C E}	2
{B D E}	1

↓ Scan D

Item Sup L_3.

{A B D}	3
{A C E}	2
{A B E}	2
{B C E}	2.

Available Association Rules are

$A \cap B \Rightarrow D = 3/4 = 75\%$.
$A \cap D \Rightarrow B = 3/3 = 100\%$
$B \cap D \Rightarrow A = 3/3 = 100\%$
$D \Rightarrow A \cap B = 3/3 = 100\%$.
$A \cap D \Rightarrow B$. $= 3/4 = 75\%$
$B \cap D \Rightarrow A$. $= 3/4 = 75\%$.

So, $A \cap D \Rightarrow B$
$B \cap D \Rightarrow A$
$A \cap B \Rightarrow D$
} are Strong Association.

Section – II Date: 25/11/2021 Signature of Student: Page No: 07

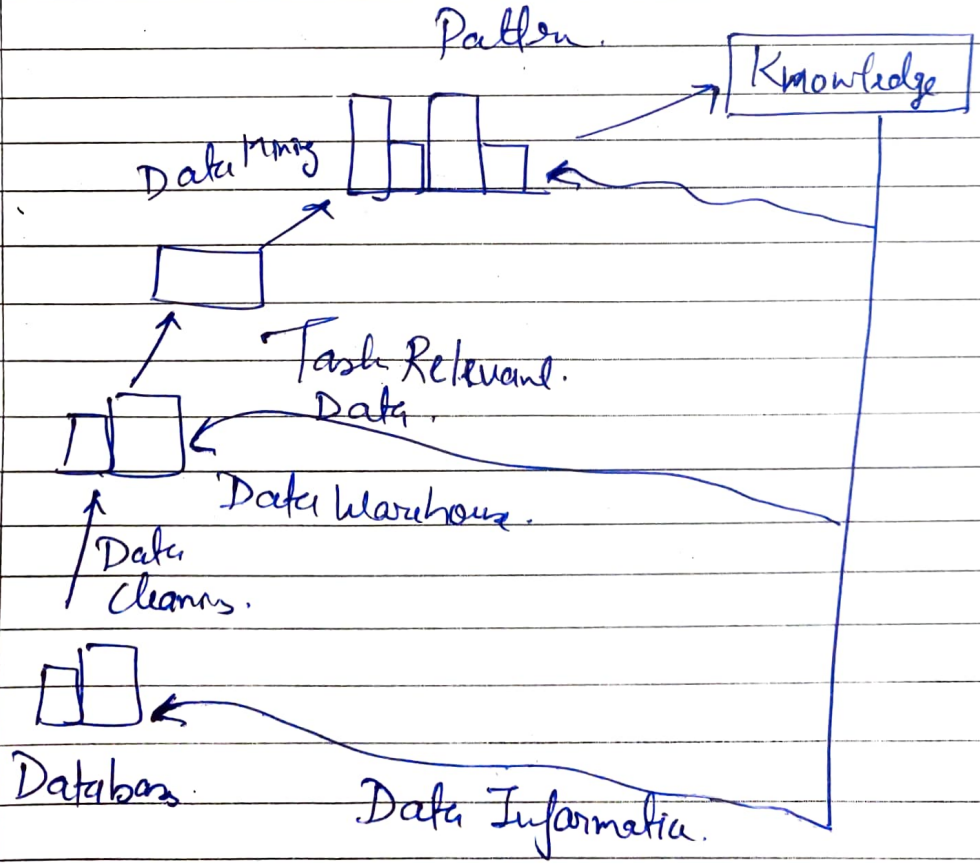
Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q3 a.

KDD - Knowledge Discovery in Database.

- Volume of Information is increasing everyday that we can handle business transaction & different data.
- So, we need a System that will be capable of extracting essential of information available & that can generated report & making decision.
- It refers to the nontrivial extraction of implicit provided & potentially useful information from data stored in database.



Section – II Date: 25/11/2021 Signature of Student: Page No: 06.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

① Data Cleaning :- Removal of Noisy & irrelevant data.

② - Cleaning with Data discrepancy detection & Data Transformation

② Data Integration :- Hetrogenous data form multiple Source combined in Comman Source.

- Data Integration using ETL

③ Data Selection - The proces where data relevant to analysis decided & retrived from data collection.

③ - It uses Clustering

④ Data Transformation - The proces of transforming data into appropriate form required by mining.

⑤ Data Mining - Techniques that are applied to extract patterns.

⑥ Pattern Evaluation :- Identifying strictly increasing pattern representing knowledge based.

⑦ Knowledge Representation :- Technique which utilizes visualization tool to represent data mining result.

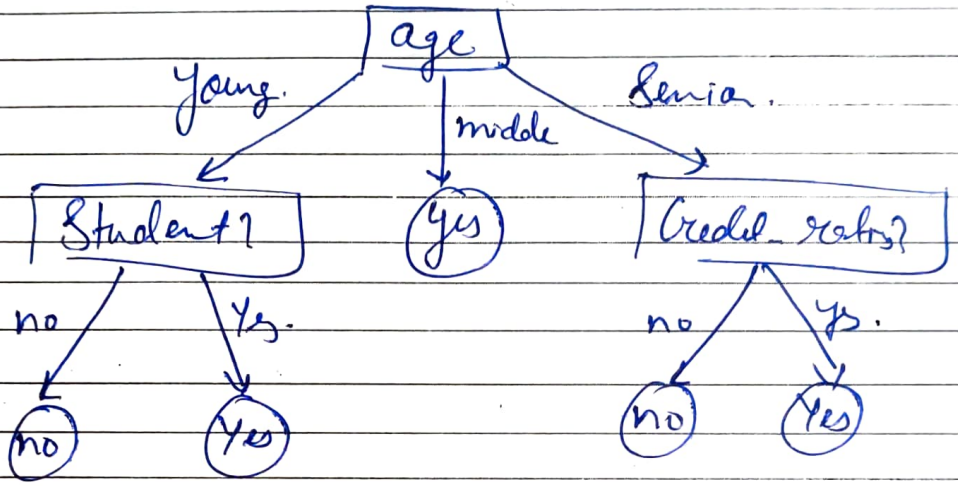
Section – II Date: 25/11/2021 Signature of Student: Page No: 09.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q3 b

A Decision Tree is a structure that include a root node, branches & leaf nodes. Each internal node denots a test on an attributs, each branch denots the outcom of a test, & each leaf node holds a clas label.



→ Benefits of decision Tree

- It does not require any domain knowledge
- It is easy to comprehend
- The learning & classification steps of a decision tree are simple & fast.

Section – II Date: 25/11/2021 Signature of Student: Page No: 10

— Working.

Create a node N;

if tuples in D are all of the same class, C then
return N as leaf node labeled with C;

if attribute_list is empty then
return N as leaf node majority class in D;

apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;

if splitting_attribute is discrete-valued &
multiway splits allowed

attribute_list = splitting attribute;
for each outcome j of splitting criterion

let D_j be = set of data tuples.

If D_j is empty then
attach a leaf
Class in D to node N;
else.
attach the node returned by generate.
decision tree
end if.
return N;

Q3 C

(i) Dimension & Fact Table.

- date_key — Calenda_date.
  day_no.
  day_overall
  day_of_week.
  mount_no_overall
  Year
  Time.
- Customa_key — Cust_full_name
  Cust_gender.
  Cust_title.
  Cust_birthdate.
  Cust_Shipto_name.
  Cust_postalcode
  Cust_numbe
  Cust_addl
  Cust_payment
  Cust. Credit Status.
- product_key — prod_numbe
  prod_Size
  prod_brand.
  prad_descriptn.
  prod_manufacture.
- Ship_mode — Ship_mode_name
  Carria_name.
  Ship_Ontact_numbe
  Ship_Contact_name

date_key INT

Calendar date DATE
day_no INT
day_week INT.
day_week TEXT
month_overall
year TEXT.

Product_key INT.
prodct_no TXT
product_Sig TXT
prodt_brand TXT
prod_des TXT
prod_nam TXT

Fact Table
date_key
Custom_key.
product_key.
Ship_mode.
Total amount
Total quantity
Total Discount.

Custom_key INT.
Cust_full_nam TXT
" _gender TXT
" _title TXT
" _ShipTo_Nam TXT
Cust_postalcode TXT
Cust_no INT
Cust_add TXT
Cust_paym TXT
Cust_Credit Sta TXT

Ship_mode INT
Ship_mode_no
Carrier_nam
Ship_contact_no
Ship_Contact_nam

Ship.

Product

date.