INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal - 500 043, Hyderabad, Telangana

Examinations Control Office

Examination	B TECH VI SEMESTER END EXAMINATIONS REGULAR JUNE 2025 REG UG20		
Month & Year	1-Jun	**Date**	20/06/2025
Course Name	DATA MINING AND KNOWLEDGE DICOVERY		
Course Code	ACIC01	**E-Code**	7918

Instructions to Evaluators

- Evaluators should spend at least 3-5 minutes on one answer booklet during the evaluation.
- Evaluators should cross check that marks are allotted for all the attempted questions.
- The marks should be assigned fairly according to the mark distribution specified in the scheme of evaluation.
- For questions that were attempted incorrectly, evaluators are required to award zero marks.
- The evaluator must give a proper justification in case of any mistakes identified in the marks provided.

START WRITING FROM HERE

Q.No.

1.a Types of Attributes in a dataset:

A dataset is a data source which consists of information / data about the specific class. So let us see the type of attributes found in a dataset.

(i) Nominal Attributes :-

These nominal attribute have fixed values, like a enum

examples:-

1. User Roles = { user, admin, manager etc.,}
2. WeatherStatus = { cloudy, sunny, windy, rainy, etc }
3. ReviewStatus = { pending, approved, denied }.

Here as we see in the above example there are a fixed values that can be in the Nominal Attributes. They are mainly useful in classifications.

(ii) Ordinal Attributes -

The ordinal attribute, are ordered values, They specify the order in the dataset.

Q.No

Examples :

1. MarksGrades = { A+, A, B+, B, C+, C etc}
   $A^+ > A > B+ > B > C^+ > C$ etc.,
2. TimeStamps = { 11:00PM, 12:00PM, 9:00AM) etc

Interval Attributes :-

Interval Attribute are order attribute with no "real zero", And each two consigutive values have a fixed interval between them

Examples :

1. Temperature = {0°C, 5°C, 1°C, 100°C etc.}
2. Time {0.00Am, 12:00PM, 6:50 Am etc}

Ratio Attributes :-

Ratio attribute are order and have real zeros. They could be numeric values or can be decimal values

Examples :-

1. Marks = {0, 10, 12, 5, 15, 21, 25 etc.}
2. Speed = {0.0, 0.01, 0.51, 0.65, 1.51 etc} (m/sec)

Q.No.

Binary Attributes:-

As the name specifies the Binary attributes consists only Binary values (Yes (or) No) or (True (or) False).

Examples:-

1. isEligible = { Yes or No }
2. isPassed = { True / False }

Now, let us see a dataset above attribute types

index	UserId	Role	Grade	Marks	isEligible
1	001	Student	A+	95	T
2	012	Teacher	A+	100	F
3	156	Student	C	61	F
4	119	Student	B+	73	T
5	017	Student	A	85	T
(Ordinal)	(Interval)	(~~ordin~~ (Nominal)	(ordinal)	(Ratio)	(Binary)

Q.No.	
1.b	Given, Dataset of patient records. AIM, Build a predictive Model to predict patient readmission rates. Process of Building a predictive Model :- 1. Data Cleaning :- Dataset need to be cleaned by identifying the missing value and managing the outlier in the data. This will help → Reduce Noise → Improve data quality → Minimized Data Redundancy. 2. Data Integration :- In this process, the data from various sources is integrated in to a single dataset.. This will help to decentralize the data, as the data is retrived from various spot.

Q.No.

3. Data Transformation :-

In this step, the data is transformed in to the useful format to make the process easy to build the model. Here techniques like Smoothing, Normalization and Text-vectorization are used.

4. Pattern Detection :-

By using the datamining algorithms the pattern that are intresting and relavent to the model are taken And irrelavent data is discarded.

5. Pattern Evaluation :-

Model evaluation metrics are used (F2-score etc.,) to check the accuracy and efficiency of the model.

6. Model Representation :-

The frontend tools like, dashboards, Graphs are used to visualize the model and Use it to predict the future events.

Q.No.

So, I would employ these following data mining techniques,

(i) data cleaning
(ii) data integration.
(iii) data transformation.
(iv) pattern recognition
(v) modal Evaluation
(vi) and modal representation,

to predict the patient readmission probabilit-ies based on the dataset provided by the health-care Organization.

Q.No.

2.b Given,

The value of the income attribute,

2000, 3000, 4000, 6000, 10,000.

Now, let us perform the Normalization methods specified.

(1) Do min-max - Normalization :-

In min-max normalization the data will be transforme between '0' - '1'

$$\text{minmaxNorm} = \frac{\text{item} - (\text{minimum\_value})}{(\text{max\_val}) - (\text{min\_value})}$$

(a) min_value = 2000

max_value = 10,000

- for 2000 $\Rightarrow \frac{2000 - 2000}{10,000 - 2000} = \frac{0}{8000}$

$\Rightarrow 0$

- for 3000 $\Rightarrow \frac{3000 - 2000}{10000 - 2000} = \frac{1000}{8000}$

$\Rightarrow 0.125$

Q.No.

- for 4000 → $\frac{4000 - 2000}{10000 - 2000} = \frac{2000}{8000}$

$\Rightarrow 0.25$

- for 6000 → $\frac{6000 - 2000}{10000 - 2000} = \frac{4000}{8000}$

$\Rightarrow 0.5$

- for 10,000 → $\frac{10,000 - 2000}{10,000 - 200} = \frac{8000}{8000}$

$\rightarrow 1.0$

∴ Min Max Normalized values

$\{0, 0.125, 0.25, 0.5, 1.0\}$

(ii) Z-Score Normalization -

$$Z\text{-Score} = \frac{\text{item} - \text{mean}}{S.D}$$

S.D → Standard deviation

$$\text{Mean} = \frac{\Sigma x_i}{n} =$$

$$= \frac{2000 + 3000 + 4000 + 6000 + 10,000}{5}$$

Q.No.

$$= \frac{25000}{5}$$

$\therefore \bar{x} = 5000$

$\therefore$ mean $= 5000$

$$SD = \sqrt{[\Sigma (x_i - \bar{x})]^2 / n}$$

$$= \sqrt{\left[(2000-5000)^2 + (3000-5000)^2 + (4000-5000)^2 + (6000-5000)^2 + (10000-5000)^2\right]/5}$$

$$\Rightarrow \sqrt{\frac{3000^2 + 2000^2 + 1000^2 + 1000^2 + 5000^2}{5}}$$

$$\Rightarrow \sqrt{\frac{(3+2+1+1+5)^2 \times (1000)^2}{5}}$$

$$= \frac{12000}{\sqrt{5}} = 536.6$$

SD = 536.6

(a) $z(2000) = \frac{|2000 - 5000|}{536.6} = \frac{3000}{536.6} = -5.6$

(b) $z(3000) = \left(\frac{3000 - 5000}{536.6}\right) = \frac{2000}{536.6} = -3.72$

(c) $z(4000) = \left(\frac{4000 - 5000}{536.6}\right) = \frac{1000}{536.6} = -1.86$

(d) $z(6000) = \left(\frac{6000 - 5000}{536.6}\right) = \frac{1000}{536.6} = 1.86$

(e) $z(10,000) = \left(\frac{10000 - 5000}{536.6}\right) = \frac{5000}{536.6} = 9.317$

Z-Score Normalized value

$\{-5.6, -3.72, -1.86, 1.86, 9.317\}$

Q.No.

(iii) Decimal Scaling :-

$$\text{Scaled} = \frac{\text{item}}{10^{i}}$$

where "i" is the (n-1)

$n = 5$, $i = 4$

$$\text{Scale}(2000) = \frac{2000}{10000} = 0.2$$

$$\text{Scale}(3000) = \frac{3000}{10000} = 0.3$$

$$\text{Scale}(4000) = \frac{4000}{10000} = 0.4$$

$$\text{Scale}(6000) = \frac{6000}{10000} = 0.6$$

$$\text{Scale}(10,000) = \frac{10000}{10000} = 1.0$$

Decimal Scaled Value

$\{0.2, 0.3, 0.4, 0.6, 1.0\}$

Q.No.

2.a

Data Cleaning :-

Data cleaning is the improtant step in data preprocessing. Data cleaning involves reducing noisy data in the data set while identifying the Outliers. Many raw data-sets have inconsistant data with many missing values and redundent information. Here in data cleaning such missing and inconsistant data is handled. Data cleaning will Improve the quality of the data and reduce the redundency.

Steps and Cleaning tools :-

* Perform description analysis using tools like pandas in python
* find the missing values in the dataset df.dropna().sum()
* fill the missing values, df.dropna → df.fillna()
  we can also specify the method

Q.No.

Handling Missing Values:-

Missing values cause the noise in the data, so they need to be handled.

(i) Replacing with Mean:-

We replace the missing value with mean of the total values in the ~~ad~~ column.

(ii) Replace with a Constant:-

We will calculate a constant value and put in the place of the missing value

(iii) Remove (ignore) the missing:-

If there is a value missing in the row (tuple) we just ignore the tuple

And there are many othe methods to handle the missing value in the data set like medians and "Predicting Models to predict the missing Values".

Q.No.

3.a. Online Analytic Processing (OLAP):-

Online Analytic processing servers are used as bridge between the data-warehouses and the frontend tools. The OLAP servers, will perform the analytical operations on the data stored in the Storage tire of the architecture where Data Marts, Metadata repositories and data-warehouse are located. And the analyzed information is ready to be used by the frontend tools to visualize the Information.

There are three Online Analytic processing server architectures, they are

(i) Relational Online Analytic Processing (ROLAP):-

The ROLAP servers are built upon the Relational database System, which consists of the Normalized data. like Tables, attributes, and relations. And high level complex SQL (Structured query language) quires are used to perform the operations, on the ROLAP Server.

Q.No.

~~Rolap~~ ROLAP Advantages :-

→ Easy to implement

→ Suitable for small data warehouses

ROLAP DisAdvantages :-

→ Inefficient for large data

→ Complex quires of SQL.

(ii) MultiDimentional Online Analytic Processing (MOLAP) :-

The MOLAP servers are flexible and used for operations on multi-dimentional data querying. Like data stored in the form of Cubes (data Cubes) and Data mining Query Language (DMQL) are used to communicate with the Gateways.

MOLAP Advantages :-

→ ~~Hard to use~~ by easy to use

→ Suitable for more data.

MOLAP DisAdvantages :-

→ Inefficient for small data

→ And need good understanding of DMQL.

Q.No.

(iii) Hybrid Online Analytic Processing :- (HOLAP) :-

HOLAP uses both the feature of ROLAP and MOLAP, ie., the data can be in the Normalized formal like tables and relation, or data could be multi-dimention-al and stored in the data-cubes.

Advantage of HOLAP :-

→ Effective and efficient for both small and heavy data, and provide accurate results

DisAdvantage of HOLAP :-

→ Complex to implement and use.

→ Need experience to maintain the server

* The OLAP server can perform the following operation

→ Roll up (Top to Bottom to up)

→ Drill down (Top to Bottom)

→ Slice and Dice (Single dimation and sub dim-)

→ Pivote (Rotate for a new perspective).

Q.No.

3.b Given,

The data warehouse consists of three dimensions.

- Time
- Doctor
- Patient.

and, two Measures

+ Count
+ charge

(i) Classes of schema that are popularly used for modelling data warehouse :-

① Star Schema :-

In star schema there is one fact table and multiple dimention tables refering to the fact-Table.,

fact-Table :- The table that stores the mesurable values

dimention-Table :- The table that stores the descriptive data

Q.No.

(b.) Snow flake Schema :-

Snow flake schema is an extention of the star schema, that have multiple sub dimention connection to the existing dimention with only on single fact table.

(c.) Galaxy Schema :-

The galaxy or fact constulation schema is the complex merge of multiple snowflake - star schemas. In the Galaxy schem there are two - or - more fact tables that share the same dimention Tables.

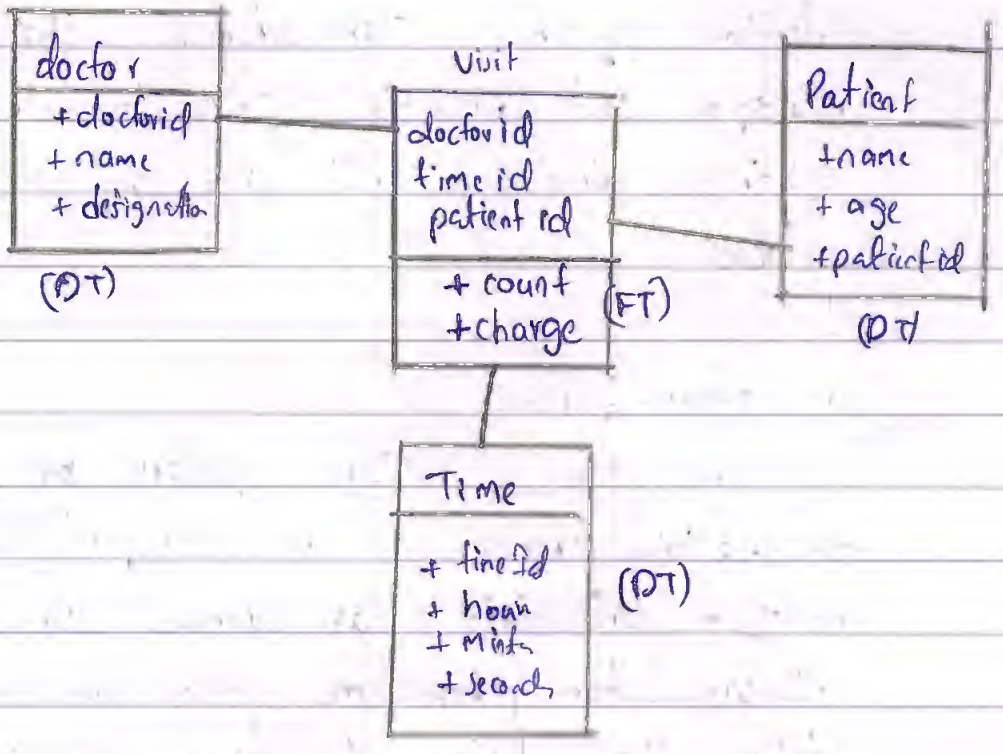
(ii) (a) ~~Star schema~~ :-

Paatient

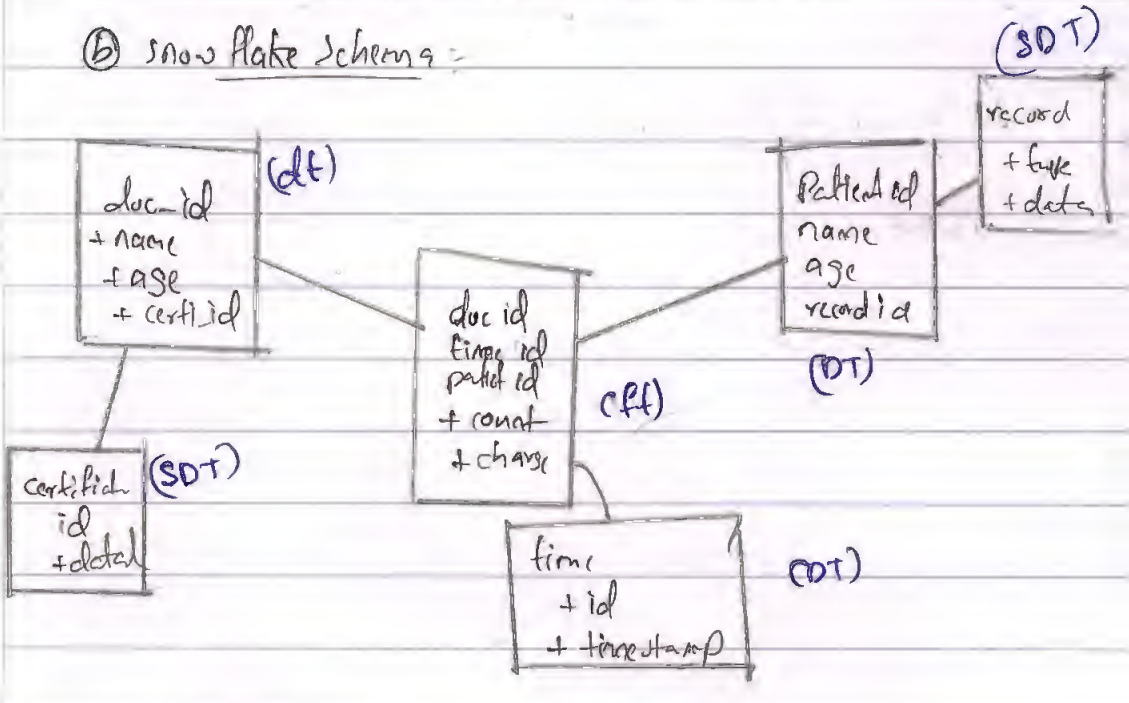
Hospibel

Q.No.

(ii) @ Star - schema :-

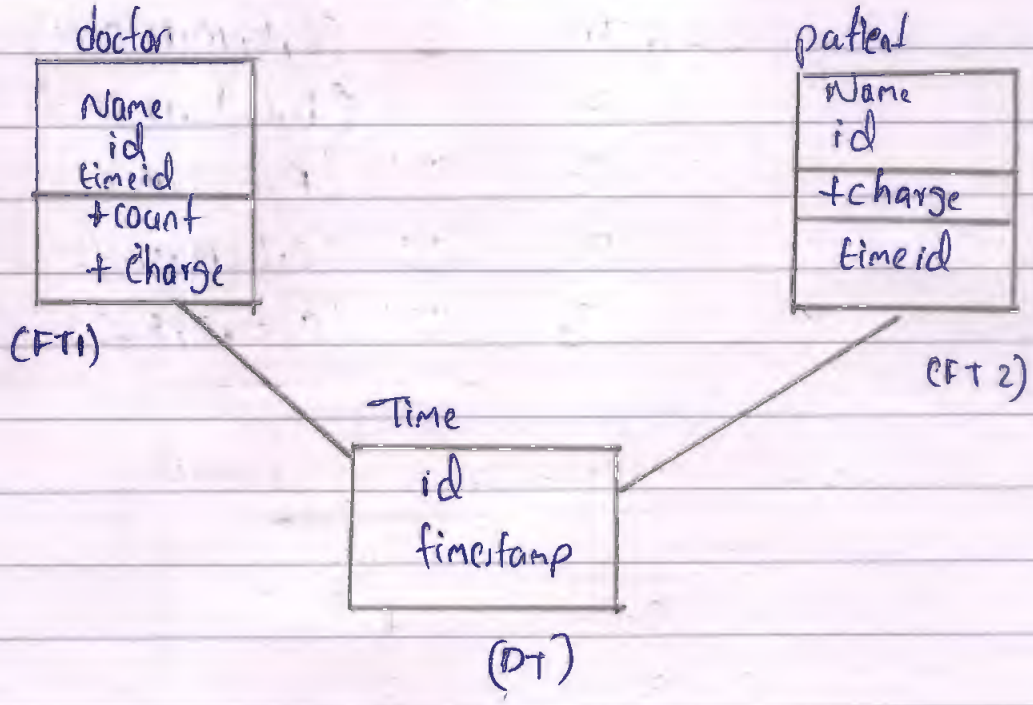


(b) snow flake schema :-



Q.No.

(c) Galaxy or Fact Constellation Schema:-



Q.No.

5.b Construction of fp-tree for given transaction.

Given,

Minimum Support Threshold (MST) = 3.

Transcation id	items
T_1	{E, K, M, N, O, Y}
T_2	{D, E, K, N, O, Y}
T_3	{A, E, K, M}
T_4	{C, K, M, U, Y}
T_5	{C, E, I, K, O}

Step 1 :- write frequency of each item

item	count
A (x)	1
C (x)	2
D (x)	1
E	4
I (x)	01
K	5
M	3
N (x)	2
O	3
U (x)	1
Y	3

remove item with count less than MST = 3

Q.No.

step 2:- write priority

item	count	priority
E	4	2
K	5	1
M	3	3
O	3	4
Y	3	5

[K > E > M > O > Y]

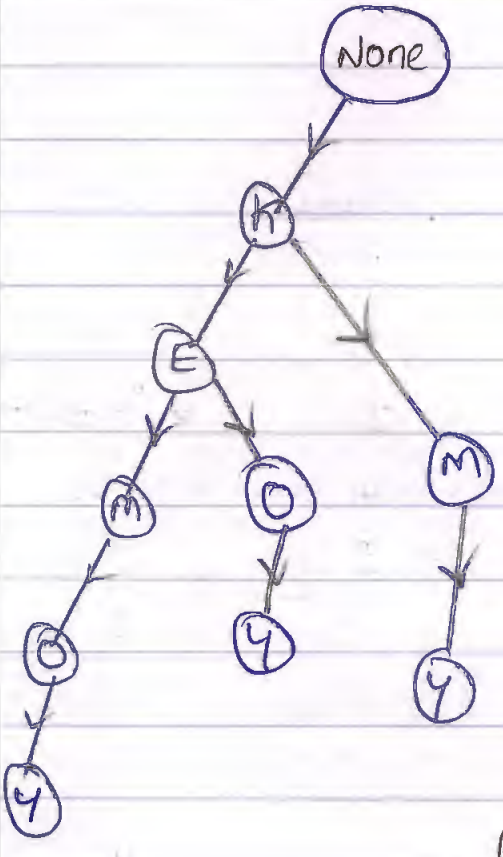
step 3:- Write Transaction in order of priority of item

id	items in Order
T_1	K, E, M, O, Y
T_2	K, E, O, Y
T_3	K, E, M
T_4	K, M, Y
T_5	K, E, O

step 4:-

Now, our next step is to draw the Fp-tree.

Q.No.



fp-tree

Frequent pattern :-

(i) None → K → M → Y

(ii) None → K → E → O → Y

(iii) None → K → E → M → O → Y.

Q.No.

7.b. Given data point

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$,

$B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$

$C_1(1,2)$, $C_2(4,9)$, $C_3(4,9)$.

and initial centroid A_1, B_1, C_1

applying K-mean algorithm

Data points	C1(2,10)	(5,8)	(1,2)	cluster
$A_1(2,10)$	0	3.60	5.1	1
$A_2(2,5)$	5	4.25	3.1	3
$A_3(8,4)$	8.6	5	6.5	2
$B_1(5,8)$	4.7	0	4.81	2
$B_2(7,5)$	7.5	3.6	6.2	2
$B_3(6,4)$	4.13	4.12	5.38	2
$C_1(1,2)$	1.8	7.21	0	3
$C_2(4,9)$	3.8	1.41	7.61	2
~~$C_3(4,9)$~~				

New mid point (centroid)

So, (2,10), (6,6.5), (3,7)

Q.No.

datapoint	(2,10),	(6,6.5)	(3,7)	New	
(2,10)	0	4.1	2.6	1	
(2,5)	5	4.2	1.8	3	3
(8,5)	8.6	2.8	5.8	2	2
(5,8)	4.7	1.6	2.2	2	2
(7,5)	7.15	2.1	3,2	2	2
(6,4)	4.13	1.6	(3,2)	2	2
(1,2)	1.8	5.8	(2,1)	1	3
(4,9)	3.8	1.56	(1.1)	3	2

repeat the steps until New clust = old clste

we get,

cluste 1 = $\{A_1, C_1\}$
cluste 2 = $\{A_3, B_1, B_2, B_3\}$
cluste 3 = $\{A_2, C_2\}$

Q.No.

7.c Different type of data used in the cluster Analysis:-

Cluster Analysis is the process of diving the data into different clusters based on the Methods like, partioning, hirarchi, Density based, Grid and Model based. The data is unsupervised and the clusters are detected based on the above method. And Cluster Analysis can be done on any type of data types like

- Nominal data, used for simple clusting based on classified value
- Ordinal data, Order of the data without real Numerical value
- Binary data like yes or No
- Text data, we can perform cluster analysis on the text data also, too. Many application use to cluster data like Meta, Google
- Image, also can be clustred into groups based on their feature

Q.No.

Typical Requirement of of Clustering :-

* Data need to be preprocess and cleaned with out noise and missing.

* Data should be formated like text to vectors and Images to Matrix for easy clustering.

* Data should be unlabled to perform clustering. if data is labled then classification is best option.

Q.No.

Q.No.	

Q.No.

Q.No.	

Q.No.	

Q.No.	

Q.No.

$Z\text{-score} = \frac{x_i - \mu}{SD}$

$SD = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$

$\left((x_1-y_1)^p + (x_2-y_2)^p\right)^{1/p}$