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	7871

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.

7b) Ans: To perform a k-means technique to from clusters for given data.

We have to used Euclidean distance as the basics to ~~pt~~ perform clusters.

Euclidean distance formulae:

$$x_1 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Given: $A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$.

Initial centers are A_1, B_1, C_1.

Round-1

Points	Initial Centers.				New
	$A_1(2,10)$	$B_1(5,8)$	$C_1(1,2)$	Clusters	Cluster
$A_1(2,10)$	0	3.605	8.062	1	
$A_2(2,5)$	5.0	4.242	3.162	3	
$A_3(8,4)$	8.485	5.0	7.280	2	
$B_1(5,8)$	3.605	0	7.211	2	
$B_2(7,5)$	7.071	3.605	6.708	2	
$B_3(6,4)$	7.211	4.123	5.385	2	
$C_1(1,2)$	8.062	7.211	0	3	
$C_2(4,9)$	2.236	1.414	7.615	2	

Q.No.

i) Three cluster centers after first round

Cluster 1 elements = A_1

Cluster center = (2, 10)

Cluster 2 elements = A_3, B_1, B_2, B_3, C_2

Cluster center =

$$\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$$

$$= (6, 6)$$

Cluster 3 elements = A_2, C_1

Cluster center =

$$\left(\frac{2+1}{2}, \frac{5+9}{2}\right)$$

$$= (1.5, 7)$$

Cluster centers after the first round are (2, 10), (6, 6), (1.5, 7)

Q.No.

Round 2

Points	Cluster Centers			Old Cluster	New Cluster
	(2,10)	(6,6)	(1.5,7)		
$A_1(2,10)$	0	5.656	3.041	1	1
$A_2(2,5)$	5.0	4.123	2.061	3	3
$A_3(8,4)$	8.485	2.828	7.158	2	2
$B_1(5,8)$	3.605	2.236	3.640	2	2
$B_2(7,5)$	7.071	1.414	5.852	2	2
$B_3(6,4)$	7.211	2.0	5.408	2	2
$C_1(1,2)$	8.062	6.403	5.024	3	3
$C_2(4,9)$	2.236	3.603	3.201	2	1

Since Clusters of round 2 and round 1 don't match Redo the centers and check again.

Cluster 1 center $= \left(\frac{4+2}{2}, \frac{10+9}{2}\right) = (3, 9.5)$

Cluster 2 center $= \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5, 5.25)$

Cluster 3 center = unchanged = (1.5, 7)

Round 3

Points	Cluster Centers			Old	Nec
	(3, 9.5)	(6.5, 5.25)	(1.5, 7)	Cluster	Clust
A_1(2,10)	1.118	6.543	3.041	1	1
A_2(2,5)	4.609	4.506	2.061	3	3
A_3(8,4)	7.438	1.952	7.158	2	2
B_1(5,8)	2.5	3.132	3.640	2	1
B_2(7,5)	6.020	0.55	5.852	2	2
B_3(6,4)	6.264	1.346	5.408	2	2
C_1(1,2)	7.762	6.388	5.024	3	3
C_2(4,9)	1.118	4.506	3.201	1	1

New Cluster does not match old so repea

$$C1 \text{ Center} = \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right) = (3.67, 9)$$

$$C2 \text{ Center} = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7, 4.33)$$

$$C3 \text{ Center} = (1.5, 7)$$

Q.No.

Round 3

Points	Cluster Centers			Old Cluster	New Cluster
	(3.67, 9)	(7, 4.33)	(1.5, 7)		
$A_1(2,10)$	1.946	7.559	3.041	1	1
$A_2(2,5)$	4.334	5044	2.061	3	3
$A_3(8,4)$	6.614	1.053	7.158	2	2
$B_1(5,8)$	1.664	4.199	3.640	1	1
$B_2(7,5)$	5.204	0.67	5.852	2	2
$B_3(6,4)$	5.516	1.053	5.408	2	2
$C_1(1,2)$	7.491	6.436	5.024	3	3
$C_2(4,9)$	0.33	5.550	3.201	1	1

Since Both Old and New Cluster are same.

Final Clusters are:

Cluster 1 = $A_1(2,10)$, $B_1(5,8)$, $C_2(4,9)$

Cluster 2 = $A_3(8,4)$, $B_2(7,5)$, $B_3(6,4)$

Cluster 3 = $A_2(2,5)$, $C_1(1,2)$

Q.No.

7a) Ans- In data mining a cluster standords for a ~~collect~~ collection objects grouped together which have similar characteristi-cs.

Steps to form a clusters in data mining:

1) Step 1 is the first step where you have to arrange the data in an order.

2) Second step is to deside the process to form clusters. Clusters can be formed by many types. Some of the methods are K-means, k-Medoids. Data density, etc.

3) After selecting the process / method for clustering smooth the data. Through this step one can remove all the unwanted data present in the raw data.

Q.No.

4) To apply the process here for example k-means clustering.
In this clustering we cluster the data based on partationing the data in equi-distance way.
We find centers of each cluster and check if all the elements arranged in that cluster are closer to that cluster center or not.

For cluster analysis we use different types of data each time to analyze the cluster.
This all depends on the type of the cluster.
Some elements used are:

Partationing method: Cluster are formed based on distance. eg:- k-means

~~Grg~~ Grid method: We check at what grid the cluster forms.

Density Method: We check the range/width of the cluster.

Q.No.

Multi level /Highearchy method : We divide cluster
based on levels.

2b) Ans:-

Given data:-

2000 , 3000, 4000, 6000, 10000

~~Decimal~~

Mean for given data:

= 2000 + 3000 + 4000 + 6000 + 10000 /5

= 5000

SD = |5000 - 2000| = 3000, |5000 - 3000| = 2000
|5000 - 4000| = 1000, |5000 - 6000| = 1000, |10000 -
5000| = 5000.

Sum of SDs = 3000 + 2000 + 1000 + 1000 + 5000
= 12000

MAD = Sum of SD / Number of elements.
MAD = 12000/5
MAD = 2400

Q.No.

Z- Score normalizations

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

$$= \sqrt{[3000^2 + 2000^2 + 1000^2 + 1000^2 + 5000^2 /5]}$$

$$= \sqrt{12000^2 / 5}$$

$$= \sqrt{28800000}$$

$$x' = 5366.56$$

Min-Max normalizations

~~Min-Max~~ $\overset{*}{x} =$ Max-Min / mean

$x' = 10000-2000/5000$

$x' = 8000/5000$

$x' = 1.6$

Decimal ~~Sec~~ Scaling for income attribute is

$$\frac{1.6}{10} = 0.16.$$

Q.No.

2a) Ans:-

Data Cleaning:

In Data Mining Data Cleaning process plays an important role before organizing the data.

This process helps us in many many ways and makes the data robust.

Data Cleaning is the process where all the impurities present in the dataset ~~are~~ are removed.

Impurities in data like missing values, same values, etc are all solved in this process.

These impurities ~~get~~ cause the data set to mal function. To avoid this companies always clean the data before ~~store~~ storing the data into their servers.

If there are missing values in the dataset companies simply delete the row from the data set.

To Avoid repeation companies use unique attributes by which they can

Q.No.

prevent some data frome entering the dataset.
By doing this companies can avoid error while fetching data of a person.

If some data is found system would get confuised. So companies provide unique ids for each • user / consumer using their product to avoid overlapping of data.

In doing this way companies can save both cost and save space while storing data.

Q.No.

5b) Ans:-

To construct an FP-tree for the given data set.

Given:-

Minimum support count =3

Transaction IDA	Items
T1	{E, K, M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O, O}

Items	Count	Priority
A	1	
C	2	
D	1	
E –	4	2
I	1	
K –	5	1
M –	3	4
N	2	
O –	4	3
Y –	3	5

Since min support count=3

K, E, O, M, Y

Combinations:

KEO,M, Y

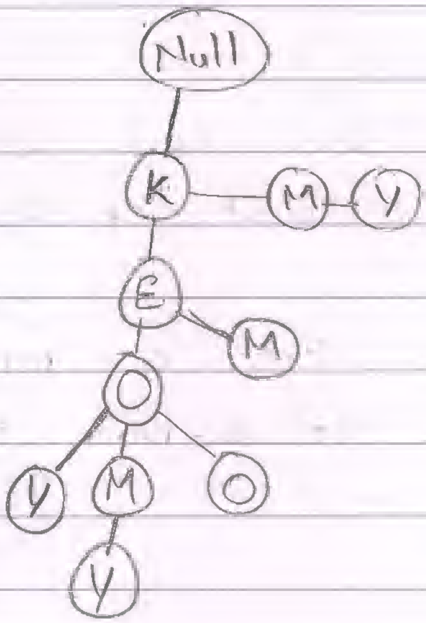
K, E, O, Y

K, E, M

K, M, Y

K, E, O, O

Tree :-



K - 1, 2, 3, 4, 5

E - 1, 2, 3, 4

O - 1, 2, 3

M - 1

Y - 1

Y - 1

M - 1

M - 1

Y - 1

O - 1

Checking:

K=5, E=4, O=3+1=4, M=1+1+1=3, Y=1+1+1=3

It matches to the count.

This is the final F-P tree.

Q.No.

50) Ans:

Decision tree is a tree where all the branches of the tree are present.

This is a tree generated by training tuples. Various tuples are trained to generate a decision tree. Steps to generate a decision tree are:

Step1: To arrange all the tuples in an order.

Step2: To break down each tuple and write the count of each element.

Step3: To choose either a top down or a bottom up method to construct the tree.

Q.No.

30) Ans:- OLAP is the Online Analytical Processing. This helps us to ~~toning~~ mine and analyze the data in a data warehouse easy. This is used for large sets of data. It is used to store data in data warehouse.

There are different types of OLAP :-

1) MOLAP:

This is used to analyze data individually by level by level.

This uses multi level approach to analyze data.

2) ROLAP:

This uses relations in data base to analyze data.

It does not care about levels it only cares the relation for the element.

3) HOLAP:

This is a hybrid.

This is best of both worlds it uses both ROLAP and MOLAP to analyze data.

Q.No.

MOLAP is used when the data is vertical and lpnear.

ROLAP is used in dbms and to stor data in servers and analyze them.

HOLAP is used when the data contains vertical levels and also relations.

Q.No.

1b>
Ans

I would employee the clustering ~~th~~ technique as there would be more and more number of patients in ~~a~~ health care organization.

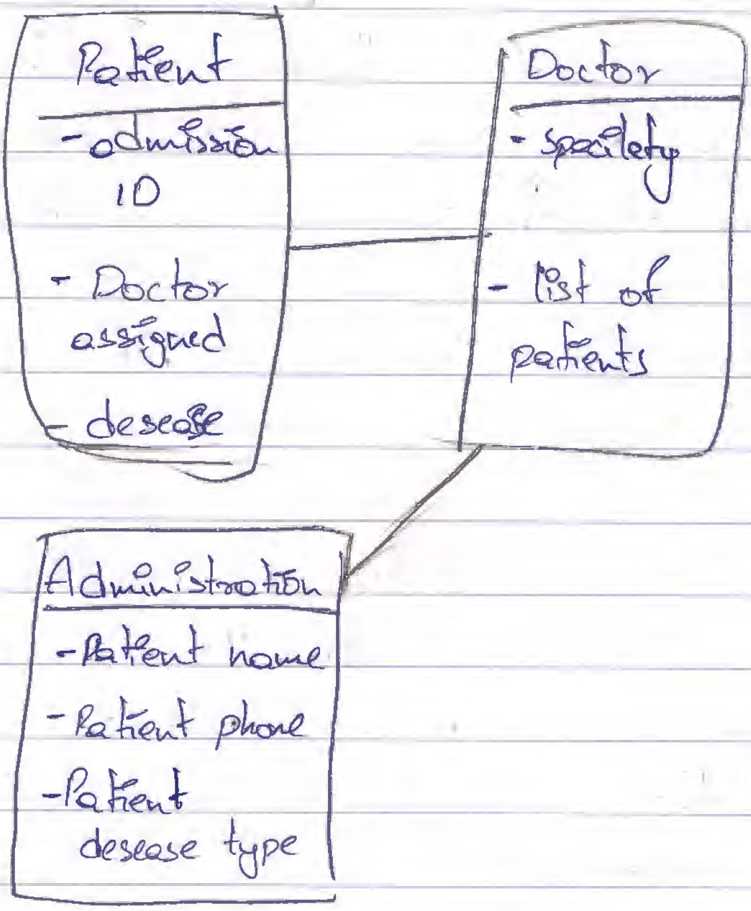
We can form clusters like cardio patients, ent patients, etc.

By this the allocation of doctors and staff would be easy. It would also be easy to study a patients case file on all the related patients who had some symptoms.

This would be easy for organizations to maintain all the required medicines for each patient.

A predictive Model is like a an ER Diagram where you would predict the regular responce of a patient.

Q.No.


Patient
-admission ID
- Doctor assigned
- desease
Doctor
- specialety
- list of patients
Administration
-Patient name
-Patient phone
-Patient desease type

Q.No.

3b) Ans:

Three schema popularly used for modelling a data in data warehouses are:-

1) Role down:-

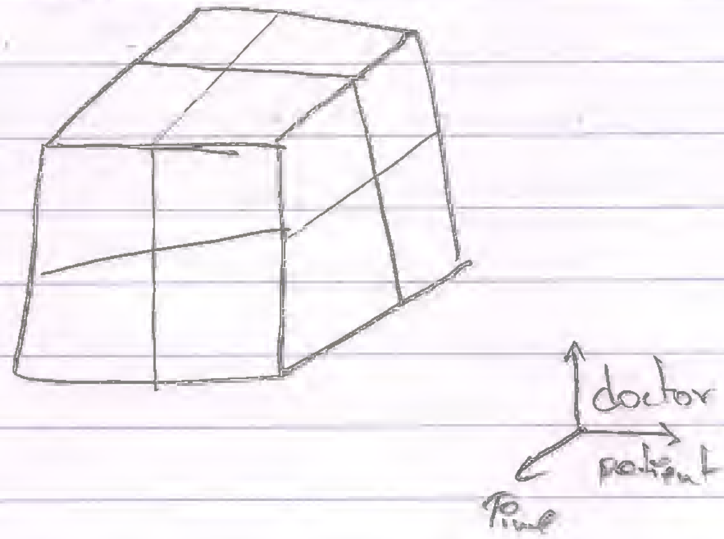
It ~~exden~~ extends the data set down words. It can add more colums and reduce the columns from bottom

2) Role Up:-

It can decrease the columns from sides

3) Dice :-

It forms equal rows on colums on each side.



Q.No.

10)
Ans:-

~~Hea~~

Many types of attributes can be found in a data set. Some of the attributes are:-

1) Primary attributes:
It is a unique ID used to identify an element.

2) Required attribute:
It is an attribute which is required to be filled to enter your da

3) Temperary attributes:
This is an attribute which holds a temperory ~~value~~ value until you decide the value of that element. to prevent any errors from occuring.

Q.No.	

Q.No.

Q.No.

I A R E

EDUCATION FOR LIBERATION

Q.No.	

Q.No.

Q.No.

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.

Q.No.	

Q.No.

ROUGH WORK

Content written here will not be considered for valuation