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**	7839

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) Attributes of the data are the properties of the data. These attributes define the different types of data in a dataset. A dataset consists of different types of attributes. These are, ordinal attributes - the attributes that give the ranking information so that we can rank the attributes based on the attributes. A dataset can be organised using the ordinal attribute. These attributes are used to classify the dataset based on the rankings so that we can perform different operations easily. Ex: Ranks of students in an exam, Grades of quality of items, these are the examples of ordinal Attributes. The normal attributes are the attributes where they only provide the information to only distinguish one dataset from an another. The attributes just provide the nominal data. The attributes are of no other use than to provide basic information. These attributes are regarded as the basic attributes since they are only used to identify the dataset. These are the examples - Roll numbers of the students studying in college, Names of the

Q.No.	
	people etc. Binary attributes are the attributes where the attributes of the dataset has either one of only 2 values - 1/0. These attributes declare if the attribute of a dataset are present or not. If the features are present, then its 1, if absent, then its 0. Therefore, binary attributes are only used to denote 2 values. Ex: Male/Female, If a student is present/absent etc. Numerical attributes are the attributes where only numerical data like integers are used to define the data. Ex: no. of chairs etc. Arbitary attributes are used to define a dataset using multiple data like for example to define a student dataset, we use, roll no - alphanumeric, marks - integer, name etc. These are the different types of attributes.
1(b)	If a healthcare organisation wants to predict patients readmission rates, we have analyse a dataset consisting of patient's medical records. Using the given dataset, we need to make use of it and

Q.No.

build a predictive model so that we can perform analytical processes. First, we need to gather the required information like the patient records. After gathering the required information, we need to clean the data. This is known as data cleaning. In this step, we will identify any outliers and try to clean the data. We need any redundandy data present to be eliminated and build the data free from errors. Then we move onto the preprocessing of data where we use different types of algorithms to pre-process the given dataset. Then we design the data as per our requirements so that we can perform data mining operations. Data mining is the process in which we used analytical models to extract any recognise patterns in the data that were previously unknown and we can gain meaningful insights from the data. After preforming the data mining operations on the dataset, we can gain valuable insights in the patient records and we can able to recognise and identify the patterns. Thus, relying on the

Q.No.	
	patterns, we can build a predictive model that can help the organisation to predict patient readmission rates. Lastly, we can infer on the data and make analytical process on the patient records. These are the data mining techniques that I would employ.
2(a)	Data cleaning is one the most important and the foremost process of data mining process. Data cleaning is the process in which we smooth the data, remove any outliers, identify and fix any gaps in the data and many processes among others. Data cleaning refers to handling the data and making it suitable for the processing of the data and perform any operations on the data. Therefore, it is useful for and one the most important steps in the data mining process. Let us now discuss each aspect of the data cleaning process now in detail. Data cleaning helps in identifying the outliers. Outliers are the data which exist outside of the regular area where

Q.No.

majority of the data occurs. Outliers are therefore, troublesome because they can skew the data and lead to less accurate results. Therefore, we need to identify the outliers and eliminate them. We can employ different techniques like normalisation etc to remove the outliers. We need to perform such techniques on the datasets to clean the data. After identifying the outliers, we need to clean them using techniques like normalisation etc. This is one part of the data cleaning process. After removing the outliers, we need to find if there are any gaps in the data. If there are any gaps present in the data, then we need to fill the gaps because then these gaps will rise the issue of inconsistent data and therefore maybe not suitable to perform data mining operations. Therefore, to prevent any such mishappenings, we need to identify if there are any gaps present in the data and after identifying such data, we need to fill the gaps in the data. We need to use different algorithms to make an educated guess of the missing data and

Q.No.

fill the gaps. We also need to smoothen the data so that it is suitable to perform data mining techniques. To handle the missing values, we can employ different techniques like using normalisation techniques like min-max where we take the maximum and minimum elements in a dataset and perform normalisation to check the integrity of the data. We also use mean, median and different scaling operations to handle missing values.

2(b) Given : 2000, 3000, 4000, 6000, 10000

range : [0.0, 1.0]

Min-Max normalisation : min : 2000

max : 10000

$$= \frac{(2000 + 10000)}{5} = \frac{8000}{5} = 1600$$

=> Normalisation : (2000 - 1600) + (3000 - 1600) + (4000 - 1600) + (6000 - 1600) + (10000 - 1600)

= 17000

Q.No.

$\Rightarrow \frac{17000}{25000} = 0.68$ is the min-max normalisation.

(sum of values)

2-score normalisation:

2-score $\Rightarrow$ mean of data $= \frac{2000+3000+4000+6000+10000}{5}$

$= 5000$

Normalisation $= \left(\frac{5000}{25000}\right) = \frac{1}{5}$

$= 0.2$ is z-score normalisation

Decimal scaling: total income $= 1000+2000+4000+6000+1000$

$= 25000$

decimal scaling $= \frac{25000}{10^5} = \frac{25000}{100000} = \frac{1}{4}$

$= 0.25$ is decimal scaling

3(a) OLAP stands for OnLine Analytic Processing. OLAP server is used to perform the analytical process on the data so that we can extract meaningful information from the data

Q.No.

and we recognise and extract the different patterns. There are three different types of OLAP servers based on the architecture and they are Multidemetional OLAP servers (MOLAP), Relational OLAP servers (ROLAP) and Hierarchical OLAP servers (HOLAP). All of these servers differs in the architecture of Online Analytical processing engine such that each of the servers performs differently based on their respective architectures. MOLAP, or also refered to as the Multidimensional Online Analytical processing, is the servers where the data exists in many different dimensions and not only in a single dimension. The MOLAP, the data is stored in the form of cubes where we perform different operations like slicing, dicing, extracting etc. The data in MOLAP is a structured data. In MOLAP, the dataset is stored in the form of cubes where the summary of the dataset is present above and if we want to get to the individual data, it present in the bottom

Q.No.

layer and we employ drill-down technique to get the data. It has many different dimensions based on the data. It is easier to implement than other OLAP architectures but it requires complex queries. It is suitable where there are multiple dimensions in a data~~base~~warehouse. ROLAP, or Relational OLAP is a OLAP in which the RDBMS (relational dbms) system is made use of an we can use the sql commands to perform the operations. ROLAP is interbwined with the RDBMS so that the relational dbms operations can be performed. It is consists of highly structured data and therefore, querying is easy and can handle large datasets. HOLAP or hierarchial OLAP is the architechure where the datasets are present in a hierarchial manner and we can access the datasets based on the priority of the dataset. It uses the tree datastructure for analytical processes therefore, it is much more complex than other OLAP architectures and it is efficient.

3(b) If a datawarehouse consists of three dimensional time, doctor, patient and

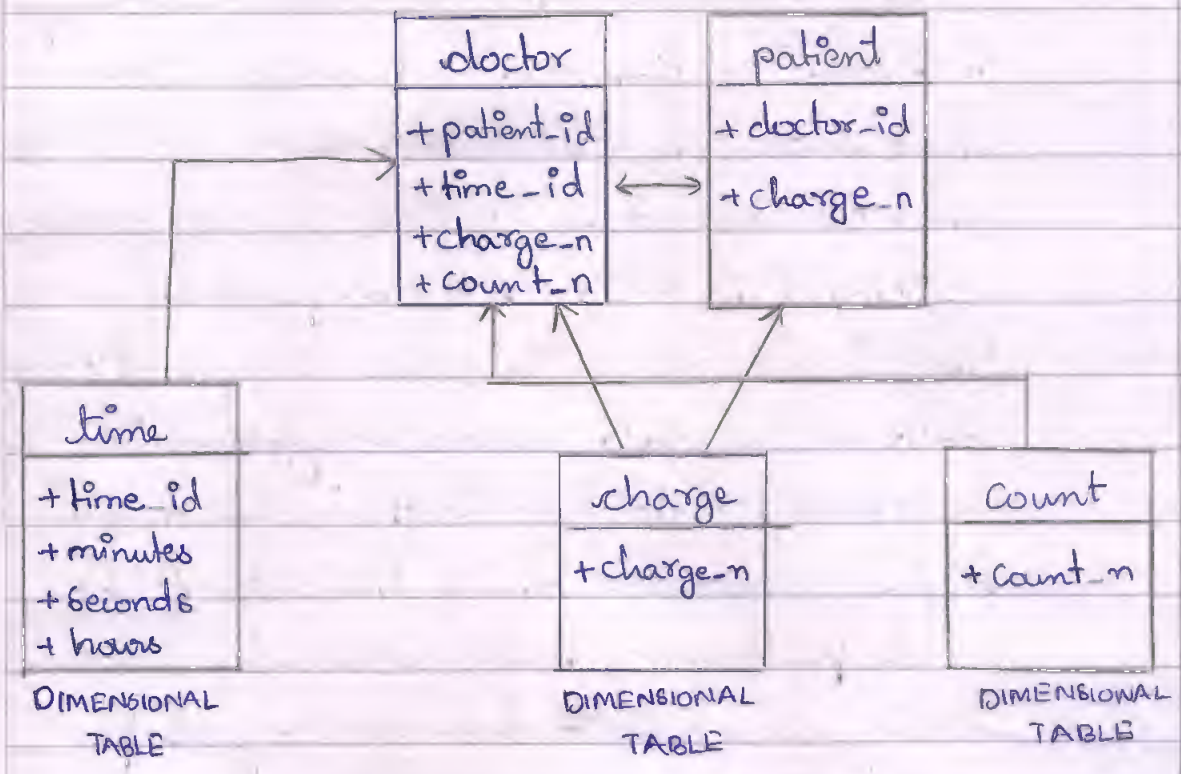
Q.No.

two measures, count and charge, where charge is the fee that a doctor charges a patient for a visit, the three classes of schemas that are popularly used for modelling data warehouses are star schema, snowflake schema and constellation schema classes. These are different classes that are used to popularly for the modelling of data warehouses. The star schema class is the schema in which there is a single flat table and multiple dimension tables. All dimension tables are then connected to the main table like a star structure where it acts as a centre node for all of the destination tables. The star schema is an outdated model, that very few people use because it limits the number of attributes. Therefore, snowflake is the widely used because it is volatile and in snowflake, similarly, dimension tables are connected to the main table, they also have sub-dimension tables. The dimension tables, therefore will extend beyond the main table

Q.No.

and becomes bigger like the snowflake. It is a widely used schema. If the schema has more that one table, then it called a constellations because the main tables along with their own dimensional tables, are also connected with each other.

(ii)



Constellation schema.

7(a) There are different types of data used for cluster analysis. Clustering is important to perform data analysis. Clustering is the

Q.No.

process, in which related data is grouped together. The different types of data are Hierarchial, Grid based, Model-based, Density based and Statistical data. In data we used cluster the data based on the algorithms where we group the similar data together. We use algorithms like K-mean and K-medoids clustering method. In K-mean method, the data is clustered by the mean of the clusters and we repeat until there is no change. K-Medoid is similar but we perform until there is no decreasing value change. Hierarchial data is the data where agglomerative - we combine data from bottom down and divise - where we divide the data. It has no shape. Density based, we cluster the high density areas and it has arbitary shape. Grid based - we perform on the shape rather than the data. All of this clustering is based on the different types of data.

The data is of different types — hierarchial data, binary data, numerical data,

Q.No.

miscellaneous data etc. In hierarchial data, the data is organised and is ranked in a hierarchy system. So we use bottom up or top down approach. We have numerical data, where the data is in the type of integers only so numeric values that are used for the count or numbers in the data. Binary data is the data that only consist of 0/1's therefore is used to represent only the presence or absence of something. Other types of data include, alphanumeric codes where both characters & numerals are present and images or text files are present in the different types of data. For data clustering, the data should be structured and it differs for each clustering technique mentioned above.

7(b) Centeroid: A1 (2,10), B1 (5,8), C1 (1,2)

Q.No.

	A1 ~~A2~~	~~A3~~ B1 ~~B2~~	~~B3~~ C1 ~~C2~~	Centre
A1	0	3.6	8.06	1
A2	5	4.2	3.16	3
A3	10	5	7.28	2
B1	3.6	0	8.9	1
B2	7.07	3.6	6.7	2
B3	7.21	4.1	5.3	2
C1	8.06	7.21	0	3
C2	2.23	1.41	7.6	2

Centre : (4.5, 7.5)
(6.25, 5.5)
(1.5, 3.5)

(ii)

~~A~~	A1	B1	C1	old	new
A1	0	3.6	8.06	1	1
A2	5	4.2	3.16	3	3
A3	10	5	7.28	2	2
B1	3.6	0	8.9	1	1
B2	7.07	3.6	6.7	2	2
B3	7.21	4.1	5.3	2	2
C1	8.06	7.21	0	3	3
C2	1	1.41	7.6	2	1

Q.No.

	A1	B1	C1	old	new
A1	0	3.6	8.06	1	1
A2	5	4.2	3.16	3	3
A3	10	5	7.28	2	2
B1	3.6	0	8.9	1	1
B2	7.07	3.6	6.7	2	2
B3	7.21	4.1	5.3	2	2
C1	8.06	7.21	0	3	3
C2	~~2.23~~ 1	1.41	7.6	~~2~~ 1	1

∴ The final clusters are: {A1, B1, C2}, {A3, B2, B3}, {A2, C1}

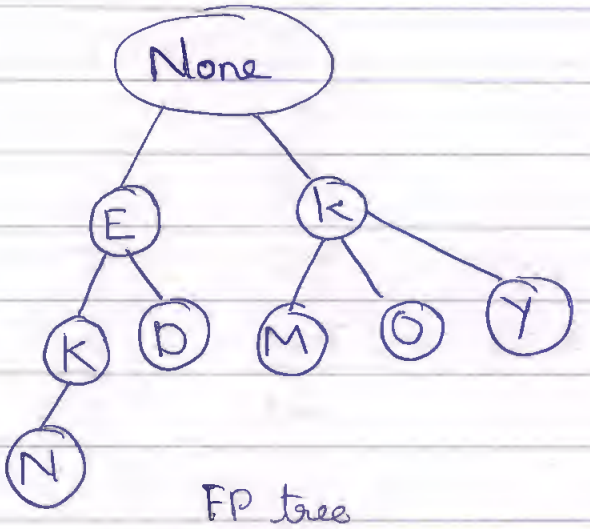
5(b)

tID	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}

Minimum support count = 3

Q.No.	

clusters: {E, K} {K, O} {K, M}
{K, Y}



FP tree

5 (a) Decision tree induction is an induction where we make use of decision tree to find out the patterns. In decision tree, the data is stored in the form of a tree with it holding different values. We use the decision tree to find out the patterns. Decision tree is therefore generated so that we can recognise the patterns. Decision tree is generated from the training tuples. Training tuples are nothing but the dataset that is used for the processing and generate the decision tree. Decision tree are made so that we can

Q.No.

transform them to the rules and we can perform any operations. Training tuples are the datasets where the data has been generated and it is used for training the decision tree. The training tuples are used to as a dataset because the decision tree holds the data which has been generated only from training tuples because the training tuples are used to generate the decision tree with the help of an algorithms. Decision tree induction refers to generation of tree. We let the training tuples undergo algorithms so that a tree is formed and unnecessary data is pruned and only the required decision tree is inducted.

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

Q.No.	

$\frac{2000+10000}{5}$, $\frac{12000}{5}$ = 4000

$-4000 - 3000 + 4000 = -1000$

$16+36 \quad 5$

$1+16 \quad 16+64$

$25+25+100+600+1600+3600+7600$

$4+9 \quad 16+36 \quad 36+9 \quad 9+4$

$-2000 - 1000 + 3000 + 6000 = 2000$

2-base → $\frac{25000}{4+1\ 5}$, $\frac{50000}{25000}$ = 0.2

$1+1$

$9+9=\sqrt{8}$

$1+9=9+16$

$36+64$

$A1,B1 = \sqrt{(5-2)^2+(10-8)^2} = 9+4$

$A1,C1 = \sqrt{(2-1)^2+(10-2)^2} = \sqrt{1+64}$

$A2,A1 = \sqrt{5^2} = 5$

A2

A1(2,10)

B1(5,8)

B1(1,2)

$\left(\frac{2+7}{2}, \frac{10+5}{2}\right) = (4.5, 7.5)$

$\left(\frac{8+7+6+4}{4}, \frac{4+5+4+9}{4}\right)_2$

$(6.25, 5.5)$

$\left(\frac{2+1}{2}, \frac{5+2}{2}\right)$ $(1.5, 3.5)$