Printed page:03 …. **Subject Code: AMTCSE0114**

Roll No: | | | | | | | | | | | | |

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

MASTER OF TECHNOLOGY (M. Tech)

(SEM: 1 Theory Examination (2020-2021)

SUBJECT NAME: DATA WAREHOUSING & DATA MINING

Time: 3 Hours **Max. Marks:70**

General Instructions:

- All questions are compulsory. Answers should be brief and to the point.
- This Question paper consists of 03 pages & 8 questions.
- It comprises of three Sections, A, B, and C. You are to attempt all the sections.
- Section A -Question No- 1 is filling the blanks carrying 1 mark each, Question No- 2 is true/false carrying 2 mark each. You are expected to answer them as directed.
- Section B - Question No-3 is Long answer type -I questions with external choice carrying 4 marks each. You need to attempt any five out of seven questions given.
- Section C - Question No. 4-8 are Long answer type –II (within unit choice) questions carrying 7 marks each. You need to attempt any one part *a* or *b.*
- Students are instructed to cross the blank sheets before handing over the answer sheet to the invigilator.
- No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.

SECTION – A

1.	**Fill in the blanks.**		**[5x1=5]**	**CO**
	a.	Data warehouse systems provide multidimensional data analysis capabilities, collectively referred to as _____.	**(1)**	**CO 1**
	b.	_____methods apply transformations to obtain a reduced or "compressed" representation of the original data.	**(1)**	**CO 2**
	c.	A hierarchical method can be classified as being either _____ or _____based on how the hierarchical decomposition is formed.	**(1)**	**CO 4**
	d.	A _____can contain additional-dimensions and measures for multimedia information, such as color, texture, and shape.	**(1)**	**CO 5**
	e.	_____ is the process of finding a model that describes and distinguishes data classes or concepts.	**(1)**	**CO 3**
2.	**State whether the following statements are true or false. Give reasons in support of your answer.**		**[5×2=10]**	**CO**
	a.	Holdout, random sampling, cross-validation, and bootstrapping are typical methods used for rule based classification.	**(2)**	**CO 3**
	b.	Cluster analysis is a means for discovering and grouping together (clustering) sets of observations that are closely related	**(2)**	**CO 4**
	c.	Binary attributes are special case of ordinal attributes.	**(2)**	**CO 1**
	d.	When searching/or similarities in multimedia data, we can search based on either the data description or the data content.	**(2)**	**CO 5**
	e.	The term "noise" has a technical meaning in data mining referring to the distortion of data from their true value and/or the addition of spurious objects	**(2)**	**CO 2**

SECTION – B

CO

3. Answer any five of the following- **[5x4=20]**

a.	What do you mean by noisy data? Briefly compare Data cleaning, data transformation concepts. You may use an example to explain your point(s).	**(4)**	**CO 1**
b.	Why is *naive Bayesian classification* called "naive"? Briefly outline the major ideas of naive Bayesian classification.	**(4)**	**CO 3**
c.	What do you mean by data reduction? Why it is used? Write names of 3 data reduction strategies.	**(4)**	**CO 2**
d.	Briefly describe and give examples of *hierarchical* methods to clustering.	**(4)**	**CO 4**
e.	Give an application example where global outliers, contextual outliers, and collective outliers are all interesting.	**(4)**	**CO 4**
f.	What is the use of Weka?	**(4)**	**CO 5**
g.	Illustrate the implementation of Data Warehouse by using efficient Data cube computation with the help of suitable example.	**(4)**	**CO 5**

SECTION – C

CO

4 Answer any one of the following- **[5×7=35]**

a. The following table consists of training data from an employee database. The data have been generalized. For example, "31 : : : 35" for *age* represents the age range of 31 to 35. For a given row entry, *count* represents the number of data tuples having the values for *department*, *status*, *age*, and *salary* given in that row. **(7)** **CO 3**

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let *status* be the class label attribute.

i. How would you modify the basic decision tree algorithm to take into consideration the *count* of each generalized data tuple (i.e., of each row entry)?
ii. Use your algorithm to construct a decision tree from the given data.
iii. Given a data tuple having the values *"systems," "26 . . . 30,"* and *"46–50K"* for the attributes d*epartment*, *age*, and s*alary*, respectively, what would a naıve Bayesian classification of the *status* for the tuple be?

b. How MOLAP, ROLAP, HOLAP are related or different from each other? Distinguish OLTP and OLAP? **(7)** **CO 3**

5. Answer any one of the following-

a. Suppose that the data mining task is to cluster points (with .x, y/ representing location) into three clusters, where the points are **(7)** **CO 4**

A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9)

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

i. The three cluster centers after the first round of execution.

ii. The final three clusters.

b. Explain DBSCAN. Prove that in DBSCAN, the density-connectedness is an equivalence relation. **(7)** **CO 4**

6. Answer any one of the following-

a. Suppose that a data warehouse for *Big University* consists of the four dimensions *student*, *course*, *semester*, and *instructor*, and two measures *count* and *avg grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination. **(7)** **CO 1**

i. Draw a *snowflake schema* diagram for the data warehouse.

ii. Starting with the base cuboid [*student*, *course*, *semester*, *instructor*], what specific ***OLAP*** *operations* (e.g., roll-up from *semester* to *year*) should you perform in order to list the average grade of *CS* courses for each *Big University* student.

iii. If each dimension has five levels (including all), such as "*student* < *major* < *status* <*university* < all", how many cuboids will this cube contain (including the base and apex cuboids)?

b. Data warehouse can be modelled by either a *star schema* or a *snowflake schema*. Briefly describe the similarities and the differences of the two models, and then analyse their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer. **(7)** **CO 1**

7. Answer any one of the following-

a. Explain with diagrammatic illustration about data mining as a step in the process of knowledge discovery. **(7)** **CO 2**

b. List and discuss the steps for integrating a data mining system with a data. **(7)** **CO 2**

8. Answer any one of the following-

a. Explain Classification and Prediction Analysis of Multimedia Data. How do we measure the accuracy of a classifier? **(7)** **CO 5**

b. Explain Multidimensional Analysis and descriptive mining of Multimedia Data. **(7)** **CO 5**