Sardar Vallabhbhai Institute of Technology, Surat
Department of Computer Science and Engineering
End Semester Exam – November 2021
B. Tech- IV (CO) 7th Sem
Data Warehousing and Mining (CO415)

Date : 3rd Dec 2021 Time : 9.30 – 12.30 Marks 100

nstruction:

1. **Attempt any 10 questions**
2. No study material is allowed in the examination
3. You may need calculator for complex calculations
4. No electronics devices , wrist watch etc. allowed
5. Write New answer on new page

Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, initially A1, B1, and C1 as the centre of each cluster, respectively. where the points are

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

Use the k-means algorithm to show

(a) The three cluster centres after the first round of execution.
(b) The final three clusters after the first round of execution

Use following distance algorithm

(a) Echlin Distance
(b) Manhattan Distance

The data is given in a Table. The output attribute is life insurance promotion.

Magazine Promotion	Watch Promotion	Credit Card Insurance	Gender	Life Insurance Promotion
Yes	No	No	Male	No
Yes	Yes	Yes	Female	Yes
No	No	No	Male	No
Yes	Yes	Yes	Male	Yes
Yes	No	No	Female	Yes
No	No	No	Female	No
Yes	Yes	Yes	Male	Yes
No	No	No	Male	No
Yes	No	No	Male	No
Yes	Yes	No	Female	Yes

Use naïve Bayes classifier to determine the value of life insurance promotion for {Yes, Yes, No, Female, ? }

Construct a decision tree for the following dataset using ID3 Algorithm. Predict the mode of the Transport for the given data point : <Female, 2, standard, high> .

Gender	Car Ownership	Travel cost	Income level	Transport mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

4 The data set of an employee contains an attribute: AGE.
AGE = { 13, 15, 16, 16, 19, X, 20, 21, 22, 22, 25,25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, Y, 45, 46, 52, 70} is given in increasing order.

a. Explain imputation using KNN.
b. Find the value of the X and Y for above data using KNN
c. Find the values of X and Y using Median value replacement.

5a. Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. 5
Partition them into three bins by each of the following methods:
(a) equal-frequency (equal-depth) partitioning (b) equal-width partitioning

5b A bank wants to develop a classifier that guards against fraudulent credit card transactions. 5

a. Outline methods to address the class imbalance problem.
b. Design the framework for a quality classifier based on a large set of non-fraudulent examples and a very small set of fraudulent cases.

6 The survay of patient is held and samples are collected for two groups of patients. Sample A denotes the patients admited for Flu symptoms and Sample B denotes patents admitted for Heart issue. The overall patients are divided into 5 getogories as per their socio economic data. The analysis is to be performed amining to investigate the difference in the distribution of the patients by social class differed in these two units. (Considering Null Hypothesis)

Category	Samples	
	A	B
I	17	5
II	25	21
III	39	54
IV	42	49
V	32	25

7a Suppose an IT company has two stores that sell computers. The company recorded the number of sales each store made each month. In the past 12 months, we have the following numbers of sold computers:
Store 1:350, 460, 20, 160, 580, 250, 210, 120, 200, 510, 290, 380.
Store 2:520, 180, 260, 380, 80, 500, 630, 420, 210, 70, 440,
Compare the two stores sales performance with the help of Box Plot.

7 b Explain the statistical terms used in identifying association between different object sets: Support, Confidence, Lift.
Enlist their limitations and show the relationship between each of them with suitable example.

8 Derive an Association between different items based on the transaction detail shown in the following table.
Support Threshold = 50 % and Confidence Threshold = 60%

Transaction	List of items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

List two models for a data warehouse. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Use suitable example. 5

Demonstrate stepwise creation of the FP tree for the given data set. 5

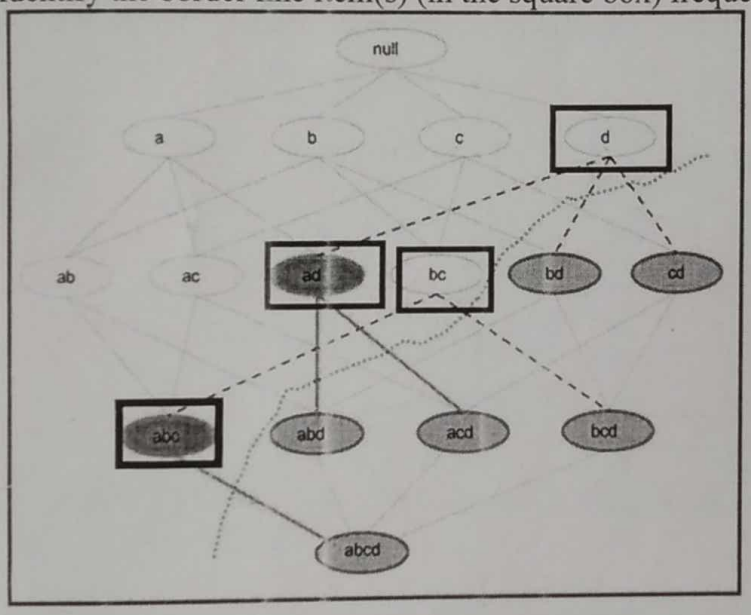
Transaction ID	Items
T1	{E, K, M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O, O}

Explain any two normalization techniques for two different data types with suitable examples

Explain measures of central tendency with example. Demonstrate their use for filling in missing values with suitable examples

Design a data warehouse for a regional weather bureau. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for more than 10 years. (Choose appropriate scheme) 5

Define Maximal Frequent Itemset. In the following graph of relationship the item(s) below the border line indicate non frequent items and items above the line indicates frequent items. Identify the border line item(s) (in the square box) frequent or not with proper justification 5



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.

(a) Draw a snowflake schema diagram for the data warehouse.

(b) 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.

(c) 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)?