20/12

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

PRN: 1032210755

Term End Examination

Dec 2023

CET2012B - Data Engineering Concepts

Question Paper ID: 027176

Faculty/School	Engineering and Technology	**Term**	Semester V
Program	TY B.Tech CSE	**Duration**	**2 Hours 30 Minutes**
Specialization		**Max. Marks**	**70**

Answer any 7 questions.

Each question has 10 marks.

Assume suitable data if necessary.

Draw appropriate diagrams if applicable.

Section - 1 (7 X 10 Marks)
Answer any 7 questions

1	Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity 1.. Time in terms of AM or PM. 2. Brightness as measured by a light meter. 3. Brightness as measured by people's judgments. 4. Angles as measured in degrees between 0◦ and 360◦ 5. Bronze, Silver, and Gold medals as awarded at the Olympics. 6. Height above sea level. 7. Number of patients in a hospital. 8. ISBN numbers for books. 9. Number of local calls in a month 10. Price of your textbook	**10 marks**	CO1	Understanding
2	A. Distinguish between noise and outliers. How to detect outliers using interquartile range (IQR) B. Calculate Interquartile Range (IQR) for the given sample size 62,63,64,64,70,72,76,77,81,81	**10 marks**	CO1	Understanding
31	Consider the data warehouse of the train application. Draw a star schema and snowflake schema for the data warehouse with hierarchies for the Passenger, train, date and station dimensions.	**10 marks**	CO3	Applying

8, 4, 1, 5, 6, 9, 1, 2.1

<table>
<tr><td>4</td><td>A. Explain the role of Metadata in Data Warehouse .
B. Elaborate the three perspectives of data warehouse metadata.</td><td><b>10 marks</b></td><td>CO3</td><td>Understanding</td></tr>
<tr><td>5</td><td>A. Explain the need of data warehouse from business analyst perceptive.
B. Discuss, how designing data warehouses is very different from designing traditional operational systems?</td><td><b>10 marks</b></td><td>CO3</td><td>Remembering</td></tr>
<tr><td>6</td><td>Write a pseudocode of Apriori Algorithm. Discuss following basic concepts of Apriori Algorithm
1. Frequent Itemsets
2. Support, Confidence
3. Join Operation
4. Prune Operation
5. Association rule generation</td><td><b>10 marks</b></td><td>CO4</td><td>Evaluating</td></tr>
<tr><td>7</td><td>Suppose that the data mining task is to cluster points (with (x, y) representing location)
into three clusters, 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).
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
(a) The three cluster centers after the first round of execution.
(b) The final three clusters</td><td><b>10 marks</b></td><td>CO4</td><td>Remembering</td></tr>
<tr><td>8</td><td>A. Why do we use Decision Trees in Data Mining? Give advantages and disadvantages of Decision Trees in Data Mining?
B. Suppose we have a dataset of students and whether they passed or failed based on two features: "Study Hours" and "Attendance."
<table>
<tr><th>Student</th><th>Study Hours</th><th>Attendance</th><th>Passed</th></tr>
<tr><td>1</td><td>2</td><td>Low</td><td>No</td></tr>
<tr><td>2</td><td>3</td><td>High</td><td>No</td></tr>
<tr><td>3</td><td>5</td><td>High</td><td>Yes</td></tr>
<tr><td>4</td><td>1</td><td>Low</td><td>No</td></tr>
<tr><td>5</td><td>4</td><td>High</td><td>Yes</td></tr>
<tr><td>6</td><td>2</td><td>Low</td><td>No</td></tr>
<tr><td>7</td><td>6</td><td>High</td><td>Yes</td></tr>
<tr><td>8</td><td>3</td><td>Low</td><td>No</td></tr>
</table>

• Calculate the entropy for the whole dataset.
• Evaluate the information gain for Study Hours and Attendance.</td><td><b>10 marks</b></td><td>CO4</td><td>Remembering</td></tr>
</table>

9 Justify is Apriori algorithm supervised or unsupervised? Apply apriori algorithm to the following data set

Given: Minimum support value is 0.3, Confidence Threshold is 60%

10 marks | CO4 | Remembering

Transaction ID	Item purchased
101	Strawberry,Litchi,Orange
102	Strawberry, Butterr fruit
103	Butter fruit, Vanilla
104	Strawberry, Litchi, Orange
105	Banana, Orange
106	Banana
107	Banana,Butter fruit
108	Strawberry, Litchi, Apple, Orange
109	Apple, Vanilla
110	Strawberry, Litchi

10 Briefly outline the major steps of Decision Tree(DT) classification.
Draw the Decision Tree and illustrate the steps with the help of below given example. 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.

10 marks | CO4 | Evaluating

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

END OF QUESTION PAPER