University of Mumbai

Examinations Summer 2022

Time: 2 hour 30 minutes Max. Marks: 80

T.E.(Information Technology Engineering)(SEM-VI)(Choice Base Credit Grading System) (R-2020-21) (C Scheme) / 89381 - Data Mining & Business Intelligence DATE: 18/5/2022 QP CODE: 91760

<table>
<tr><td>Q.1</td><td>Choose the correct option for following questions. All the Questions are compulsory and carry equal marks (2 marks each)</td></tr>
<tr><td>1.</td><td>If dimensionality reduction is performed on a record data matrix, the transformed data matrix_____</td></tr>
<tr><td>Option A:</td><td>has reduced number of rows</td></tr>
<tr><td>Option B:</td><td>has reduced number of columns</td></tr>
<tr><td>Option C:</td><td>has reduced number of both rows and columns</td></tr>
<tr><td>Option D:</td><td>has same number of rows and columns</td></tr>
<tr><td></td><td></td></tr>
<tr><td>2.</td><td>Consider the following data: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. Partition the given data with Bin size: 4. What is the output obtained after smoothing the data by Bin Boundaries.</td></tr>
<tr><td>Option A:</td><td>Bin 1: 4, 4, 4, 15 Bin 2: 21, 21, 25, 25 Bin 3: 26, 26, 26, 34</td></tr>
<tr><td>Option B:</td><td>Bin 1: 4, 4, 15, 15 Bin 2: 21, 21, 21, 25 Bin 3: 26, 26, 34, 34</td></tr>
<tr><td>Option C:</td><td>Bin 1: 4, 15, 15, 15 Bin 2: 21, 25, 25, 25 Bin 3: 26, 26, 26, 34</td></tr>
<tr><td>Option D:</td><td>Bin 1: 4, 4, 4, 15 Bin 2: 21, 25, 25, 25 Bin 3: 26, 26, 26, 34</td></tr>
<tr><td></td><td></td></tr>
<tr><td>3.</td><td>Knowledge discovery in databases is referred to</td></tr>
<tr><td>Option A:</td><td>Non Trivial process of choosing dataset</td></tr>
<tr><td>Option B:</td><td>Non Trivial process for identifying useful patterns in data</td></tr>
<tr><td>Option C:</td><td>Non Trivial process for identifying invalid patterns in data</td></tr>
<tr><td>Option D:</td><td>Non Trivial process of creating patterns in data</td></tr>
<tr><td></td><td></td></tr>
<tr><td>4.</td><td>For the given confusion matrix compute recall
Predicted data
Actual data
<table><tr><th>Cancer Classes</th><th>Yes</th><th>No</th><th>Total</th></tr><tr><td>Yes</td><td>90</td><td>210</td><td>300</td></tr><tr><td>No</td><td>140</td><td>9560</td><td>9700</td></tr><tr><td>Total</td><td>230</td><td>9770</td><td>10000</td></tr></table></td></tr>
<tr><td>Option A:</td><td>20%</td></tr>
<tr><td>Option B:</td><td>30%</td></tr>
<tr><td>Option C:</td><td>40%</td></tr>
<tr><td>Option D:</td><td>45%</td></tr>
<tr><td></td><td></td></tr>
<tr><td>5.</td><td>You are given reviews of food quality of few restaurants as Good, Average or Poor. Finding reviews of a new restaurant is an example of_____</td></tr>
<tr><td>Option A:</td><td>Classification</td></tr>
<tr><td>Option B:</td><td>Regression</td></tr>
<tr><td>Option C:</td><td>Clustering</td></tr>
<tr><td>Option D:</td><td>Association mining</td></tr>
<tr><td></td><td></td></tr>
</table>

6.	BIRCH falls under which clustering approach
Option A:	Partitioning approach
Option B:	Hierarchical approach
Option C:	Density-based approach
Option D:	Distribution based approach
7.	Given {2,4,3,10,11,12,20,25,30}, Assume k=2 and initial means are m1=4, m2=11. Apply k -means clustering technique and find its output after 1st iteration
Option A:	K1= {2,3,4,10,11,12} K2= {20,30,25}
Option B:	K1= {2,3,4} K2= {10,11,12,20,30,25}
Option C:	K1= {2,3 } K2={4,10,11,12,20,30,25}
Option D:	K1= {2,3,4,10} K2={11,12,20,30,25}
8.	In one of the frequent item-set examples, it is observed that if milk and bread are bought then eggs are also purchased by the customers. After generating an association rule among the given set of items, it is inferred
Option A:	{Milk} is antecedent and {eggs} is consequent
Option B:	{Milk} is antecedent and the item set {bread, eggs} is consequent
Option C:	The item set {milk, bread} is consequent and {eggs} is antecedent
Option D:	The item set {milk, bread} is antecedent and {eggs} is consequent
9.	For the given transactional database compute confidence for the rule Milk ⇒ Beer
Option A:	20%
Option B:	50%
Option C:	40%
Option D:	60%
10.	_____ is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations.
Option A:	Data Mining
Option B:	Data dredging
Option C:	Decision support system
Option D:	Artificial Intelligence system

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Q.2 Solve any Two Questions out of Three **Marks**

A Define data warehouse. Describe different OLAP operations in detail **10**

B Apply Naive Bayes classifier algorithm to the dataset given below, and classify the unknown data sample? **10**
Given all the previous patients I've seen(below are their symptoms and their diagnosis)

chills	runny nose	headache	fever	flu ?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Do I believe that patient with following symptoms has the flu?

chills	runny nose	headache	fever	flu ?
Y	N	Mild	Y	?

C Explain multi-level and multidimensional association rules with example **10**

Q.3 Solve any Two Questions out of Three

A Suppose we have six objects with name A, B, C, D, E and F. Apply single linkage clustering and draw dendrogram for the given data. **10**

	X	Y
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

B Suppose the data for analysis includes the attribute age. The age values for data tuples are (in increasing order): **10**
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40, 45,46,52,70

i) What is mean of data? What is median of data?
ii) What is mode of data? Comment on data's modality.
iii) What is mid range of data?
iv) Give the five point summary of the data.
v) Show box plot of the data

C What is Business Intelligence (BI)? Explain BI architecture in detail **10**

Q.4 Solve any Two Questions out of Three

A Briefly explain Bagging and Boosting of classifiers **10**

B For the table given, apply Apriori algorithm and show frequent item set and strong association rules. Assume Minimum Support of 30% and Minimum confidence of 70%. **10**

TID	Items
01	1,3,4,6
02	2,3,5,7
03	1,2,3,5,8
04	2,5,9,10
05	1,4

C What is an outlier? Describe methods used for outlier analysis. **10**

****