Paper / Subject Code: 89381 / Data Mining & Business Intelligence

TE / Sem VI / 'C' Scheme / ET.

6

(3 hours) **[80 marks]**

NOTE:

1. Question No 1 is compulsory
2. Attempt any three questions from remaining.
3. Assume suitable data if necessary and state the same.

Q.1 [20]

A) Draw Data warehousing Architecture?

B) What is noisy data? How to handle noisy data?

C) Compare and contrast between OLTP and OLAP.

D) Explain concept of information gain and gini value used in decision tree algorithm.

Q.2

A) What is Data mining? Explain KDD process with diagram. [10]

B) Consider we have age of 29 participants in a survey given to us in sorted order. [10]
5, 10, 13, 15, 16, 16, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70, 85
Explain how to calculate mean, median, standard deviation, 1st and 3rd Quartile for given data and also compute the same. Show the Box and Whisker plot for this data.

Q.3

A) Explain market Basket Analysis with example. [10]

B) Consider Training dataset as given below. Use Naive Bayes Algorithm to determine whether it is advisable to play tennis on a day with hot temperature, rainy outlook, high humidity and no wind? [10]

Outlook	temperature	Humidity	Windy	Class
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	No
overcast	cool	normal	true	Play
sunny	mild	high	false	No
sunny	cool	normal	false	Play
rain	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	high	true	Play
overcast	hot	normal	false	Play
rain	mild	high	true	No

a

Q.4

A) What is an outlier? Explain various methods for performing outlier analysis. [10]

B) Use the Apriori algorithm to identify the frequent item-sets in the following database. Then extract the strong association rules from these sets. Assume Min. Support = 50% Min. Confidence=75% [10]

Tid	a	b	c	d	e	f	g
Items	1,2,4,5,6	2,3,5	1,2,4,5	1,2,4,5	1,2,3,4,5,6	2,3,4	1,2,4,5

Q.5

A) Cluster the following eight points (with (x, y) representing locations) into three clusters: [10]

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Assume Initial cluster centers are at: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centres after the second iteration.

B) Compare star schema, Snow flakes schema and star constellation [10]

Q.6 Write short note on following (Any 4) [20]

A) Dimensional Modeling.

B) Random Forest Technique.

C) Decision Tree Induction.

D) Cross Validation.

E) DBSCAN Algorithm

****

T.E - VI Sem - ~~AI & DS~~ IT

Paper / Subject Code: 37311 / Data Mining & Business Intelligence

11/12/2023

6

(3 Hours) **[Total Marks: 80]**

NOTE:

1. Question No 1 is compulsory
2. Attempt any three questions from remaining.
3. Assume suitable data if necessary and state the same.

Q.1 A) Explain types of attributes used in data exploration **(10)**

B) Explain DBSCAN algorithm with example. **(10)**

Q.2 A) Explain K means algorithm in detail. Apply K-means Algorithm to divide the given set of values {2,3,6,8,9,12,15,18,22} into 3 clusters **(10)**

B) Compare Bagging and Boosting of a classifier **(10)**

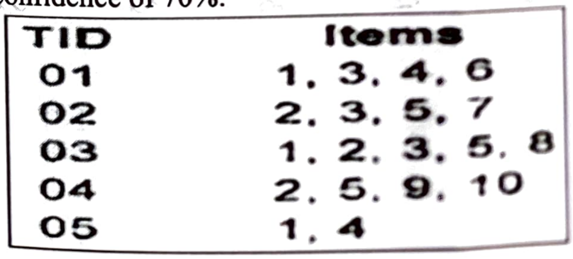
Q.3 A) Explain Multilevel and Multidimensional Association rules with suitable examples **(10)**

B) Using the given training dataset classify the following tuple using Naïve Bayes Algorithm: <Homeowner: No, Marital Status: Married, Job experience:3> **(10)**

Homeowner	Marital Status	Job experience (in years)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes

Q.4 A) Define data mining. Explain KDD process with help of a suitable diagram **(10)**

B) For the table given perform 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

5

Q.5 A) What is noisy data? How to handle it (10)
For the following data D={4,8,9,15,21,21,24,25,26,28,29,34}
Number of bins =3
Perform the following:
i. Partition into equal frequency bins
ii. Smoothing by bin means
iii. Smoothing by bin boundaries

B) Define data warehouse. Explain data warehouse architecture with help of a diagram (10)

Q.6 A) What is an outlier? List types of outliers. Describe methods used for outlier analysis. (10)

B) Design BI system for Fraud Detection? Explain all steps from data collection to decision making (10)

①

Paper / Subject Code: 89381 / Data Mining & Business Intelligence

E/sem VI/IT/CBCGS/R-20-21/CSch / DMBI

Duration: 3hrs ③ [Max Marks: 80]

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

1 Attempt any FOUR [20]

A Draw a three tier data warehousing architecture

B Data : 4, 8, 15, 21, 21, 24, 25, 28, 34
Divide data in 3 bins (equal frequency) and perform smoothing by bin means and smoothing by bin boundaries on every bin

C How to calculate correlation coefficient for two numeric attributes and also comment on the significance of this value

D Write a short note on support and confidence

E Explain the concept of information gain which is used in decision tree algorithm?

2 A Describe any two methods of data reduction [10]

B Compare star schema, snowflake schema and fact constellation [10]

3 A Write and explain Bayes classification algorithm [10]

B Write the steps of Ada-boost algorithm [10]

4 A How is data mining used in Business Intelligence? [10]

B Give the overview of partition clustering methods [10]

5 A How can we further improve the efficiency of Apriori-based mining? [10]

B Explain OLAP operations with the examples [10]

6 A Describe the classification performance evaluation measures that are obtained from confusion matrix? [10]

B Use the normalization methods to normalize the following group of data: [10]
200, 300, 400, 600, 1000
Use min-max normalization by setting min = 0 and max = 1 and z-score normalization

****

18/05/22

TE IT Sem VI 'C' Scheme DMBI QP Code: 91760

University of Mumbai

Examinations Summer 2022

6 1

Time: 2 hour 30 minutes Data Mining & Business Intelligence Max. Marks: 80

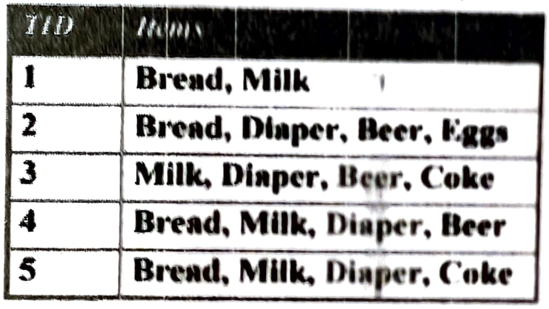
Q.1	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks (2 marks each)**
1.	If dimensionality reduction is performed on a record data matrix, the transformed data matrix______
Option A:	has reduced number of rows
Option B:	has reduced number of columns
Option C:	has reduced number of both rows and columns
Option D:	has same number of rows and columns
2.	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.
Option A:	Bin 1: 4, 4, 4, 15 Bin 2: 21, 21, 25, 25 Bin 3: 26, 26, 26, 34
Option B:	Bin 1: 4, 4, 15, 15 Bin 2: 21, 21, 21, 25 Bin 3: 26, 26, 34, 34
Option C:	Bin 1: 4, 15, 15, 15 Bin 2: 21, 25, 25, 25 Bin 3: 26, 26, 26, 34
Option D:	Bin 1: 4, 4, 4, 15 Bin 2: 21, 25, 25, 25 Bin 3: 26, 26, 26, 34
3.	Knowledge discovery in databases is referred to
Option A:	Non Trivial process of choosing dataset
Option B:	Non Trivial process for identifying useful patterns in data
Option C:	Non Trivial process for identifying invalid patterns in data
Option D:	Non Trivial process of creating patterns in data
4.	For the given confusion matrix compute recall

		Predicted data		
	Cancer Classes	**Yes**	**No**	**Total**
Actual data	Yes	90	210	300
	No	140	9560	9700
	Total	230	9770	10000

Option A:	20%
Option B:	30%
Option C:	40%
Option D:	45%
5.	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______
Option A:	Classification
Option B:	Regression
Option C:	Clustering
Option D:	Association mining

TE IT Sem VI 'C' Scheme DMBI QP Code: 91760

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



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

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

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

TE IT Sem VI 'C' Scheme DMBI QP Code: 91760

(4)

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

****