C **Reg. No. :** □□□□□□□□□□

Question Paper Code: 99209

B.E./B.Tech. DEGREE EXAMINATION, NOV 2023

Elective

Computer Science and Engineering

19UCS909- Data Mining

(Régulations 2019)

Duration: Three hours Maximum: 100 Marks

Answer ALL Questions

PART A - (5 x 1 = 5 Marks)

1. Strategic value of data mining is___________. CO1- U

(a) Cost-sensitive (b) Work-sensitive

(c) Time- sensitive (d) Technical- sensitive

2. If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the support of bread and jam is _______. CO2- App

(a) 2%. (b) 20% (c) 3%. (d) 30%.

3. Which of the following criteria is not used to decide which attribute to split next in a decision tree: CO1- U

(a) Gini index (b) Information gain (c) Entropy (d) Scatter

4. Which is needed by K-means clustering? CO1- U

(a) defined distance metric (b) number of clusters

(c) initial guess as to cluster centroids (d) all of the above

5. Data mining can be used to improve ___. CO1- U

(a) Efficiency (b) Quality of data (c) Marketing (d)All of the above

PART – B (5 x 3= 15 Marks)

6. State the various issues in data mining? CO1- U

7. What is meant by constraint based mining? CO1- U

8. What is rule based classification? How the rule is assessed? CO1- U

9. State the various requirements of clustering CO1- U

10. What is meant by text mining? CO1- U

PART – C (5 x 16= 80 Marks)

11. (a) The following data (in increasing order) for the attribute age: 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. CO2- App (16)

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].

(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 35 for age.

(iv) Comment on which method you would prefer to use for the given data, giving reasons as to why.

Or

(b) Suppose a group of 12 sales price records has been sorted as follows: CO2- App (16)

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.

(i) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(ii) How might you determine outliers in the data?

(iii) What other methods are there for data smoothing?

12. (a) Explain various kinds of Association Rules Mining CO1- U (16)

Or

(b) Describe the method of generating frequent item sets with candidate generation Using Apriori Algorithm with an example. CO1- U (16)

13. (a) Explain the concept of Bayesian network in representing knowledge in an uncertain domain with the following problem “Consider a situation in which we want to reason about the relationship between smoking and lung cancer. We’ll use 5 Boolean random variables representing "has lung cancer" (C), "smokes" (S), "has a reduced life expectancy" (RLE), "exposed to second-hand smoke" (SHS), and "at least one parent smokes" (PS). Intuitively, we know that whether or not a person has cancer is directly influenced by whether she is exposed to second-hand smoke and whether she smokes. Both of these things are affected by whether her parents smoke. Cancer reduces a person’s life expectancy”. CO2- App (16)

Or

(b) You are a robot in a lumber yard, and must learn to discriminate Oak wood from Pine wood. You choose to any one learning algorithm to classify the sample data. You are given the following (noisy) examples: CO2- App (16)

Example	Density	Grain	Hardness	Class
Example #1	Light	Small	Hard	Oak
Example #2	Heavy	Large	Hard	Oak
Example #3	Heavy	Small	Soft	Oak
Example #4	Heavy	Small	Soft	Oak
Example #5	Light	Large	Hard	Pine
Example #6	Light	Small	Soft	Pine
Example #7	Heavy	Large	Soft	Pine
Example #8	Light	Large	Hard	Pine

14. (a) Cluster the following data set consisting of the scores of two variables on each of seven individuals and **k=2** using any one Clustering method. CO2- App (16)

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Or

(b) 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). CO2- App (16)

The distance function is Euclidean distance. Find out the final cluster using any cluster algorithm.

15. (a) Discuss about the various methods in Text Mining CO1- U (16)

Or

(b) Explain how data mining is used in health care analysis CO1- U (16)