C

0400CST466052302

Pages: 3

Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth Semester B.Tech Degree Regular Examination June 2023 (2019 Scheme)

Course Code: CST466

Course Name: DATA MINING

Max. Marks: 100 **Duration: 3 Hours**

PART A

Answer all questions, each carries 3 marks.

		Marks
1	List and explain any two applications of data warehouse.	(3)
2	Describe the similarities and the differences of star schema and snowflake schema.	(3)
3	Perform data smoothing by bin means on 3 equi-width bins. Data: [20,24,23,12,15,20,31,29,35,36,32,40]	(3)
4	What is the purpose of data discretization? List any two data discretization strategies.	(3)
5	How is Gain Ratio calculated? What is the advantage of Gain Ratio over Information Gain?	(3)
6	What are the requirements for a good clustering algorithm?	(3)
7	Describe any three methods to improve the efficiency of Apriori algorithm.	(3)
8	Write about the bi-directional searching technique for pruning in pincer search algorithm.	(3)
9	Describe the following activities involved in the web usage mining i) Pre-processing activity ii) Pattern analysis	(3)
10	Differentiate between web content mining and web structure mining.	(3)

PART B

Answer any one full question from each module, each carries 14 marks.

Module I

11	a)	Explain the knowledge discovery process (KDD) in databases for finding useful information and patterns in data.	(7)
	b)	Illustrate the various stages of data mining in business intelligence with a diagram.	(7)

OR

12	a)	Describe different issues in data mining.	(6)
	b)	Suppose that a data warehouse for a university consists of the following four dimensions: **student, course, semester**, and **instructor**, and two measures: **count** and **avg_grade**. (i) Draw a snowflake schema diagram for the data warehouse. (ii) Starting with the base cuboid, what specific OLAP operations should one perform in order to list the average grade of CS courses for each University student.	(8)

Module II

13 a) Suppose that the data for analysis includes the attribute cost price and the values for the data tuples are: 100, 150, 140, 115, 190, 120, 130, 125, 135, 145, 140, 150, 165, 160, 170

(i) Use min-max normalization to transform the value of 145 for cost price onto the range [0,1]. (6)

(ii) Use Z-Score normalization to transform the value 145 for cost price where the standard deviation of cost price is 120.

b) Real-world data tend to be incomplete, noisy and inconsistent. What are the various approaches adopted to clean the data? (8)

OR

14 a) Describe the various techniques for numerosity reduction in data mining. (8)

b) 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. Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, stratified sampling. Use samples of size 5 and the strata "youth," "middle-aged," and "senior." (6)

Module III

15 a) Consider the following dataset for a binary classification problem with class label "yes" and "no".

sl.no	age	income	student	credit_ rating	Class: Risky
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

(8)

The above table shows class labeled dataset of customers in a bank. Explain information gain attribute selection measure, and find the information gain of the attribute "age".

b) Explain the concept of DBSCAN algorithm along with its advantages. (6)

OR

16 a) A database contains 80 records on a particular topic of which 55 are relevant to a certain investigation. A search was conducted on that topic and 50 records were retrieved. Of the 50 records retrieved, 40 were relevant. Construct the confusion matrix and calculate the precision and recall scores for the search. (7)

b) Explain the working of PAM algorithm with an example. (7)

Module IV

17 a) A database has six transactions. Let min_sup be 33.33% and min_conf be 60%.

TID	ITEMS
T1	Cake, Bread, Jam
T2	Cake, Bread
T3	Cake, Coke, Chips
T4	Chips, Coke
T5	Chips, Jam
T6	Cake, Coke, Chips

(8)

Find frequent itemset using Apriori algorithm and generate strong association rules from the dataset.

b) Illustrate the working of Pincer Search Algorithm with an example. (6)

OR

18 a) A database has six transactions. Let min_sup be 3.

TID	ITEMS
T1	{f, a, c, d, m, p}
T2	{a, b, c, f, m}
T3	{b, f, j}
T4	{b, c, k, p}
T5	{a, f, c, e, p, m}
T6	{f, a, c, d, m, p}

(8)

Find frequent itemsets using FP growth algorithm.

b) Describe the working of dynamic itemset counting technique with suitable example. Specify when to move an itemset from dashed structures to solid structures. (6)

Module V

19 a) List and explain the different data structures used for web usage mining? (8)

b) Write any three applications of web usage mining and explain. (6)

OR

20 a) Describe different Text retrieval methods. Explain the relationship between text mining, information retrieval and information extraction. (6)

b) Explain the different traversal patterns and discovery methods in web usage data. (8)

****