C 0400CST466052301 Pages: 3

Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth Semester B.Tech Degree Supplementary Examination October 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 out the three major features of data warehouse.	(3)
2		Describe the similarities and the differences of OLTP and OLAP.	(3)
3		Perform data smoothing by bin means on 3 equi-width bins. Data: [24,27,29,16,17,31,33,29,36,37,35,44]	(3)
4		Explain concept hierarchy with an example.	(3)
5		What are the requirements for a good clustering algorithm?	(3)
6		Discuss the issues regarding the implementation of decision tree.	(3)
7		Describe any three methods to improve the efficiency of the Apriori algorithm.	(3)
8		Define support, confidence and frequent itemset in association rule mining context.	(3)
9		Describe any two-text retrieval indexing techniques.	(3)
10		Compare and contrast the focused crawling and regular crawling techniques.	(3)

PART B

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

Module I

11	a)	Explain different OLAP operations on multidimensional data with suitable examples.	(7)
	b)	Illustrate the various stages in Knowledge discovery process with a diagram.	(7)

OR

12	a)	Explain the differences between star schema and snowflake schema in a data warehouse.	(6)
	b)	Suppose that a data warehouse consists of the three dimensions: **time, doctor**, and **patient**, and the two measures: **count** and **charge**, where charge is the fee that a doctor charges a patient for a visit.	
		i. Draw a schema diagram for the above data warehouse using star schema	(8)
		ii. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2023?	

Module II

13	a)	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, cluster sampling, stratified sampling. Use samples of size 5 and the strata "youth," "middle-aged," and "senior."	(8)
	b)	Real-world data tend to be incomplete, noisy and inconsistent. What are the various approaches adopted to clean the data?	(6)

OR

14 a) Describe the various techniques for numerosity reduction in data mining. (6)

b) Why do we need data transformation? What are the different ways of data transformation? (8)

Module III

15 a) Consider the following dataset for a binary classification problem with class label as C1 and C2.

A	B	Class Label
T	F	C1
F	F	C2
T	T	C1
T	F	C2
F	F	C2
T	F	C2
T	T	C1
F	F	C2
T	T	C2
T	T	C1

(8)

i) Calculate the gain in Gini index when splitting on A and B respectively. Which attribute would the decision tree induction algorithm choose?

ii) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

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

OR

16 a) Find the first splitting attribute for the decision tree by using the ID3 algorithm with the following dataset.

Age	Competition	Type	Class (profit)
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

(8)

b) Explain the working of SLIQ algorithm. (6)

Module IV

17 a) A database has six transactions. Let min_sup be 60% and min_conf be 80%.

TID	items bought
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

(8)

Find frequent itemsets using Apriori algorithm and generate strong association rules from a three-item dataset.

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

OR

18 a) Describe the working of dynamic itemset counting technique with suitable example. Specify when to move an itemset from dashed structures to solid structures. (8)

b) Explain the partitioning algorithm for finding large itemset and explain how it removes the disadvantage of Apriori algorithm. (6)

Module V

19 a) Explain HITS algorithm with an example. (6)

b) Describe different Text retrieval methods. Explain the relationship between text mining, information retrieval and information extraction. (8)

OR

20 a) Explain how web structure mining is different from web usage mining and web content mining? Write a CLEVER algorithm for web structure mining. (6)

b) Term frequency matrix given in the table shows the frequency terms per document.

Document/terms	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

(8)

Calculate the TF-IDF value for the term T4 in document 3.

****