Reg. No. : | | | | | | | | | | | | |

Question Paper Code:U5803

B.E./B.Tech. DEGREE EXAMINATION, NOV 2024

Fifth Semester

Inormation Technology

21UIT503 – MINING AND ANALYSIS OF BIG DATA

(Regulations 2021)

Duration: Three hours Maximum: 100 Marks

Answer All Questions

PART A - (10 x 2 = 20 Marks)

1.	Compare OLTP and OLAP.	CO1	U
2.	Apply the concept hierarchy for the dimension location (Tamilnadu, Karnataka, and Kerala).	CO2	App
3.	List few techniques to improve the efficiency of apriori algorithm.	CO1	U
4.	Explain about data cleaning.	CO1	U
5.	Let x1= (1, 2) and x2= (3, 5) represent two points. Calculate the Manhattan and Euclidean distance between the two points.	CO2	App
6.	Summarize the requirements of cluster analysis.	CO1	U
7.	Infer the characteristics of big data	CO1	U
8.	Analyse, how is big data analysis helpful in increasing business revenue?	CO3	Ana
9.	Summarize on Hive.	CO1	U
10.	In HBase, what is column family?	CO1	U

PART – B (5 x 16= 80 Marks)

11.	(a)	Explain in detail the concept of multidimensional data model.	CO1	U	(16)
		Or			
	(b)	Explain in detail on various Roll-up, Drill down, Slice, Dice, and pivot operations in a multidimensional data cube.	CO1	U	(16)
12.	(a)	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (increasing order) 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. (a) What is the mean of the data? What is the median? (b) What is the mode of the data?	CO2	App	(16)

(c) What is the midrange of the data?
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
(e) Give the five number summaries of the data.
(f) Show a box plot of the data.
(g) Calculate Inter Quartile Range.(IQR)

Or

(b) Use these methods to normalize the following group of data:200, 300, 400,600,1000 — CO2 App — (16)
(i) min-max normalization by setting min=0 and max=1
(ii) z-score normalization
(iii) z-score normalization using mean absolute deviation instead of standard deviation
(iv) Decimal Scaling

13. (a) What is decision tree? Explain how classification is done using decision tree induction for the following table consists of training data from an employee database. The data have been generalized. For example, "31.....35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age and salary given in that row. — CO2 App — (16)

Department	Status	Age	Salary	Count
sales	senior	31.....35	46K.....50K	30
sales	junior	26.....30	26K.....30K	40
sales	junior	31. . . 35	31K. . . 35K	40
systems	junior	21. . . 25	46K. . . 50K	20
systems	senior	31. . . 35	66K. . . 70K	5
systems	junior	26. . . 30	46K. . . 50K	3
systems	senior	41. . . 45	66K. . . 70K	3
marketing	senior	36. . . 40	46K. . . 50K	10
marketing	junior	31.....45	41K...45K	4
Secretary	senior	46....50	36K.....40K	4
Secretary	junior	26.....30	26K.....30K	6

Let status be the class label attribute.
Use Your algorithm to construct a decision tree from the given data.

Or

	(b)	Consider five points{x1,x2,x3,x4,x5} with the following co-ordinates as a two dimensional sample for clustering: x1=(0,2), x2=(1,0), x3=(2,1), x4=(4,1) and x5=(5,3). Illustrate the k-means algorithm on the above data set. The required number of cluster is two, & initially clusters are formed from random distribution of samples: c1={x1, x2, x4} and c2= {x3, x5}.	CO2 App	(16)
14.	(a)	Explain the core components of Hadoop	CO1 U	(16)
		Or		
	(b)	List and explain significance of Map Reduce.	CO1 U	(16)
15.	(a)	Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ; Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so.	CO2 App	(16)
		Or		
	(b)	To create a table named employee with fields named Id, Name, Salary, Designation, and Dept. Generate a hive query to retrieve the employee details who earn a salary of more than Rs 30000.	CO2 App	(16)