Reg. No. : □□□□□□□□□□□□

Question Paper Code: U5803

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

Fifth Semester

Information 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. What is meant by multidimensional data? CO1-U

2. What is market basket analysis? CO1-U

3. What are the data reduction strategies? CO1-U

4. 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. CO2-App
   (a) What is the mean of the data? What is the median?
   (b) What is the mode of the data?
   (c) What is the midrange of the data?

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. What are the requirements of cluster analysis? CO1-U

7. How would you transform unstructured data into structured data? CO1-U

8. Which hardware configuration is most beneficial for hadoop jobs? CO1-U

9. Difference between Hbase and Hive CO1-U

10. Is it possible to execute Hive queries from a script file? CO1-U

PART – B (5 x 16= 80 Marks)

11. (a) Explain in detail the concept of multidimensional data model. CO1-U (16)

Or

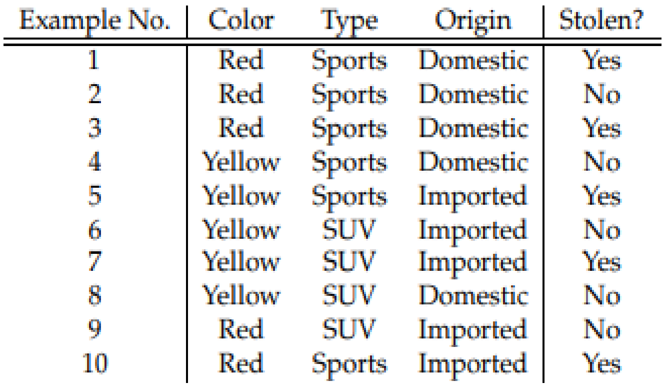
(b) Explain the process of data warehouse design. CO1 -U (16)

12. (a) Explain in detail the various steps involved in data preprocessing. CO1-U (16)

Or

(b) Explain with diagrammatic illustration that data mining as a step in the process of knowledge discovery. CO1 -U (16)

13. (a) Classify the given training data using Navie Bayes Classifiers CO2-App (16)



Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Predict the class label of the Stolen for the following test data.
Test data ={Color='red', Type='SUV', Origin='Domestic'}

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 in detail about the different types of data in big data analytics. CO1-U (16)

Or

(b) Explain in detail about Hadoop distributed file system architecture with neat diagram. 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. CO1-U (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. CO1-U (16)