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

Question Paper Code: U5803

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

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.	List the two interesting measures of an association rule.	CO1-U
2.	A retail company wants to build a data warehouse to track sales data. What are the different factors that they need to consider when designing the data warehouse?	CO2-App
3.	Give few techniques to improve the efficiency of Apriori algorithm.	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. What is the mean of the data? What is the median? i. What is the mode of the data? ii. What is the midrange of the data?	CO2-App
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	List the types of data used in cluster analysis	CO1-U
7	What are the characteristics of big data?	CO1-U
8	How would you transform unstructured data into structured data?	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 architecture of data warehousing. CO1-U (16)

Or

(b) Write detailed notes on various OLAP operations in a multidimensional data cube. CO1 -U (16)

12. (a) Apply the Apriori algorithm for discovering frequent item sets for mining association rules of the following table. Use 0.3 for the minimum support value. Illustrate each step of the Apriori algorithm. CO2-App (16)

Trans ID	Items Purchased
101	milk, bread,eggs
102	milk, juice
103	juice,butter
104	milk,bread,eggs
105	coffee,eggs
106	coffee
107	coffee , juice
108	milk, bread,cookies,eggs
109	cookies, butter
110	milk , bread

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) 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) Explain in detail about Pig architecture with neat diagram. CO1-U (16)

Or

(b) Explain the difference between Pig, Hive and HBase in detail. CO1-U (16)