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

Question Paper Code: 95803

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

Fifth Semester

Information technology

19UIT503– Mining and Analysis of Big Data

(Regulation 2019)

Duration: Three hours | Maximum: 100 Marks

Answer ALL Questions

PART A - (10 x 2 = 20 Marks)

1.	State why data preprocessing is an important issue for data warehousing and data mining?	CO1- U
2.	Differentiate between OLTP vs. OLAP	CO1- U
3.	What is market basket analysis?	CO1- U
4.	Why are decision tree classifiers so popular?	CO3- 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 categories of clustering methods.	CO1- U
7.	What is Big Data?	CO1- U
8.	What are the characteristics of big data?	CO1- U
9.	What is Hive?	CO1- U
10.	Mention key components of Hive architecture.	CO1- U

PART – B (5 x 16= 80 Marks)

11.	(a)	Use these methods to normalize the following group of data:200, 300, 400,600,1000 (a) min-max normalization by setting min=0 and max=1 (b) z-score normalization (c) Decimal Scaling	CO2- App	(16)

Or

(b) 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. Answer the following: CO2- App (16)

(a) Use smoothing by bin means to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(b) How might you determine outliers in the data?

What other methods are there for data smoothing?

12. (a) Apply FP growth for discovering frequent item sets for mining association rules of the following table. 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) 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

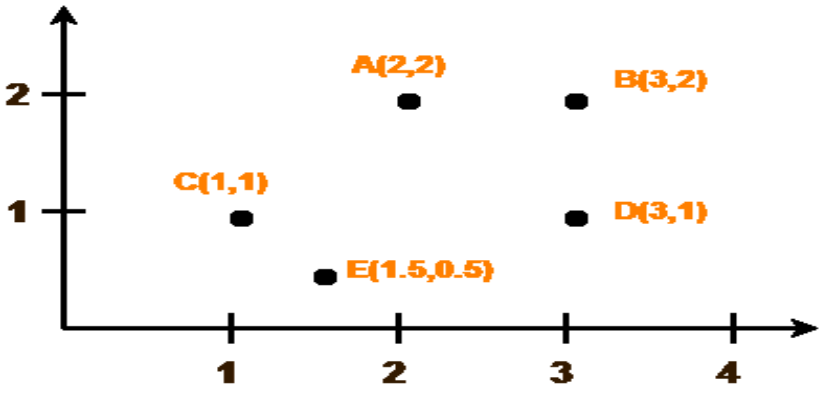
13. (a) Obtain regression equation of Y on X and estimate Y when X=55 from the following CO3- Ana (16)

Dataset

X	40	50	38	60	65	50	35
Y	38	60	55	70	60	48	30

Or

(b) Use K-Means Algorithm to create two clusters. Compare the cluster results with the K-mediods. CO3- Ana (16)



14.	(a)	What is Bigdata? Describe the main features of a big data in detail.	CO1-U	(16)
		Or		
	(b)	Explain the main characteristics of Big Data.	CO1-U	(16)
15.	(a)	ustomers 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)	Compare Pig and SQL. How SQL is differ from HiveQL.	CO1-U	(16)