Reg. No. : □□□□□□□□□□

Question Paper Code: 95803

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

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. Apply the concept hierarchy for the dimension location (Tamilnadu, Karnadaka, and Kerala) CO2-App

2. Differentiate between OLTP vs. OLAP CO1- U

3. What is market basket analysis? CO1- U

4. What are the things suffering the performance of Apriori candidate generation technique. 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. Considering the K-median algorithm, if points (0, 3), (2, 1), and (-2, 2) are the only points which are assigned to the first cluster now, what is the new centroid for this cluster? Justify. CO3-Ana

   A.(0,2) B.(2,1) C.(2,0) D.(1,2)

7. What is Big Data? CO1- U

8. What are the characteristics of big data? CO1- U

9. What is Hive? CO1- U

10. Define Sharding. CO1- U

PART – B (5 x 16= 80 Marks)

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

Or

(b) Explain in detail about the following techniques: CO1- U (16)

(a) Data Cleaning techniques

(b) Normalization techniques and

(c) Data Transformation Techniques.

12. (a) Consider the data about weather in given table below CO2- App (16)

Week	Weather	Parents	Money	Decision (category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Shopping
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Apply Navie Bayesian Classification algorithm to the above training set and predict the class label of the unknown test set

X1=(week=w11,Weather=Rainy, Parents=Yes, Money=Rich, Decision=?)

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

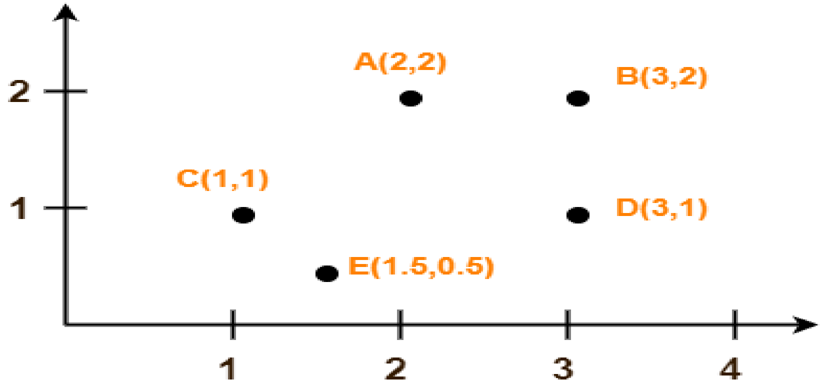
13. (a) Consider five points{x1,x2,x3,x4,x5} with the following co-ordinates as a two CO3- Ana (16)

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}.Compare the cluster results with the K-mediods

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

Or

(b) Compare Pig and SQL. How SQL is differ from HiveQL. CO1-U (16)