Reg. No. : □□□□□□□□□□

Question Paper Code: 95803

B.E./B.Tech. DEGREE EXAMINATION, NOV 2022

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. 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?

3. What is market basket analysis? CO1- U

4. What are the things suffering the performance of Apriori candidate generation technique. CO1- U

5. What are the requirements of cluster analysis? 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. What are challenges of Big Data? CO1- U

9. Define Sharding. CO1- U

10. Difference between Hbase and Hive 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 CO2- App (16)

(b) z-score normalization

(c) Decimal Scaling

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) 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) A mobile company conducted a survey about the selection of Mobile phones and the survey results are given below. CO2- App (16)

✓Predict the choices of the customers using Naïve Bayes Algorithm

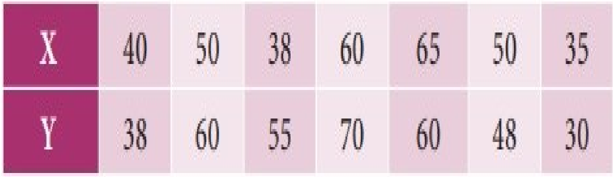
✓Compare the actual choice and predicted choice for any one tuple & test the accuracy of prediction.

Dataset

Features	Cost	Class
Good	High	Buy
Moderate	Moderate	Buy
Good	Moderate	Buy
Good	High	Buy
Moderate	Moderate	Buy
Moderate	High	Not Buy
Moderate	Moderate	Not Buy
Good	High	Not Buy
Moderate	High	Not Buy
Moderate	Moderate	Not Buy

13. (a) Obtain regression equation of Y on X and estimate Y when X=55 from the following CO2- App (16)

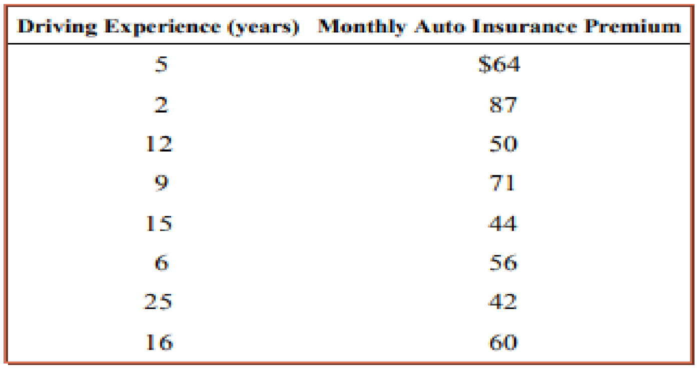
Dataset



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

Or

(b) A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums. CO2- App (16)



Driving Experience (years)	Monthly Auto Insurance Premium
5	$64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

(a) Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? Do you expect a positive or a negative relationship between these two variables?

Compute SSxx, SSyy, and SSxy.

14. (a) Explain the benefits of big data processing. CO1-U (16)

Or

(b) Explain in detail about the different types of data in big data analytics. 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)