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

# CS 505 (E4)

## Third Semester M.Sc. Degree Examination, December 2018/January 2019
## COMPUTER SCIENCE
## Data Mining Techniques (Elective – I) (Repeaters)

Time : 3 Hours Max. Marks : 70

***Note :*** *Answer **any five** questions. **All** questions carry **equal** marks.*

1. a) Describe the steps involved in data mining when viewed as a process of knowledge discovery. **7**

   b) What are the difference between mining small amount of data and large amount of data ? Give examples. **7**

2. a) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in 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. **8**

      i) Compute mean, median, mode and midrange of the data.

      ii) Give the five-number summary of the data.

      iii) Show a boxplot of the data.

   b) What is data visualization and the need for data visualization ? Describe any one method for data visualization. **6**

3. a) In real-world data, tuples with missing values for some attributes are a common occurrence. List out the reasons for missing values and describe any two methods for handling this problem. **7**

   b) What is normalization ? **7**

      i) Normalize the two attributes based on z-score normalization.

      ii) Calculate the correlation coefficient to know whether these two attributes are correlated.

| **age** | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| **%fat** | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |

**P.T.O.**

4. a) What is data warehousing ? Describe the different models of data warehousing. **8**

   b) Distinguish between distributive, algebraic and holistic measures. Give examples. **6**

5. a) For the following transaction data, compute frequent itemsets using FP-growth algorithm. Let min-sup = 3. **9**

| Tid | Items |
|---|---|
| T1 | {a, b} |
| T2 | {b, c, d} |
| T3 | {a, c, d, e} |
| T4 | {a, d, e} |
| T5 | {a, b, c} |
| T6 | {a, b, c, d} |
| T7 | {a} |
| T8 | {a, b, c} |
| T9 | {a, b, d} |
| T10 | {b, c, e} |

   b) Illustrate the limitations of support-confidence framework for the evaluation of association rules. How are the limitations overcome ? **5**

6. a) Why is näive Bayesian classification called “näive” ? Explain the major ideas of näive Bayesian classification. **6**

   b) Explain :
      i) Confusion matrix
      ii) Precision and recall
      iii) Sensitivity and specificity
      iv) ROC curves. **8**

7. a) Describe clustering by partitioning method. Explain any one partitioning method for clustering. **7**

   b) What are ensemble methods for classification ? Explain any one popular ensemble method for classification. **7**

8. a) What are outliers ? Explain any one method of identifying outliers. **7**

   b) Comment on “Curse of Dimensionality”. **7**

---