# BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
## Work Integrated Learning Programmes (WILP) Division
### M. Tech -Data Science and Engineering
### I Semester 2019-20
### End-Semester Examination

| | | |
|---|---|---|
| Course No | : | **DSECF ZC415** |
| Course Name | : | **Data Mining** |
| Nature of Exam | : | Open Book |
| Max Marks | : | 50 |
| Duration | : | 2-Hour |
| Date of Exam | : | |

No. of page: 2
No. of questions: 4

**Note:**

1. Please read and follow all the instructions given on the cover page of the answer booklet.
2. **Start each answer from a fresh page. All parts of a question should be answered consecutively**.
3. Please ensure that your answers cover necessary technical details, avoiding unnecessary text and diagrams.

---

**<u>Question 1</u>**

a) Given below age distribution of population in a town. Calculate the approximate median age. List assumptions made. **[4 marks]**

| Age range | 1-5 | 5-15 | 15-20 | 20-50 | 50-80 | 80-100 |
|---|---|---|---|---|---|---|
| Population | 200 | 450 | 300 | 1500 | 700 | 44 |

b) Consider the following scenario:
"Sam is trying to get into a Medical college for Post-graduation in India. Before applying for any college/university, he needs to take an exam for that particular college/university. Therefore, he decided to take two exams - AIIMS PG and JIPMER PG". The summary statistics of the results for each exam are given below:

| Exam | Mean | Standard Deviation |
|---|---|---|
| AIIMS PG | 151 | 10 |
| JIPMER PG | 25.1 | 6.4 |

Sam took both the exams and scored 172 in AIIMS PG and 37 in JIPMER PG.
Based on the above data, you need to figure out, in which exam did he do relatively better? Explain how did you arrive at this conclusion? **[5 marks]**

c) After mining a transaction database for frequent itemsets, there is only one largest frequent itemset of size 10. Let N be the total number of frequent itemsets (including the one of size 10). What is the minimal value of N? **[3 marks]**

**Question 2**

Below is the partial transaction table of a grovery store. The letters in the table are codes for products sold by the store. **[5+3+3=11 marks]**

| Transaction ID | List of Items |
|---|---|
| T1 | F, A, C, D, G, I, M, P |
| T2 | A, B, C, F, L, M,O |
| T3 | B, F, H, J, O, W |
| T4 | B, C, K, S, P |
| T5 | A, F, C, E, L, P, M, N |

a) Construct a FP Tree corresponding to the set of transactions given in the above table. Consider the minimum support count is 3.

b) Find the frequent itemset(s) generated by the algorithm.

c) Assuming confidence level of 50%, find the most interesting association rule involving three items

**Question 3**

Consider the following data points: **[6+3+3+3=15 Marks]**

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

a) Compute a distance matrix using Euclidean distance measure.

b) Identify the clusters that could be formed using DBScan Algorithm, assuming Epsilon is 2, min-point is 2?

c) What is the impact on the number of clusters if Epsilon is increased to $(10)^{1/2}$ (i.e. square root of 10)?

d) What conclusion can you draw from part 2) and 3)?

**Question 4**

a) Consider the distance matrix provided for data objects. The outlier score of an object is the inverse of the density around an object. The density of an object is equal to the number of objects that are within a distance of 3 units from the object. Identify the outlier using the density-based outlier detection method. **[10 marks]**

| | A | J | M | C | P | L |
|---|---|---|---|---|---|---|
| A | 0 | 12 | 3 | 4 | 1 | 2 |
| J | 12 | 0 | 2 | 8 | 7 | 10 |
| M | 3 | 2 | 0 | 9 | 6 | 5 |
| C | 4 | 8 | 9 | 0 | 5 | 1 |
| P | 1 | 7 | 6 | 5 | 0 | 2 |
| L | 2 | 10 | 5 | 1 | 2 | 0 |

b) Explain any one shortcoming of content-based filtering? **[2 marks]**