

Reg No.: \_\_\_\_\_

Name: \_\_\_\_\_

**APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY**

B.Tech Degree S8 (R,S) Exam April 2025 (2019 Scheme)

**Course Code: AIT426****Course Name: MINING OF MASSIVE DATASETS****Max. Marks: 100****Duration: 3 Hours****PART A***Answer all questions, each carries 3 marks.*

Marks

- |    |  |     |
|----|--|-----|
| 1  | Explain Bonferroni's Principle and its significance in data mining.                                      | (3) |
| 2  | Describe the difference between statistical modeling and machine learning in the context of data mining. | (3) |
| 3  | Explain the MapReduce implementation of the selection operation ( $\sigma C(R)$ ).                       | (3) |
| 4  | How do grouping and aggregation work in MapReduce?   | (3) |
| 5  | Explain the six rules that must be followed when representing a stream using buckets.                    | (3) |
| 6  | What are some fundamental issues associated with stream processing?                                      | (3) |
| 7  | How the CURE algorithm begins in clustering.   | (3) |
| 8  | What do you mean by Cluster Analysis?  | (3) |
| 9  | How do clustering social network graphs help in improving advertising targeting?                         | (3) |
| 10 | What is the "Matching Problem" and how does it impact advertising effectiveness?                         | (3) |

**PART B***Answer any one full question from each module, each carries 14 marks.***Module I**

- |    |  |     |
|----|--|-----|
| 11 | a) Discuss the role of machine learning in data mining and give an example where machine learning is beneficial and another where it is not.       | (7) |
|    | b) Summarize the computational approaches to modeling in data mining and discuss how they differ from statistical and machine learning approaches. | (7) |

**OR**



- 12 a) Discuss the concept of power laws in data mining and provide an example of a phenomenon that follows a power law distribution. (7)
- b) Explain the significance of secondary storage considerations in data mining. How does it affect the design of algorithms? (7)

### Module II

- 13 a) Design MapReduce algorithms to take a very large file of integers and produce as output: (7)
- (a) The largest integer.
  - (b) The average of all the integers.
  - (c) The same set of integers, but with each integer appearing only once.
  - (d) The count of the number of distinct integers in the input.
- b) How does the MapReduce framework streamline the computation of Selections for efficient data processing? (7)

### OR

- 14 a) How do relational algebra operations, including selection, projection, union, intersection, difference, natural join, and grouping-and-aggregation, contribute to the field of data mining? (7)
- b) Explain the MapReduce computation schematic, illustrating the essential components and their interactions. (7)

### Module III

- 15 a) Explain The Flajolet-Martin Algorithm. (7)
- b) What is the primary function of a Data-Stream-Management System, and why is it significant in the context of managing and processing data streams? (7)

### OR

- 16 a) How does the Datar-Gionis-Indyk-Motwani algorithm facilitate query answering in the context of data stream processing? (7)
- b) How do extensions to the counting of ones in a data stream enhance the basic counting operation? (7)

### Module IV

- 17 a) Explain the process of initializing buckets in the BDMO Algorithm. How often are new buckets created, and what determines the size of each bucket? (7)



- b) Perform a hierarchical clustering of the following one-dimensional set of points: (7)  
2, 5, 11, 15, 22, 30, 35, 42, and 50. Assume clusters are represented by their centroid (average), and at each step, merge the clusters with the closest centroids.

**OR**

- 18 a) In the stream model, what challenges arise when the statistics of the stream elements vary with time? (7)  
b) Explain the point assignment algorithm with an example. (7)

**Module V**

- 19 a) How does the concept of "degrees" in social network graphs contribute to our understanding of node centrality and influence within the network? (7)  
b) Illustrate the matches and perfect matches in the matching problem. (7)

**OR**

- 20 a) Explain the Adwords Problem. (7)  
b) Explain social networks as graphs with examples. (7)

\*\*\*\*