(93)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

M.Tech. (DS)-II SEMESTER | T2 EXAM - MARCH 2026

COURSE CODE (CREDITS): 22M11CI213 (3) MAX. MARKS: 25

COURSE NAME: Big Data Analytics (BDA) MAX. TIME: 90 Min.

COURSE INSTRUCTORS: D. Gupta

***Note:** 1) All questions are compulsory. Marks and COs for each question are indicated. 2) **Answer the questions in the given order.** 3) **Be concise and write neatly.** 4) Use of a calculator is allowed.*

Q. No.	Question	CO	Marks
Q. 1	Given a 1D dataset: {2, 4, 10, 12}, K = 2, initial centroids = 2 and 10 a) Perform one complete iteration of K-Means (assignment + update). b) Perform the next iteration and check for convergence. c) Explain whether the final clustering depends on initial centroid selection.	5	4
Q. 2	Given vectors: X = (1,2,3), Y = (2,4,6) a) Compute cosine similarity. b) Explain why magnitude does not affect cosine similarity. c) In which applications is cosine similarity preferred over Euclidean distance?	4	4
Q. 3	Data stream mining operates under strict constraints such as single-pass processing, limited memory, and approximate query responses. a) Explain the data stream model and its key constraints with suitable examples. b) Compare Bloom Filter and Count-Min Sketch in terms of: • functionality • error guarantees • memory efficiency	3	4
Q. 4	a) Explain the curse of dimensionality and how it affects distance-based similarity measures? b) Why does the PCY algorithm use a bitmap instead of storing full hash bucket counts after Pass 1? c) Why is a fading function necessary in data stream clustering? Explain how it helps in handling concept drift. d) Why does the Apriori algorithm require multiple database scans, and how does this impact its performance on large-scale datasets?	4, 5	[2*4 = 8]
Q. 5	a) What is a dendrogram? How is it used to determine the number of clusters? b) Explain the Elbow Method for determining optimal number of clusters. c) Why is selecting the correct value of K important in clustering? Propose a strategy to select K in real-world datasets.	5	[2+1 +2 =5]