Register No : _____

THE KAVERY ENGINEERING COLLEGE

(An Autonomous Institution)

UG/PG DEGREE END SEMESTER THEORY EXAMINATIONS, APRIL-MAY 2026

Second Semester

Computer Science and Engineering

PCSE2416-BIG DATA MINING AND ANALYTICS

Regulations - 2024

Duration : 3 Hrs ***Maximum : 100 Marks***

PART – A(ANSWER ALL QUESTIONS) (Marks 10*2=20)

Q.No	*Questions*	*BLT*	*CO*	*Marks*
1.	What is the conceptual difference between statistical machine learning and statistics for machine learning?	L2	1	2
2.	List and brief the three main types of statistical models.	L1	1	2
3.	Explain the working principle of the K-Nearest Neighbor (K-NN) Algorithm and how it determines the class of a given test data point.	L1	2	2
4.	State the difference between character-level shingling and word-level shingling**.** In what scenarios would one be preferred over the other?	L1	2	2
5.	What are the key challenges associated with processing streaming data?	L1	3	2
6.	What mathematical functions (such as exponential decay functions) are commonly used for implementing Decaying Windows	L1	3	2
7.	Indicate the role of the damping factor (d) in PageRank and its impact on convergence and result stability.	L2	4	2
8.	Find the two key strategies used to detect and mitigate Link Spam in large-scale web data processing.	L4	4	2
9.	What is clustering in data mining, and how is it different from classification?	L2	5	2
10.	What is hierarchical clustering, and name its two main types?	L2	5	2

PART – B (ANSWER ALL QUESTIONS) (Marks 5*16 = 80)

Q.No		*Questions*	*BLT*	*CO*	*Marks*
11.	a	Explain the concept of Feature Extraction in detail .Describe the various techniques used for feature extraction with suitable examples.	L2	1	16
		Or			
	b	What is a Distributed File System (DFS)? Explain its architecture, key components, and working principles.	L2	1	16
12.	a	You are given the following dataset of 5 two-dimensional points representing user locations in a recommendation system: • A (2, 3) • B (5, 4) • C (1, 1) • D (6, 7) • E (3, 5) Answer the below question:	K6	2	16

a) Calculate the Euclidean distance between all pairs of points.
b) Create a distance matrix showing the distances between each pair.
c) Identify the two most similar and the two most dissimilar points.

Or

b	i	Show that the minhash function family is (d1, d2, 1 − d1, 1 − d2)-sensitive	K6	2	8
	ii	Given two documents D1 = {“His brow trembled”} and D2 = {“The brown emblem”}, compute all k-shingles for each document with k = 4.	K6	2	8
13. a	i	What is stream filtering in the context of data stream processing? Explain its importance with respect to real-time data. Describe various techniques used for filtering data streams, such as Bloom filters and sampling.	K2	3	16

Or

b	i	Describe the working of the Flajolet-Martin (FM) algorithm for estimating the number of distinct elements. Include: • The use of hash functions • The concept of trailing zeroes • How multiple hash functions improve accuracy	K2	3	10
	ii	Explain the problem of counting distinct elements in a data stream. Why can’t we use traditional set-based methods in high-throughput environments?	K2	3	6
14. a	i	You are given the following transactional dataset containing itemsets purchased by customers in a retail store:	K3	4	16

Transaction ID	Items Purchased
T1	A,B,C
T2	A,C
T3	A,D
T4	B,E
T5	A,B,C,E
T6	B,C

a) Using the A-Priori algorithm, find all the frequent itemsets with a minimum support threshold of 50%. Show all intermediate steps including candidate generation, pruning, and support count calculations. (12 marks)

b) From the frequent itemsets, generate all possible strong association rules with a minimum confidence threshold of 70%. (4 marks)

Or

b	i	You are given a stream of item IDs representing items viewed by users on an e-commerce platform: A, B, C, A, B, D, A, E, A, B, C, A, B, B, D, C, B Assume the system has limited memory and can only store a maximum of 3 counters at any time. Use the Misra-Gries (Limited Pass) Algorithm (also known as the Frequent Elements Algorithm) to: a) Identify the elements that may be frequent (i.e., occurring more than ⌊n/k⌋ times, where k = 4). Show each step of the algorithm, including counter insertions, decrements, and final count estimates. (12 marks) b) Based on the final output, list all elements that could be frequent and verify which of them actually meet the frequency condition. (4 marks)	K3	4	16

15. a i For the given data, compute two clusters using K-means algorithm for clustering where initial cluster centers are (1.0, 1.0) and (5.0, 7.0). Execute for two iterations. K3 5 16

Record Number	A	B
R1	1.0	1.0
R2	1.5	2.0
R3	3.0	4.0
R4	5.0	7.0
R5	3.5	5.0
R6	4.5	5.0
R7	3.5	4.5

Table 15.1

Or

b i You are given a distance matrix that shows the pairwise Euclidean distances between five data points: **a, b, c, d, and e**. K3 5 16

Using the agglomerative hierarchical clustering method with single linkage (i.e., the distance between two clusters is the minimum distance between any two points in the clusters), perform the clustering in four steps until all data points are merged into one cluster.

For each step:

- Identify which two clusters are merged
- Explain why they are merged (based on the smallest distance)

Use the following distance matrix:

$$\begin{pmatrix} a & b & c & d & e \\ 0 & 4 & 3 & 6 & 11 \\ 4 & 0 & 5 & 7 & 10 \\ 3 & 5 & 0 & 9 & 2 \\ 6 & 7 & 9 & 0 & 13 \\ 11 & 10 & 2 & 13 & 0 \end{pmatrix}$$

******