Enrolment No./Seat No_____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE- SEMESTER–VII EXAMINATION – WINTER 2025

Subject Code:3174302 **Date:15-11-2025**

Subject Name:Mining of Massive Datasets

Time:10:30 AM TO 01:00 PM **Total Marks:70**

Instructions:

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**
4. **Simple and non-programmable scientific calculators are allowed.**

Q.1	(a)	Define Data Mining. Explain its significance in real world scenario.	03
	(b)	Differentiate between data mining and statistical modeling.	04
	(c)	Describe the steps involved in the data mining process.	07
Q.2	(a)	Explain roles and responsibilities of Name node.	03
	(b)	Discuss how physical organization of compute nodes influences the performance of a distributed file system.	04
	(c)	Apply the MapReduce framework to solve the word count problem for a large dataset. Explain each phase in detail.	07
		OR	
	(c)	Summarize how the MapReduce algorithm handles matrix-vector multiplication	07
Q.3	(a)	Describe how Jaccard similarity is used for similarity estimation	03
	(b)	Describe the concept of link spam and its impact on search engine algorithms.	04
	(c)	Summarize the concept of Bloom Filters and how they perform membership tests.	07
		OR	
Q.3	(a)	Explain in brief PageRank works in the context of link analysis.	03
	(b)	Explain how MinHashing is applied to estimate similarity between large datasets.	04
	(c)	Explain how sampling and filtering techniques are used in data stream mining	07
Q.4	(a)	Compare hierarchical clustering and K-means clustering	03
	(b)	Describe how the CURE algorithm handles outliers in clustering large datasets.	04
	(c)	Explain how the A-Priori algorithm identifies frequent item sets in a dataset. How does the algorithm use the concept of "support" to prune the search space?	07
		OR	

Q.4	(a)	What is the significance of Market Basket analysis?	03
	(b)	Compare K-means and CURE Clustering Algorithms.	04
	(c)	Explain the challenges of applying traditional clustering algorithms like K-means in non-Euclidean spaces	07
Q.5	(a)	Explain how clustering can be applied in non-Euclidean spaces	03
	(b)	How would you apply dimensionality reduction techniques to improve the scalability and accuracy of the recommendation engine for large-scale data?	04
	(c)	Describe PageRank algorithm to rank web pages in a large-scale network. How would you modify the algorithm to detect the effects of link spam in the web graph?	07

OR

Q.5	(a)	What is the significance of social-network graph analysis technique?	03
	(b)	Give real-time scenario where recommendation system excels.	04
	(c)	Apply a moment estimation technique to track the frequency of items in a data stream. How would you implement a method to estimate higher-order moments?	07

****