AR 24 Reg. No | | | | | | | | | | |

GANDHI INSTITUTE OF ENGINEERING AND TECHNOLOGY UNIVERSITY, ODISHA, GUNUPUR

(GIET UNIVERSITY)

Ph.D. (First Semester) Examinations, December – 2024

23SPPECS1014– Data Mining and Data Warehousing

(CSE/CSA)

Time: 3 hrs Maximum: 70 Marks

The figures in the right hand margin indicate marks.

	Answer *ANY FIVE* Questions. (14 x 5 = 70 Marks)	**Marks**
1.a.	Define OLAP and OLTP and highlight their differences.	8
b.	Explain the key components of data warehouse architecture.	6
2.	What are partitioning methods in cluster analysis? Give an example of an algorithm.	14
3.a.	Explain the difference between agglomerative and divisive hierarchical clustering? With an example?.	7
b.	Explain how temporal-based frequent patterns differ from traditional frequent patterns.	7
4.	A time series T has seasonal components with periods P=12. If the values for the last 24 months are: [10,15,12,14,16,18,17,20,22,24,23,25,13,18,14,16,18,20,19,22,24,26,25,27]: Decompose the time series into its trend and seasonal components using the average method. Predict the next 6 months' values assuming no significant change in trend or seasonality.	14
5.a.	Consider the sales data for a retail store recorded weekly for 20 weeks: [100,150,120,140,160,200,180,220,240,260,200,150,120,140,160,200,180,220,240,260]. Use autocorrelation to determine the period of repeating patterns. How would you use this information to forecast the next 5 weeks of sales?	7
b.	A social network graph with 100 nodes has an average degree of 5. If 20% of the nodes form a dense community, what is the expected number of edges in this community if its average degree is 10?	7
6.a.	Explain how web documents can be classified using SVM (Support Vector Machines), and discuss its advantages over rule-based classification methods.	7
b.	A distributed algorithm for frequent pattern mining generates partial results: Node 1: {A: 30, B: 50, C: 40} Node 2: {A: 20, B: 30, C: 10} Node 3: {A: 50, B: 20, C: 30} Combine the results to compute the global frequency of each pattern.	7
7.	Explain the concept of HITS (Hyperlink-Induced Topic Search) and compute the authority and hub scores for the following graph: A→B,B→C,C→D,D→A.	14
8.a.	Explore how real-time data warehousing is being utilized for streaming data applications.	7
b.	Explain how ensemble methods like Random Forest and XGBoost handle imbalanced datasets.	7

---End of Paper---