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

GANDHI INSTITUTE OF ENGINEERING AND TECHNOLOGY UNIVERSITY, ODISHA, GUNUPUR

(GIET UNIVERSITY)

Ph.D. (First Semester) Examinations, June – 2025

23WPPECS1014 - Data Mining and Data Warehouse

(CSE)

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.	How does the Apriori algorithm work for frequent pattern mining?	8
b.	What is data mining, and how does it differ from machine learning?	6
2.	What is the difference between supervised and unsupervised learning in the context of classification?	14
3.a.	Compare the K-Nearest Neighbors (KNN) algorithm with Support Vector Machines (SVM) for classification?	7
b.	How does the GSP (Generalized Sequential Pattern) algorithm work for temporal pattern mining?	7
4.	Given a time-series dataset *T*=[5,7,6,8,10,12,11,13,14,16], Calculate the moving average for a window size of 3, Identify and remove noise using a smoothing technique?	14
5.a.	A sequence of temperature measurements recorded hourly is given as: [30,32,31,30,29,30,32,31,30,29,30,32,31,30]. Identify the periodicity in the sequence, Use Fourier Transform to confirm the periodic pattern?	7
b.	In a dataset streamed over time, there are 10,000 records with the following distribution: Class 1: 9500 records, Class 2: 500 records. Apply SMOTE (Synthetic Minority Oversampling Technique) to balance the classes. How many new records for Class 2 are required? Discuss how an incremental learning algorithm can address this imbalance in a streaming context.	7
6.a.	Explain the three main categories of web mining: web content mining, web structure mining, and web usage mining.	7
b.	Consider a set of videos tagged with keywords: Video 1: {sports, basketball} Video 2: {sports, tennis} Video 3: {music, classical} Group the videos into clusters based on their tags using Jaccard similarity.	7
7.	A set of images on a webpage contains metadata (e.g., file name, size, tags, and captions). How would you extract and categorize the images based on their tags using a web scraper?	14
8.a.	Explain the role of cloud-based distributed data warehouses like Amazon Redshift and Google BigQuery in modern analytics?	7
b.	A dataset has 95% of samples belonging to Class A and 5% to Class B. Explain how undersampling and oversampling can be applied to balance the dataset?	7

---End of Paper---