Subject Code:24CT11RC19 **R-24** **Reg No:** | | | | | | | | | | | |

GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN
(Autonomous)
(Affiliated to Andhra University, Visakhapatnam)
II B.Tech. - II Semester Regular Examinations, April – 2026
DATA WAREHOUSING & DATA MINING
[CSE (AI&ML)]

1. All questions carry equal marks
2. Must answer all parts of the question at one place

Time: 3Hrs. **Max Marks: 70**

UNIT-I

1. a. Distinguish between Knowledge Discovery in Databases (KDD) and Data Mining. Illustrate the KDD process with a neat diagram and explain each step briefly. [7M]
   b. Explain the various applications of Data Mining in the fields of healthcare, retail, and finance. Also discuss the major issues in Data Mining. [7M]

OR

2. a. Explain the different types of Data Attributes — Nominal, Ordinal, Interval, and Ratio — with examples. How does attribute type influence the choice of statistical measure? [7M]
   b. Compute the mean, median, mode, variance, and standard deviation for the following dataset: {4, 8, 15, 16, 23, 23, 42, 4, 8, 15}. Interpret the data distribution from these statistics. [7M]

UNIT-II

3. a. Explain the ETL (Extract, Transform, Load) process in Data Warehousing. Discuss common challenges encountered during each phase and how they are addressed. [7M]
   b. Describe Data Transformation techniques: normalization, aggregation, generalization, and attribute construction. Explain Min-Max normalization and Z-score normalization with examples. [7M]

OR

4. a. Explain the concept of Data Cube and describe the OLAP operations — Roll-up, Drill-down, Slice, Dice, and Pivot — with a real-world example for each. [7M]
   b. Illustrate with examples how Attribute-Oriented Induction (AOI) is used for data generalization. Differentiate AOI-based summarization from Data Cube-based OLAP analysis. [7M]

UNIT-III

5. a. Using the following training dataset, construct a Decision Tree using the ID3 algorithm. Compute Information Gain for each attribute at the root and identify which attribute is selected as the root node. [7M]

Training Dataset

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

   b. Discuss the concepts of Overfitting and Tree Pruning. Compare the following pruning strategies and explain when each should be preferred: [7M]
      (i) Pre-Pruning (Early Stopping)
      (ii) Post-Pruning (Reduced Error Pruning)
      (iii) Cost-Complexity Pruning

OR

6. a. Explain about Naive Bayes Classifier with suitable example? [7M]
   b. Describe the following model evaluation and selection techniques. State which is preferred for small datasets and why: [7M]
      (i) Holdout Method
      (ii) K-Fold Cross-Validation
      (iii) Bootstrap Sampling
      (iv) ROC Curve and Area Under the Curve (AUC)

UNIT-IV

7. a. Explain the Apriori Principle and describe how it is used to prune the candidate itemset search space. Discuss the limitations of the Apriori algorithm in terms of time complexity, number of database scans, and memory consumption. [7M]
   b. Define and differentiate between Maximal Frequent Itemsets and Closed Frequent Itemsets with illustrated examples. Explain why closed itemsets provide a compact representation without losing information. [7M]

OR

8. a. Construct an FP-Tree for the transaction database given below (min. support = 50%). Mine all frequent itemsets from the FP-Tree showing each step clearly. [7M]

Transaction Dataset

Transaction ID	Items Purchased
T1	Coffee, Bread, Biscuits
T2	Coffee, Tea, Biscuits, Cake
T3	Tea, Biscuits, Juice
T4	Coffee, Tea, Bread, Biscuits
T5	Coffee, Cake, Juice
T6	Tea, Bread, Biscuits, Cake

   b. Discuss Rule Generation in Association Analysis. Explain the role of the Anti-Monotone property and Confidence-Based Pruning in efficiently generating strong association rules from a list of frequent itemsets. [7M]

UNIT-V

9. a. Explain the concept of Cluster Validity. Describe the internal and external cluster evaluation criteria. Discuss the Silhouette Coefficient as a measure of clustering quality with a suitable example. [7M]
   b. Apply the Agglomerative Hierarchical Clustering (AGNES) algorithm to the dataset below using Single Linkage with Euclidean distance. Show all merge steps and draw the dendrogram. Identify two final clusters. [7M]

Dataset

Point	X	Y
A	1	2
B	2	1
C	4	3
D	5	4
E	3	5

OR

10. a. Analyse the DBSCAN algorithm. Explain Core Points, Border Points, and Noise Points with neat diagrams. Discuss how the parameters epsilon (ε) and MinPts affect the clustering outcome. [7M]
   b. Discuss the following additional issues in K-means clustering: sensitivity to initial centroid selection, handling of outliers, and performance on non-spherical or varying-density clusters. Propose and explain a variant of K-means that resolves at least two of these issues. [7M]