

**Gandhi Institute of Engineering and Technology University, Odisha, Gunupur
(GIET University)**



M. Tech (Second Semester - Regular) Examinations, May – 2026

**24MCSPC12001 – Data Preparation and Analysis
CSE**

Time: 3 hrs

Maximum: 60 Marks

Answer ALL questions

(The figures in the right hand margin indicate marks)

PART – A**(2 x 5 = 10 Marks)**Q.1. Answer **ALL** questions

- | | CO # | Blooms
Level |
|---|------|-----------------|
| a. What is a Confusion Matrix? List its components | CO1 | K1 |
| b. What is a ROC Curve? What does it represent? Give an example. | CO2 | K1 |
| c. Define Missing Data. Elaborate the terms MAR and MNAR. | CO3 | K2 |
| d. What is Clustering? State its purpose. | CO4 | K1 |
| e. Define Data Preparation. List any four steps involved in data preparation. | CO5 | K1 |

PART – B**(10 x 5 = 50 Marks)**Answer **ALL** the questions

- | | Marks | CO # | Blooms
Level |
|---|-------|------|-----------------|
| 2. a. Explain performance evaluation of classifiers using Confusion Matrix and ROC Curve with diagrams. | 10 | CO1 | K2 |
| (OR) | | | |
| b. A multinational company combines employee data from different countries for workforce analysis. The dataset contains different date formats, currencies, and languages. Explain the role of data integration and transformation. | 10 | CO1 | K2 |
| 3.a. Describe methods for handling missing data (MAR, MNAR) with suitable examples. | 5 | CO2 | K2 |
| b. (i) Use decimal scaling normalization for: {9,85,760,42} | | | |
| (ii) Normalize value 150 using Min-Max normalization from: {50,100,150,200,250} to the range [0,1]. | 5 | CO2 | K3 |
| (OR) | | | |
| c. Discuss different types of Data Transformation methods with examples. | 5 | CO2 | K2 |
| d. Consider the transaction dataset: | | | |

TID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Butter
T5	Milk, Bread, Butter

Find the support and confidence for the rule: Milk→Bread

- | | | | |
|---|---|-----|----|
| 4.a. What are the key principles of designing visualizations? Discuss in detail. | 5 | CO3 | K2 |
| b. A classifier predicts all samples as Positive for a dataset containing: 90 Positive instances and 10 Negative instances Construct the confusion matrix and | 5 | CO3 | K3 |

calculate: (i) Accuracy (ii) Precision (iii) Recall (iv) F1-score

(OR)

- c. Explain Support Vector Machines (SVM). Discuss its working and advantages. 5 CO3 K2
- d. Given $X = [2, 4, 6, 8]$ $Y = [5, 10, 15, 20]$. Calculate Pearson correlation coefficient 5 CO3 K3
- 5.a. Consider a dataset with: 8 Positive samples and 8 Negative samples. After splitting by attribute A: Left node: 6 Positive, 2 Negative
Right node: 2 Positive, 6 Negative. Find the information gain. 5 CO4 K3
- b. An e-commerce company collected customer purchase data from different countries. The dataset contains missing values, duplicate records, and inconsistent date formats. Explain how data preparation techniques can be used before analysis. 5 CO5 K2

(OR)

- c. A bank dataset contains:

Income	Credit Score	Loan Approved
High	Good	Yes
Medium	Good	Yes
Low	Poor	No
Medium	Poor	No
High	Good	Yes

10 CO5 K3

Apply Naive Bayes classifier for: (Income=Medium, Credit Score=Good)

- 6.a. Explain Linear Regression and Multiple Linear Regression. Derive the Ordinary Least Squares (OLS) method. 5 CO6 K2
- b. Classifier A: TPR = 0.9, FPR = 0.3
Classifier B: TPR = 0.8, FPR = 0.1
Which classifier is better? Justify 5 CO6 K3

(OR)

- c. Write a short note on
(i) Batch Gradient 5 CO6 K2
(ii) Stochastic Gradient
- d. Describe the steps involved in designing an effective information visualization. 5 CO6 K3

--- End of Paper ---