

--	--	--	--	--	--	--	--	--	--

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

Ph.D. (Second Semester-Summer) Examinations, April – 2026

23SPPECS2012 – Data Preparation and Analysis
(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. | List out various scalability and real time issues of data gathering and preparation methods during data pre-processing techniques. | 8 |
| b. | Write the functionality of structured, unstructured and semi-structured data format. | 6 |
| 2. a. | What is data pre-processing? Explain the various methods used in the data cleaning phase with suitable examples. | 7 |
| b. | Explain Big Data Analytics briefly. Discuss the important design decisions involved in choosing appropriate file formats for Big Data processing. | 7 |
| 3.a. | What is consistency checking in data pre-processing? Explain its role and importance after performing data cleaning techniques. | 7 |
| b. | What is data transformation? Explain the various techniques of data transformation used during data pre-processing with suitable examples. | 7 |
| 4. | Explain the differences between MCAR, MAR, and MNAR missing data mechanisms. Illustrate each with examples and implement appropriate techniques in Python or R to handle missing data. | 14 |
| 5.a. | Discuss the suitable data cleaning strategies required prior to clustering. How can clustering be applied to enhance the performance and accuracy of a linear regression model? | 7 |
| b. | Explain the novel methods used for multivariate data projection and structure visualization. Discuss their significance in analysing high-dimensional data with suitable examples. | 7 |
| 6.a. | Describe the step-by-step process of performing data visualization using Python or R, illustrating with a sample dataset. | 7 |
| b. | Apply the K-means clustering algorithm on the given dataset for K = 2. Use initial cluster centers C1 (2, 4) and C2 (6, 3). Perform the necessary iterations and show the final clusters. Data points: a (2, 4), b (3, 3), c (5, 5), d (6, 3), e (4, 3). | 7 |
| 7. | Write a short notes of the followings : | 14 |
| | i) LDA ii) PCA iii) Gradient Descent | |
| 8.a. | Draw and explain the architecture of a Convolutional Neural Network. | 7 |
| b. | Define Machine Learning. Explain the different perspectives of Machine Learning and discuss the key issues and challenges involved in its implementation. | 7 |

---End of Paper---