BITS Digital
First Semester 2025-2026

Comprehensive Test

Course No. :
Course Title : Data Preprocessing
Nature of Exam : Closed Book (No Internet)
Weightage : 40%
Duration : 2.5 Hours
Date of Exam :

No. of Pages = 1
No. of Questions = 7

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. Read each question carefully and write to-the-point answer.
3. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
4. Assumptions made if any, should be stated clearly at the beginning of your answer.
5. Show all the calculations/derivations in fair and **box/highlight the final answer.**

Q.1.1 Explain the difference between “Interval attribute” and “Ratio attribute”. Provide one example for each. [5Marks]

Q.1.2 Explain the difference between "stratified sampling" and "cluster sampling" with an example for each. [5Marks]

Q.2.1 Which one should be handled first in data mining: [5 Marks]

a. remove noise and then outliers,
b. remove outliers and then the noise,

Justify your answer.

Q.2.2 Assume six students have obtained the following marks: [5Marks]
10, 15, 20, 25, 30

a. Compute the z-score for each student using z-score normalization.
b. Identify which students have scores within ±0.5 standard deviations of the mean.

[5Marks]

Q.3.1 Forward Selection (Attribute Subselection Method) is a lossy or lossless reduction technique? Justify the answer. [03Marks]

Q.3.2 Discuss the mathematical similarity and dissimilarity between the Simple Matching Coefficient (SMC) and the Jaccard index. [07Marks]

Q.4 Given the set {8, 15, 49, 3, 24, 2, 36, 11, 1, 42, 4, 50}, perform data transformation using:
a) Mean Binning with total of 3 bins
b) Boundary Binning with total of 3 bins
c) Min-Max Normalization
d) Decimal Scaling [10Marks]
