WEST BENGAL STATE UNIVERSITY

B.Sc. Honours 5th Semester Examination, 2022-23

CMSADSE02T-COMPUTER SCIENCE (DSE1/2)

DATA MINING

Time Allotted: 2 Hours | Full Marks: 40

The figures in the margin indicate marks of question.
Candidates should answer in their own words and adhere to the word limit as practicable.
All symbols are of usual significance.

GROUP-A

1. Answer any ***four*** from the following: 2×4 = 8

(a) What is the need of data warehousing?
(b) What is market basket analysis?
(c) Define support and confidence in Association Rule Mining.
(d) Write the steps in data preprocessing.
(e) What is meant by Outlier? How Outliers are detected using Data Mining?
(f) Differentiate between Clustering and Classification.

GROUP-B

Answer any *four* from the following 8×4 = 32

2. Consider the following set of frequent 3-itemsets:
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}.
(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.
(b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

3. Describe the various phases in knowledge discovery process with a neat diagram.

4. (a) What are key issues in hierarchical clustering? Explain.
(b) Explain about the basic Agglomerative Hierarchical clustering algorithm.

5. Suppose that 1000 people attended a disease prediction test. Among 300 patients having heart related disorders, 280 of them tested positive, 20 tested negative. Among the 700 people, without having any heart diseases, 685 tested negative and 15 tested positive. Find accuracy, precision, recall and specificity.

6. (a) What are different distance measure?
   (b) Explain partition algorithm.
   (c) Describe the working principle of the decision tree generation algorithm.

7. Write short notes: (any *two*)
   (a) Cross validation
   (b) Confusion matrix
   (c) Type I and Type II error
   (d) Three-tier architecture in data warehousing.

——×——