

Code: R7421504

R7**B.Tech IV Year II Semester (R07) Supplementary Examinations March/April 2013****DATA WAREHOUSING & DATA MINING**
(Common to CSSE & ECC)

Time: 3 hours

Max Marks: 80

Answer any FIVE questions
All questions carry equal marks

- 1 (a) Briefly discuss about data integration.
(b) Briefly discuss about data transformation.
- 2 Compare the advantages and disadvantages of eager classification versus lazy classification.
- 3 Write the FP-growth algorithm for discovering frequent item sets without candidate generation. Explain an example.
- 4 Explain the following:
 - (a) Spatial association analysis.
 - (b) Sequential pattern mining.
 - (c) Latent semantic indexing.
 - (d) Term frequency matrix.
- 5 Explain the following terms in detail:
 - (a) Concept description.
 - (b) Variance and standard deviation.
 - (c) Mean, median, and mode.
 - (d) Quartiles, outliers and box plots.

Contd. in Page 2

R7

Code: R7421504

- 6 Describe each of the following clustering algorithms in terms of the following criteria:
(i) Shapes of clusters that can be determined;
(ii) input parameters that must be specified; and (iii) limitations.
- (a) k-means
 - (b) k-medoids
 - (c) CLARA.
 - (d) BIRCH.
 - (e) ROCK.
 - (f) CHAMELEON.
 - (g) DBSCAN.
- 7 Write a short note on following:
- (a) Missing values.
 - (b) Noisy data.
 - (c) Inconsistent data.
 - (d) Data cube aggregation.
- 8 (a) Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8):
(i) Compute the Euclidean distance between the two objects.
(ii) Compute the Manhattan distance between the two objects.
(iii) Compute the Minkowski distance between the two objects, using $q=3$.
(b) Explain about statistical-based outlier detection and deviation-based outlier detection.
