Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (COMPUTER SCIENCE & ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL/MAY 2017

SUBJECT: DATA WAREHOUSING AND DATA MINING [CSE 4007]

REVISED CREDIT SYSTEM
(29/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL FIVE** questions.
- Missing data may be suitable assumed.

1A. What is Data Mining? Differentiate between Database Processing and Data Mining Processing. Also, give any two queries, each, which could be handled by Database and Data mining respectively. **3M**

1B. Suppose that a hospital tests the age and body fat data for 18 randomly selected adults with the results given in Table 1B:

Table 1B

age	22	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.6	32.9	41.2	35.7

i. Calculate the mean, median and mode of age and % fat
ii. Draw the boxplots for age and %fat. **4M**

1C. Explain the following Data Reduction Techniques:
i. Data Cube Aggregation ii. Attribute Subset Selection iii. Decision Tree Induction **3M**

2A. What are the computational challenges of Apriori algorithm? How it can be improved by modifying the algorithm using hash based technique, transaction reduction and sampling? **3M**

2B. Transactional Data for the Retail shop XYZ is given in the Table 2B:

Table 2B

TID	T1	T2	T3	T4	T5	T6	T7	T8	T9
List of Item IDs	A,B, E,	B,D	B,C	A,B, D	A,C	B, C	A, C	A,B, C, E	A,B, C

Generate frequent patterns using FP-tree(Min. Support = 2(22%)) **4M**

2C. Why we have to augment the support-confidence framework with a pattern interestingness measure? Illustrate any two interestingness measures used. **3M**

3A. Give basic algorithm for inducing decision tree from training tuples. Explain, how the information gain can be used for attribute selection. **4M**

3B. How ROC curves are used in comparing two classification models? Plot the ROC curve to compare two classifier models whose probability for the predicted class of each test tuple is given in the Table 3B:

Table 3B

Classifier C1:

Test Tuple	1	2	3	4	5	6	7	8	9	10
Actual Class	P	N	N	P	P	N	P	P	N	N
Probability for P class	.9	.55	.7	.4	.83	.45	.3	.6	.2	.51

Classifier C2:

Test Tuple	1	2	3	4	5	6	7	8	9	10
Actual Class	P	N	P	P	P	N	N	P	N	N
Probability for P class	.75	.45	.85	.65	.57	.46	.52	.62	.3	.33

3M

3C. How the rules are assessed in a rule based classifier? Illustrate with an example. How the rule based classifier predicts the class label for a given tuple?. **3M**

4A. Given distance matrix in the Table 4A:

Table 4A

Item	**A**	**B**	**C**	**D**	**E**
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Find the corresponding dendogram using
i) single link ii) complete link and iii) average link **3M**

.

4B. Discuss the four cases that are examined for each of the nonrepresentative objects, which will affect the cost function, in k-medoids clustering algorithm, when non-representative object, o_{random} swapped with one of the medoids, o_j,, to determine whether a nonrepresentative object, o_{random}, is a good replacement for a current representative object, o_j. **3M**

4C. Briefly discuss the different intrinsic and extrinsic methods to measure the clustering quality. **4M**

5A. Briefly explain Top-Down and Bottom-Up approach of building Data Warehouse. List the advantage and dis advantages of each approach. Is it possible to adopt a compromise approach in building a data warehouse? Also, give pictorial representation of five different data warehouse architecture types. **4M**

5B. Suppose that a data warehouse for *Big University* consists of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg_grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

(a) Draw a *snow flake schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*student; course; semester; instructor*], what specific *OLAP operations* should one perform in order to list the average grade of *CS* courses for each *Big University* student.

(c) If each dimension has five levels (including all), such as *student* < *major* < *status* < *university* < all", how many cuboids will this cube contain (including the base and apex cuboids)? **3M**

5C. What is Pivot table? How it can be extended to deal with hierarchies. Give an SQL query to create the pivot table on Sales(item_name, color, size, quantity). **3M**