Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

III SEMESTER M.C.A

MAKEUP EXAMINATIONS,

DEC 2017

SUBJECT: DATA WAREHOUSING AND DATA MINING (MCA-5102)

REVISED CREDIT SYSTEM
(/ /2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A.	What is Data Mining? Explain with a neat diagram the architecture of the typical Data Mining System.	**5**
1B.	Suppose that the data for analysis includes the attribute age. The age values of the data tuples are : 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70. (i) Compute the 5 number summary. (ii) Clean the data by finding and eliminating outliers if any. (iii) Draw a box plot for the cleaned data. (iv) Use smoothing by bin means to smooth data using bins of depth size 3.	**3**
1C.	Differentiate between a data mart and a data warehouse.	**2**
2A.	For the following transaction data set, (i) Find all frequent item sets for minimum support of 25% using the Apriori method. (ii) Find all association rules with a minimum confidence of 80 %.	**5**

Transaction Id	I1	I2	I3	I4	I5	I6	I7	I8	I9
T1	1	0	0	0	1	1	0	1	0
T2	0	1	0	1	0	0	0	1	0
T3	0	0	0	1	1	0	1	0	0
T4	0	0	1	0	0	0	0	0	0
T5	0	0	0	0	1	1	1	0	0
T6	0	1	1	1	0	0	0	0	0
T7	0	1	0	0	0	1	1	0	1
T8	0	0	0	0	1	0	0	0	0

2B. Differentiate between subjective and objective measures of pattern interestingness. **3**

2C. What is the need for concept hierarchies? Create a concept hierarchy for the attribute "Location". **2**

3A. The data warehouse for a UNIVERSITY consists of the following 4 dimensions- STUDENT, COURSE, SEMESTER, INSTRUCTOR and 2 measures –COUNT and AVG_GRADE (average grade). **5**

(i) Assume attributes and draw a STAR schema diagram for the UNIVERSITY warehouse.

(ii) What OLAP operations are required to extract the average grade of all students studying MCA course in 2nd semester?

3B. What does the confusion matrix represent? Define the following Classification Accuracy Measures and compute them from the confusion matrix provided below. **3**

(i) Accuracy Rate (ii) Misclassification Rate

(iii) sensitivity (iv) specificity

classes	buy_computer = yes	buy_computer = no	total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
total	7366	2634	10000

3C. What is the need for the Laplacian correction in the Naïve Bayesian classification method? **2**

4A. The following table shows the relationship between the amount of fertilizer used and The height of a plant.

(i) Calculate a simple linear regression equation using Fertilizer as the descriptor and Height as the response.

(ii) Predict the height when fertilizer is 9.5

5

Fertilizer	10	5	12	18	14	7	15	13	6	8	9	11	16	20	17
Height	0.7	0.4	0.8	1.4	1.1	0.6	1.3	1.1	0.6	0.7	0.7	0.9	1.3	1.5	1.3

4B. Differentiate between web content mining and web structure mining. **3**

4C. Differentiate between supervised learning and unsupervised learning techniques. Give examples. **2**

5A. Consider the following distance matrix and perform agglomerative clustering on the 5 data points. Visualize using a dendrogram. **5**

	p1	p2	p3	p4	p5
p1	0	0.10	0.41	0.55	0.35
p2	0.10	0	0.64	0.47	0.98
p3	0.41	0.64	0	0.44	0.85
p4	0.55	0.47	0.44	0	0.76
p5	0.35	0.98	0.85	0.76	0

5B. Given two data points X= (12, 32, 27, 17) and Y= (14, 20 , 46, 8) .

(i). Represent them as a data matrix.

(ii). Represent them as a distance matrix using

1. Euclidean distance between the data points
2. Manhattan distance between the data points.
3. Minkowski distance between the data points using q = 3.

3

5C. How outliers are spotted using the Density based clustering technique? **2**