Reg. No.

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL

III SEMESTER M.C.A

END SEMESTER EXAMINATIONS,

NOV/DEC 2017

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

REVISED CREDIT SYSTEM
(17/11/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A.	Define Data Mining. Explain with a neat diagram, the steps involved in the Knowledge Discovery in Databases process.	5
1B.	Consider a data set with the values 230,375, 428, 705, 1500. Perform Data transformation on each of the above values with: (i) the min-max normalization method by setting min = 0 and max = 5 (ii) the decimal scaling method	3
1C.	What is an ordinal attribute? Give appropriate examples.	2
2A.	The following are the set of 6 transactions representing the purchase of books. Let minimum support be = 2. (i) Find frequent itemsets using FP-growth algorithm or the Apriori algorithm. (ii) Generate all the association rules with a minimum confidence of 80 %.	5

	T1 ANN, CC, TC, CG T2 CC, D, CG T3 ANN, CC, TC, CG T4 ANN, CC, D, CG T5 ANN, CC, D, TC, CG T6 CC,D,TC	
2B.	What are the key features of a Data Warehouse?	3
2C.	What is the need for concept hierarchies? Create a concept hierarchy for the attribute "Order Date".	2
3A.	Details of cars such as, its color, type of car and its origin is recorded. The variable 'stolen' indicates whether the car was stolen or not. Predict the class label "Stolen" for a car with attributes – color – Red, type – Sports and origin – Imported using the Naive Bayesian method.	5

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

3B.	How is the attribute selection measure – "Information Gain" computed for numeric attributes?	3
3C.	How are hyper planes computed in the Support vector machine (SVM) classifier?	2
4A.	Consider a table of the following observations. (i) Draw a scatter plot to illustrate the relationship between x & y. (ii) Use the method of "simple non-linear regression" to predict y value for x = 14.6.	5
4B.	What strategies could be adopted for separation of test and training dataset for classifiers?	3
4C.	List any two measures that quantify the accuracy of prediction algorithms.	2
5A.	Consider the following 8 data points with (x,y) representing location. A1 (2,10) , A2 (2,5) , A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C1 (1,2) & C2 (4,9) (i) Cluster the data using the k-means algorithm, into 3 clusters assuming A1, B1 and C1 to be initial cluster centres. (ii) Visualize the final clusters (last iteration only).	5
5B.	List the different types of data that are utilized by various web mining algorithms.	3
5C.	Differentiate between global outliers and contextual outliers using appropriate examples.	2

x	3	6	9	8	10	11	12	13	13.5	14	14.5	15	15.2	15.3
y	4	5	7	6	8	10	12	14	16	18	22	28	35	42