Student Exam No._____

GANPAT UNIVERSITY

B. Tech. Semester: Sem-VII

(Information Technology/Computer Engineering) Engineering

Regular / Remedial Examination Nov-Dec 2016

2CE703/2IT703 Data Mining & Data Warehousing

Time: 3 Hours **Total Marks:** 70

Que. – 1

(A) Explain Data Mining as a step in the process of knowledge discovery. **[4]**

(B) Discuss in Detail with example: **[4]**

I. ROLLUP OPERATION

II. SLICE AND DICE OPERATION

(C) Given 1-dimensional data set X={-5,0,23.0,17.6,9.23,1.11} normalize the data set using, **[4]**

i)Min-Max Normalization[0,1]

ii) Min-Max Normalization[-1,1]

OR

Que. – 1

(A) What are the major issues in Data Mining? **[4]**

(B) Explain smoothing, Aggregation, Generalization and Normalization with suitable Example. **[4]**

(C) Explain six methods to fill Missing Values in database. **[4]**

Que. – 2

(A) Explain Mining Frequent Item sets Using Vertical Data Format (ECLAT). **[3]**

(B) A database has five transactions. Let min sup = 60% and min conf = 80%. Find all frequent itemsets using apriori and FP-growth, respectively. Compare the efficiency of the two mining processes. **[8]**

TID	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

OR

Que. – 2

(A) Explain Confusion matrix for evaluating performance of classifier accuracy. **[4]**

(B) How rules are extracted from decision tree? **[4]**

(C) Define multilevel association rules, Multidimensional association rules. **[3]**

Que. – 3

(A) Here 2 *2 contingency table with summarizing the transactions **[6]**

with respect to game and video purchases. Find out χ^2, Cosine, All-Conf, Lift.

	game	game	Srow
video	4,000	3,500	7,500
video	2,000	500	2,500
Scol	6,000	4,000	10,000

(B) Find out the gain for age, Income, Student, Credit_Rating and predict the root node. Predict a class label of an unknown tuple X= {age ='<=20', Income= 'Medium', Student='Yes', Credit_Rating='Fair'} using Naïve Bayesian Classification. **[6]**

Age	Income	Student	Credit_rating	Class: Buys Laptop
>30	Medium	No	Excellent	No
<=20	High	No	Fair	No
21..30	High	Yes	Fair	Yes
<=20	High	No	Excellent	No
21..30	Medium	No	Excellent	Yes
21..30	High	No	Fair	Yes
<=20	Medium	Yes	Excellent	Yes
>30	Medium	No	Fair	Yes
>30	Medium	Yes	Fair	Yes
>30	Low	Yes	Fair	Yes
<=20	Low	Yes	Fair	Yes
>30	Low	Yes	Excellent	No
21..30	Low	Yes	Excellent	Yes
<=20	Medium	No	Fair	No

Section –II

Que. – 4

(A) Explain difference between OLAP and OLTP? **[6]**

(B) Explain Stars, Snowflakes, and Fact Constellations Schemas for Multidimensional Databases with Diagram. **[6]**

OR

Que. – 4

(A) Explain the architecture of Data Warehouse. **[6]**

(B) Describe difference between single linkage algorithm and complete linkage algorithm. **[6]**

Que. – 5

(A) Given the samples X1 = {1, 0}, X2 = {0, 1}, X3 = {2, 1}, and X4 = {3, 3}. Suppose that samples are randomly clustered into two clusters C1 = {X1, X3} and C2 = {X2, X4}. **[6]**

a) Apply one iteration of K-means partitional clustering algorithm, and find a new distribution of samples in clusters. What are the new centroids? How you can prove that the new distribution of samples is better than the initial one?
b) What is the change in a total square error?
c) Apply the second iteration of K-means algorithm and discuss the changes in clusters.

(B) Create student.arff file having following details. **[5]**
Attributes→ sub1 sub2 sub3 sub4 sub5
The numeric value of rank should be given to each subject.
The detail is given below. 1-poor, 2-satisfactory, 3-average, 4-good, 5-excellent. Enter at least 10 records

OR

Que. – 5

(A) Suppose that a patient record table contains the attributes name, gender, fever, cough, test-1, test-2, test-3, and test-4, where name is an object identifier, gender is a symmetric attribute, and the remaining attributes are asymmetric binary. Find the Jaccard coeffients. **[6]**

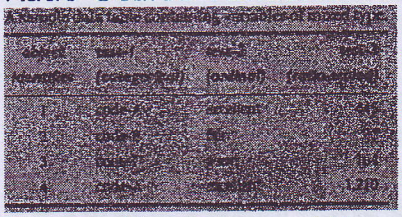
Table A relational table where patients are described by binary attributes.

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(B) Explain decision tree induction with example. **[5]**

Que.- 6

(A) Explain Interval-Scaled Variables, Categorical, Ordinal, and Ratio-Scaled Variables with suitable example. **[6]**



(B) Explain the Balanced Iterative Reducing and Clustering Using Hierarchies. Show how effective is BIRCH? Where, C1= (2, 5), (3, 2), (4, 3) and C2= (5, 2), (2, 3), (3, 4). Show CF1, CF2 and CF3. **[6]**

END OF PAPER