Exam No: _____

GANPAT UNIVERSITY

M. TECH SEMESTER - II

REGULAR EXAMINATION– APRIL-JUNE 2016

3CE203/3IT203: Data Mining & Data Warehousing

TIME: 3 HRS

TOTAL MARKS: 60

Instructions: (1) This Question paper has two sections. Attempt each section in separate answer book.
(2) Figures on right indicate marks.
(3) Be precise and to the point in answering the descriptive questions.

SECTION: I

Q.1 **Answer the following**

A Draw the KDD Process, Data mining Architecture and explain it in detail. (10)

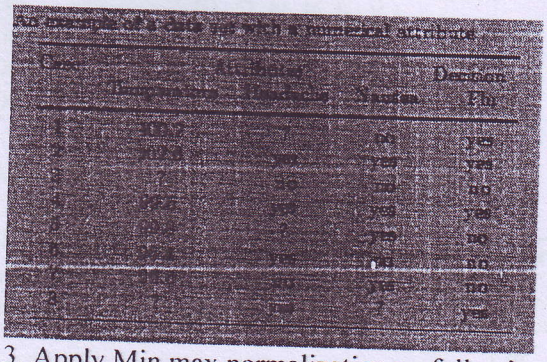
B Explain the major issues in Data Mining.

OR

Q. 1 Why Preprocess the Data and answer the following. (10)

1. Replacing missing attribute values by the attribute mean below table.

2. Replace Missing data with most common value of an attribute.



3. Apply Min max normalization on following table for Age and Salary attribute.

Id	Gender	Age	Salary
1	F	27	19K
2	M	51	64K
3	F	52	100K
4	F	33	55K
5	M	45	45K

4. Convert Categorical Attributes to Numerical Attributes for wind attribute (ref. Q-2 B table)

5. Explain data Discretization.

Q.2 **Answer the following**

A What is Lattice of cuboid? Draw a diagram of lattice of cuboid of starting from 0-D to 4-D for various dimensions. (10)

B Explain Confusion matrix for evaluating performance of classifier accuracy.

OR

Q.2 **Answer the following**

A Suppose that an airline company is making a loss in almost all classes for all routes for the year 2015 which cover a wide range from 12,08,234 (LOSS) to 2,50,231 (LOSS). A user wishes to have a concept hierarchy for loss automatically generate. Suppose that the data within the 5th percentile and 95th percentile are between 11, 55, 400 (LOSS) and 3, 00, 189 (LOSS). Apply 3-4-5 rule up to second level in hierarchy. (5)

B The weather attributes are Outlook, Temperature, Humidity, and Wind Speed. They can have the following values: Outlook = {sunny, overcast, rain}, Temperature = {hot, mild, cool}, Humidity ={high, normal}, Wind = {weak, strong}. Construct decision tree using classification algorithm and Decide on which day you can play tennis. Attribute <play Tennis> has 2 values {yes,no} (5)

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Q.3 Explain Following Type of data in clustering analysis using distance matrix and example: (10)
Interval-scaled variables, Binary variables, Nominal, ordinal, and ratio variables, Variables of mixed types. Use below table for reference.

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

SECTION: II

Q.4 A database has four transactions. Let min sup = 60% and min con f = 80%. Cust_ID TID items bought (in the form of brand-item category) (10)

T100 {King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
T200 {Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
T300 {Westcoast-Apple, Dairyland-Milk,Wonder-Bread, Tasty-Pie}
T400 {Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

OR

Q.4 A The "Popular Kids" dataset also divided the students' responses into "Urban," "Suburban," and "Rural" school areas. Is there an association between the type of school area and the students' choice of good grades, athletic ability, or popularity? Use chi square method and lift method. (5)
A two-way table for student goals and school area appears as follows:

Goals	School Area: Rural	Suburban	Urban	Total
Grades	57	87	24	168
Popular	50	42	06	098
Sports	42	22	05	069
Total	149	151	35	335

B. Explain Hash algorithm for following table. (5)

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

Q.5 **Explain Following clustering methods :** (10)
1. RObust Clustering using links (ROCK),
2. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
   where C1= (2, 5), (3, 2), and (4, 3) and C2 = (2, 3), (3, 1), and (6, 3)
3. Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling

OR

Q.5 Explain Agglomerative Hierarchical Clustering with single-link or complete-link algorithms. Apply agglomerative hierarchical clustering and Plot the Dendograms. ***Find*** similarity between samples using Jaccard's Coefficient. Where objects are: ***A = (1, 0, 1, 1, 0), B = (1, 1, 0, 1, 0), C = (0, 0, 1, 1, 0), D = (0, 1, 0, 1, 0), E = (1, 0, 1, 0, 1), F = (0, 1, 1, 0, 0).*** (10)

Q.6 A Explain Inter transaction Association Rule Mining using following table where w=2. (5)

Date	TCS	SBI	ONGC	Tata Steel	Asian Paints	Titan
31-12-2014	383.85	151.85	1047.75	399.4	752.3	381.65
30-12-2014	382.55	151.65	1044.9	396.3	749.95	377.1
29-12-2014	386.9	154.3	1069.2	404.1	757.55	373.95
26-12-2014	378.35	149.3	1046.3	398.25	727.05	367.25
24-12-2014	376.85	146.15	1035.85	395.25	728.45	368.75
23-12-2014	383.65	145.85	1044.9	394.8	739.25	374.85
22-12-2014	390.35	145.5	1071.9	404.05	745.55	381.95
19-12-2014	376	137	1056.5	405.45	733	376.6
18-12-2014	367.55	138.6	1037.6	401.6	744.95	377.25
17-12-2014	367.25	128.95	1025.7	393.45	731.3	355
16-12-2014	366.75	131.95	1029.85	389.65	758.85	360

B. Explain following term: 1. Web Mining 2. Text Mining 3. Data Mining privacy (5)

----------END OF PAPER----------