B. TECH. (CBCS - 2014 COURSE) SEM - VIII (COMPUTER ENGG.) : SUMMER - 2018

SUBJECT: DATA MINING AND KNOWLEDGE DISCOVERY

Day: **Saturday**
Date: **09/06/2018**

S-2018-4671

Time: **02.30 PM TO 05.30 PM**
Max Marks: 60

N.B.:

1) All questions are **COMPULSORY**.
2) Use of non programmable calculator is **ALLOWED**.
3) Figures to the right indicate **FULL** marks.
4) Assume suitable data, if necessary.

Q.1 What is data cleaning? Describe in detail the different methods for data cleaning. **(10)**

OR

Q.1 How data mining systems can be classified? Describe the different data mining functionalities with example. **(10)**

Q.2 What is OLAP Server? Explain its architecture in detail. Also state the differences between OLTP and OLAP systems. **(10)**

OR

Q.2 What is the importance of multidimensional views of data and data cubes? Why they are used? Explain in detail the various data cube implementations and operations. **(10)**

Q.3 Trace the results of using the Apriori algorithm on the given database samples below with support threshold = 33.34% and confidence threshold = 60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rule that are generated and highlight the strong ones, sort them by confidence. **(10)**

Transaction	Itemsets
T1	Hotdogs, Buns, Ketchup
T2	Hotdogs, Buns
T3	Hotdogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hotdogs, Coke, Chips

OR

Q.3 Illustrate with an example the process of mining frequent patterns without candidate generation. **(10)**

P.T.O

Q.4 What is classification concept in data mining? Justify by giving an example the use of rule based classifiers for classification? **(10)**

OR

Q.4 Consider the training samples shown in the following table for a binary classification problem. **(10)**

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

i) What is the entropy of this collection of training samples with respect to positive class?

ii) What are the information gain of a_1 and a_2 relative to these training examples?

iii) For a_3 which is a continuous attribute, compute the information gain for every possible split.

iv) What is best split (among a_1, a_2 and a_3) according to the information gain?

Q.5 What is the difference between data mining and knowledge discovery? Explain in detail the KDD lifecycle. **(10)**

OR

Q.5 Explain in detail the integration of KDD system with Database/Data warehouse system. **(10)**

Q.6 What is cluster analysis? Illustrate using an example the k-means algorithm. **(10)**

OR

Q.6 Describe in brief using example the following approaches to clustering: **(10)**

i) Model based clustering methods,

ii) Constraint based methods.

* * * * *