Code No: 127CD **R15**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, May/June - 2019
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 Hours **Max. Marks: 75**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 25 marks. Answer all questions in Part A. Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks and may have a, b, c as sub questions.

PART- A

(25 Marks)

1.a)	List and define the characteristics of Data warehouse.	[2]
b)	Give a brief note on Fact-Less-Facts.	[3]
c)	What do you mean by Data Cleaning?	[2]
d)	What are the limitations of data mining?	[3]
e)	Mention the importance of Association Rule Mining.	[2]
f)	Define frequent sets, confidence and support	[3]
g)	What is classification?	[2]
h)	Write the need for tree pruning in decision tree induction?	[3]
i)	Differentiate between clustering and classification.	[2]
j)	How are outliers detected using data mining?	[3]

PART-B

(50 Marks)

2.a) Draw the Data warehouse Architecture and explain its Components.
b) Explain Star and Snow-Flake Schemas. [5+5]

OR

3.a) Give a note on OLAP Operations.
b) What are the differences between the MOLAP and ROLAP models? Also list their similarities. [5+5]

4. Explain the following with examples
a) Aggregation
b) Dimensionality reduction
c) Feature subset selection. [10]

OR

5.a) What steps you would follow to identify a fraud for a credit card company.
b) List and define the measures of Similarity and Dissimilarity. [5+5]

6. A database has four transactions. Let min_sup=60% and min_conf=80%

TID	date	items_bought
100	10/15/99	{K, A, B, D}
200	10/15/99	{D, A, C, E, B}
300	10/19/99	{C, A, B, E}
400	10/22/99	{B, A, D}

a) Find all frequent items using apriori & FP-growth, respectively. Compare the efficiency of the two meaning process.
b) List all of the strong association rules (with support s and confidence c) matching the following metarule where X is a variable representing customers, and item i denotes variables representing items (e.g., "A", "B", etc.): Vx Є transactions, buys (X,item1) ^ buys (X,item2) =>buys(X,item3)[s,c]. [10]

OR

7.a) What is more efficient method for Generalizing association rule? Explain.
b) Describe a data set for which sampling would actually increase the amount of work. In other words it would be faster to work on full data set. [5+5]

8. Construct a decision tree with root node Type from the data in the table below. The first row contains attribute names. Each row after the first represents the values for one data instance. The output attribute is Class. [10]

Scale	Type	Shade	Texture	Class
One	One	Light	Thin	A
Two	One	Light	Thin	A
Two	Two	Light	Thin	B
Two	Two	Dark	Thin	B
Two	One	Dark	Thin	C
One	One	Dark	Thin	C
One	Two	Light	Thin	C

OR

9.a) Explain in detail the Naive-Bayes Classifier.
b) List the characteristics of K- Nearest neighbor classification. [5+5]

10.a) Differentiate Agglomerative and divisive Hierarchical Clustering.
b) Explain Partitioning Clustering-K-Means Algorithm with an example. [5+5]

OR

11.a) Compare the performance of various outlier detection approaches.
b) Give the classification of various clustering methods in data mining. [5+5]

--ooOoo--