USN | | | | | | | | | | |

10IS74

Seventh Semester B.E. Degree Examination, Dec.2018/Jan.2019
Data warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

Note: ***Answer FIVE full questions, selecting at least TWO questions from each part.***

PART – A

1 a. What is metadata? Derive some examples of metadata from everyday situations. **(06 Marks)**

b. What is data scrubbing? On what basis data scrubbing is done? Explain. **(06 Marks)**

c. Why and when would an enterprise implement a separate ODS and a separate data warehouse? Explain. **(08 Marks)**

2 a. Define OLAP? Differentiate between OLTP and OLAP. **(06 Marks)**

b. With an example, explain the relationship between aggregations of a 3D cube. **(06 Marks)**

c. Describe the operations roll-up, drill-down, slice and dice and pivot. **(08 Marks)**

3 a. What is data mining? Explain the process of knowledge discovery in databases. **(06 Marks)**

b. Consider the following tow binary vectors
X = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1) and
Y = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1)
Find the i) Cosine ii) SMC iii) Jaccard coefficient. **(06 Marks)**

c. Explain the various sampling approaches. **(08 Marks)**

4 a. A database has 5 transactions. Let min_sup = 60% and min_conf = 80%

Tid	Items
1	A B C D E F
2	B C D E F G
3	A D E H
4	A D F I J
5	B D E K

Generate all the frequent itemsets and the association rules using apriori algorithm. **(12 Marks)**

b. Explain the various alternative methods for generating frequent itemsets. **(08 Marks)**

PART – B

5 a. Write and explain with an example the algorithm for nearest neighbor classification. **(06 Marks)**

b. Construct a decision tree for a customer data base at car sales shop

ID	Age	Income	Student	Credit Rating	Buy Car
1	Young	High	No	Fair	No
2	Young	High	No	Good	No
3	Middle	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Good	No
7	Middle	Low	Yes	Good	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Good	Yes
12	Middle	Medium	No	Good	Yes
13	Middle	High	Yes	Fair	Yes
14	Old	Medium	No	Good	No

(06 Marks)

c. Explain the various measures for selecting the best splits with an example. **(08 Marks)**

6 a. List 5 criteria for evaluating classification methods. Discuss them briefly. **(06 Marks)**

b. What is Baye's theorem? Show how it is used as the basis of the Naïve Baye's classifier. **(08 Marks)**

c. Describe any 3 methods of estimating the accuracy of a classification method. **(06 Marks)**

7 a. Following five objects, each with two attributes, are to be clustered : A_1 (4, 4), A_2 (8,4), A_3 (15,8), A_4 (24, 4) and A_5 (24, 12).
Find the distance matrix between the objects using Manhattan distance and use the agglomerative method to build hierarchical clusters. **(12 Marks)**

b. Describe the single_link, complete_link, centroid and Ward's algorithm. Which one is used most frequently? Why? **(08 Marks)**

8 Write a note on :

a. Web content mining

b. Text mining

c. Text clustering

d. Mining spatial and Temporal Databases. **(20 Marks)**

* * * * *