Course Code : CAT 307 **KQLR/RS – 24 / 1256**

Sixth Semester B. Tech. (Computer Science and Engineering / Artificial Intelligence and Machine Learning) Examination

DATA MINING AND WAREHOUSING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) Assume suitable data wherever necessary.

(2) All questions carry marks as indicated.

1. (a) The data warehouse for wholesale furniture company has to allow analyzing the company's situation at least with respect to the Furniture, Customers and Time. Moreover, the company needs to analyze :

- ● The furniture with respect to its type (chair, table, wardrobe, cabinet...), category (kitchen, living room, bedroom, bathroom, office...) and material (wood, marble...)
- ● The customers with respect to their spatial location, by considering at least cities, regions and states.

The company is interested in learning at least the quantity, income and discount of its sales :

- ● Identify Facts, Dimensions and Measures.
- ● Draw STAR and Snowflake Schema for the above Scenario.
- ● Write SQL queries for the following :
  - ○ Find the quantity, the total income and discount with respect to each city, type of furniture and the month.
  - ○ Find the average quantity, income and discount with respect to each country, furniture material and year. 6(CO1)

(b) Explain the Datawarehouse architecture and comment on importance of metadata repository. 4(CO1)

2. (a) Differentiate between the following :—

(i) ROLAP and MOLAP.

(ii) Snowflake and fact constellation schema. 4(CO3)

(b) Suppose a student collected the price and weight of 20 products in shop with the following result :—

Price	5·89	149	59·98	129	15·89	56·99	35·75	42·19	31	125·5
Weight	1·4	1·5	2·2	2·7	3·2	3·9	4·1	4·1	4·6	4·8
Price	4·5	22	52·9	61	33	328	122	142·19	229	89·4
Weight	4·9	5·1	5·5	5·8	5·8	8·9	9·6	18·0	36·9	38·2

(i) Give 5 number summary for price.

(ii) Draw boxplot for price. Identify outliers, if any.

(iii) Draw scatter plot based on these two variables.

(iv) Calculate the Pearson correlation coefficient.

(v) Are these two variables positively or negatively correlated ? 6(CO3)

3. (a) Explain the need to create function-based indexes. Write a command to create a function-based index on emp-name column of EMPLOYEE table. 2(CO1)

(b) Explain query optimizer with respect to data warehousing. 3(CO1)

(c) State the advantage of partitioning in data-warehouse.

Write a query to create composite List-Range partitioning for the following scenario :

- Supplier table having attributes sup_id, sup_name, sup_state and time_id.
    - Perform list partitioning on state attributes and range partitioning on time-id.
    - Partition definitions for list are as below :
        - Partition East should accept values ('WB', 'JK').
        - Partition South should accept values ('TN', 'AP').
        - Partition North should accept values ('UP', 'HP').
        - Partition Temp should accept any other state.
    - Partition definitions for range are as below :
        - Partition P1 should accept values less than 01-Jan.-2018.
        - Partition P2 should accept values less than 01-April-2018.
        - Partition P3 should accept values less than 01-July-2018.

Write query to access data from partition and subpartition. 5(CO2)

4. (a) Use the dataset below to learn a decision tree which predicts if people pass Java Test (True or False) based on their previous GPA (High (H), Medium (M) or Low (L)) and whether or not they studied. GPA and Studied are two features. Passed is the target function.

GPA	Studied	Passed ?
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

Construct the decision tree using ID3 algorithm that would be learned for this dataset. 5(CO2)

(b) Discuss in detail various steps in KDD process with suitable diagram. 5(CO3)

5. (a) Generate the frequent itemsets using the FP-growth algorithm for the transaction database shown below and a minimum support s_min = 3.

TId	Items
T1	a, d, e
T2	b, c, d
T3	a, c, e
T4	a, c, d, e
T5	a, e
T6	a, c, d
T7	b, c
T8	a, c, d, e
T9	b, c, e
T10	a, d, e

5(CO3)

(b) Apply Naive Bayesian classifier for the following dataset and find class(x) by executing it in the given training set :
X = (age < 30, Income = Medium, Student = Yes, Credit_rating = Fair)

Age	Income	Student	Credit_rating	Buys_laptop
⩽30	High	No	Fair	No
⩽30	High	No	Excellent	No
31-40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31-40	Low	Yes	Excellent	Yes
⩽30	Medium	No	Fair	No
⩽30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
⩽30	Medium	Yes	Excellent	Yes
31-40	Medium	No	Excellent	Yes
31-40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

5(CO3)

6. (a) Use the k – means algorithm and Euclidean distance to cluster the following 10 examples into 3 clusters :

X1(2, 10) ; X2(2, 5) ; X3(8, 4) ; X4(9, 4) ; Y1(5, 8) ; Y2(7, 5) ; Y3(6, 4) ; Z1(1, 2) ; Z2(4, 9) ; Z3(6, 10).

Suppose that the initial seeds (centers of each cluster) are X1, X4 and Z2.

Run the k – means algorithm for 3 iteration.

At the end of each iteration, show :

(i) The new clusters. (i.e. the examples belonging to each cluster).

(ii) The centers of the new clusters.

(iii) Draw a 10 by 10 space with all the 10 points and show the clusters after each iteration. 6(CO4)

(b) Briefly outline how to compute dissimilarity between objects describe by the following type of variable using examples :

(i) Numerical Variable.

(ii) Binary Variable.

(iii) Categorical Variable. 4(CO4)