Course Code : CST 359-4 **SUVT/MS – 22 / 2515**

Sixth Semester (Computer Science and Engineering) Examination

DATA WAREHOUSING AND MINING

Time : 2 Hours] [Max. Marks : 40

Instructions to Candidates :—

(1) All questions carry marks as indicated against them.

(2) Assume suitable data wherever necessary and clearly state your assumptions.

1. (a) The Restaurants 'SR' wholesale restaurant company supplies equipment to 55 different restaurants in Mumbai, such as tables, chairs, table cloths, napkin holders, cutlery and so on, as well as kitchen equipment such as saucepans, knives and chef clothing. They wish to analyze their daily sales in terms of revenue, unit sales, costs and profit for each product and customer. They also would like to know this information by product line and product group.

(i) Design a STAR schema according to the given scenario.

(ii) Convert STAR schema into Snowflake Schema.

(iii) Bring out the difference between STAR and Snowflake Schema.

6(CO1)

2. (a) The table below gives information regarding two varieties of potatoes. Draw and compare their box plots. Which variety Type A or Type B will you advice to plant in the future and why ?

	Type A	**Type B**
Median	52 Gms.	52 Gms.
Lower Quartile	49 Gms.	51 Gms.
Upper Quartile	57 Gms.	54 Gms.
Range	14 Gms.	8 Gms.
Interquartile Range	8 Gms	3 Gms.

4(CO1)

(b) Compute Pearson's coefficient of correlation between advertisement cost and sales as per the data given below :

Advertisement Cost in 1000's	39	65	62	90	82	75	25	98	36	78
Sales in lakhs	47	53	58	86	62	68	60	91	51	84

Also plot the scatter plot and comment on your observations. 4(CO1)

3. Implement virtual column based partitioning as below :
Create table employee with attributes Emp_id, emp_name. monthly_sal, bonus. Generate Total salary as virtual column. Total salary is addition of bonus and monthly salary. Perform range partitioning on Total Salary with four partitions as below :

- Partition P1 stores salary less than 15000
- Partition P2 stores salary less than 45000
- Partition P3 stores salary less than 65000
- Partition P4 stores any salary above and equal to 65000

Insert some sample rows in the table and explain the output. 6(CO2)

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

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}

(i) Use Apriori and 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.) :
$\forall x \in \text{transaction}, \; buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \; [s, c]$

(ii) Calculate lift and leverage for first three rules generated in (i) Are these strong rules necessarily interesting ? Comment on the basis of your answer. 6(CO3)

5. (a) Apply Naive Bayesian classifier on the following dataset and classify a Red Domestic SUV.

Example No.	**Color**	**Type**	**Origin**	**Stolen ?**
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

7(CO3)

6. Use single, complete, and average link agglomerative clustering to group the data described by the following distance matrix. Produce the dendrograms.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

7(CO4)