TE(ECS) / Sem VI / R-19 / SH-25 / 09-12-25

**Paper / Subject Code: 89324 / Data Warehousing and Mining**

QP CODE: 97246

**Time: 03 Hours** **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of the remaining five questions.
3. Assume any suitable data wherever required and justify the same.

Q1 a) Every data structure in the data warehouse contains the time element. Why? [5]

b) Mention difference between OLTP Vs OLAP with respect to following features: [5]
(i) Characteristic
(ii) Data
(iii) Access
(iv) Number of users
(v) DB size

c) Calculate Accuracy, Recall and Precision with the help of following data: True Positive (TP)= 50, True Negative (TN) = 20, False Positive (FP)= 20, False Negative (FN)= 10 [5]

d) What is Support and Confidence in market basket analysis. Explain with an example. [5]

Q2 a) Information requirements are recorded for "Hotel occupancy" considering dimensions like Hotel, Room and Time. Few Facts recorded are vacant rooms, occupied rooms, number of occupants, etc. [10]
(i) Draw a star schema diagram.
(ii) Can you convert this star schema to a snowflake schema? If yes, justify and draw the snowflake schema.

b) For the given set of points identify clusters using a single linkage algorithm. Draw dendrogram. [10]

| Object | Attribute(X) | Attribute(Y) |
|---|---|---|
| A | 2 | 2 |
| B | 3 | 2 |
| C | 1 | 1 |
| D | 3 | 1 |
| E | 1.5 | 0.5 |

Q3 a) Name the set of basic transformation tasks. Give an example for each. [10]

b) A database has five transactions. Let min sup = 50% and min conf = 60%. [10]

| TID | Items |
|---|---|
| t1 | Milk, Bread, Butter |
| t2 | Bread, Butter, Sugar |
| t3 | Bread, Sugar, Potato |
| t4 | Milk, Bread, Sugar |
| t5 | Milk, Bread, Butter, Potato |
| t6 | Milk, Bread, Butter, Sugar, Potato |

Find all frequent itemsets and strong association rules using Apriori Algorithm.

**97246**

Q4 a) Describe slowly changing dimensions. What are the three types? Explain each type very briefly. [10]

b) The following table contains a training set D, of class-labeled tuples randomly selected from the AllElectronics customer database. Let buys_computer be the class label attribute. Using Naïve Bayesian classification predict the class label of a tuple X = (age = youth, income = medium, student = yes, credit rating = fair). [10]

| RID | age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Q5 a) Explain Data mining as a step in KDD with an appropriate diagram. [10]

b) Suppose data mining task is to cluster the following items into 2 clusters. Use the K-mean algorithm to cluster {3,5,11,13,4,21,35,12,30}. Write an algorithm for K-means clustering. [10]

Q6 a) What are the three major areas in the data warehouse? Relate and explain the architectural components to the three major areas. [10]

b) The following table shows the time spent writing an essay and essay grades obtained for students in an English course. [10]

| Hours spent on writing an essay | Grades |
|---|---|
| 6 | 82 |
| 10 | 88 |
| 2 | 56 |
| 4 | 64 |
| 6 | 77 |
| 7 | 92 |
| 0 | 23 |
| 1 | 41 |
| 8 | 80 |
| 5 | 59 |
| 3 | 47 |

(i) Use the method of least squares to find an equation for the prediction of a student's essay grade based on the hours spent on writing an essay in the English course.
(ii) Predict the essay grade of a student who spent 2.35 hours on writing an essay in the English course.

**********

97246