TE/ECS/SEM-VI/R-19/SH-24 Date: 12/12/24

Q.P. Code: 68676

Paper / Subject Code: 89324 / Data Warehousing and Mining

Time: 3 Hours **Total 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) Explain KDD process with diagram. **[05]**

b) Differentiate between OLTP Vs OLAP. **[05]**

c) Calculate Accuracy, Recall and Precision with the help of following data: **[05]**

True Positive (TP)= 30, True Negative (TN) = 55, False Positive (FP)= 5, False Negative (FN)= 10

d) What is Support and Confidence in market basket analysis. Explain with an example **[05]**

Q2 a) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. **[10]**

(i) Draw a star schema diagram for the data warehouse.
(ii) Draw a data cube and apply OLAP operations Slice, Dice, rollup and drill down to retrieve data from data cube.

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) Write the importance of data preprocessing. Explain different data cleaning techniques. **[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.

28

Q4 a) Describe Major steps in ETL process. [10]

b) The following table contains a training set D, of the weather conditions for playing a game of golf. Let Play Golf be the class label attribute. Using Naïve Bayesian classification predict the class label of a tuple X = ("Sunny, Hot, Normal, False"). [10]

RID	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Q5 a) Explain with example different data Sampling techniques. [10]

b) Explain multidimensional association mining and Multilevel Association mining with example. [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) Explain in detail Single linear regression and Multiple linear regression. [10]

29