Ref. No.: Ex/PG/SWE/T/116A/2024

M.E. SOFTWARE ENGINEERING
FIRST YEAR FIRST SEMESTER EXAM 2024
Department of Information Technology
Advanced Databases

Full Marks 100 **Time : 3.00 Hrs.**

Answer any five question

Q.1

a) What is the structure of a decision tree and how does it function in the context of machine learning?
b) How to a set logical rules in a decision tree?
c) What criteria are commonly used for splitting nodes in a decision tree?
d) How does a decision tree handle categorical variables during the splitting process?
e) What is pruning in the context of decision trees and why is it important?
f) Can decision trees handle missing data, and if so, how?
g) Apply the principle of decision tree construction to construct a decision tree with the given data set :

SepalLength	SepalWidth	PetalLength	PetalWidth	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
6.7	3.1	4.4	1.4	Versicolor
5.6	3.0	4.5	1.5	Versicolor
6.8	3.2	5.9	2.3	Virginica
7.6	3.0	6.6	2.1	Virginica
4.8	3.4	1.9	0.2	Setosa
6.3	3.3	4.7	1.6	Versicolor
7.2	3.6	6.1	2.5	Virginica
5.0	3.2	1.2	0.2	Setosa

2+2+2+2+2+2+8=20

Q.2

a) Given a transaction log with Log Sequence Numbers (LSN) ranging from 100 to 500, if a recovery process is at LSN 300, what transactions need to be redone and undone during the recovery?
b) If a transaction is partially completed before a failure occurs, and the recovery process identifies that it needs to be rolled back, explain the steps involved in log-based rollback.
c) How does a distributed commit protocol ensure atomicity in a distributed transaction? What challenges does it address in maintaining consistency across multiple nodes?
d) Explain the trade-offs involved in choosing between blocking and non-blocking strategies in the context of distributed commit protocols. Under what circumstances would you opt for a blocking or non-blocking approach, and what are the implications for system performance and responsiveness?
e) Discuss the potential impact of network partitions on the performance and reliability of distributed commit protocols. How can these protocols be designed to handle network failures effectively? **2+4+3+6+5=20**

[Turn over

Ref. No.: Ex/PG/SWE/T/116A/2024

Q.3

a) Apply and illustrate Apriori algorithm at each step to find association rules from the dataset given below with the support threshold of 0.2 and confidence threshold 0.9.

Transaction ID	Itemsets
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, c}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, c}

b) What do you mean by Association Rule Mining?

c) How does the antimonotone property of confidence can be used in rule generation for Apriori algorithm.

d) Give maximum number itemsets and maximum number of candidate 3- itemsets with a list of 7 items.

e) With a diagram discuss Lattice of Rules.

f) What are the disadvantages of Apriori algorithm ? **10+2+2+2+2+2=20**

Q.4

a) How does Mandatory Access Control differ from other access control models?

b) Discuss the common tactics employed by Trojan horses in exfiltration of sensitive data from compromised systems. How can organizations detect and prevent data leakage through these malicious programs?

c) State the role of differential privacy in protecting statistical databases. How does differential privacy help mitigate the risk of data leakage while still providing meaningful and accurate statistical results?

d) Give an example where the system provides for location and replication transparencies, but it does not provide for fragmentation transparency.

e) Explain Min term predict, COM_MIN algorithm in the context of the given **PROJECT** table

PNO	PNAME	BUDGET	LOC
P1	Instrumentation	150000	Montreal
P2	Database Develop	135000	New York
P3	CAD/CAM	250000	New York
P4	Maintenance	310000	Paris

2+4+4+4+6=20

Q.5

a) Discuss with an example the challenges and strategies involved in integrating data from disparate sources into a cohesive data warehouse architecture.

b) Define and explain the concept of a data cube. How does it differ from traditional relational databases, and what advantages does it offer for multidimensional analysis? Illustrate with example.

c) Explain the importance of data quality and cleansing in the context of data warehouse design. What techniques can be used to ensure high-quality data?

d) Consider the following data cube shown in Figure 1representing sales data and give output for the given data cube operations *Rolling up from quarterly to yearly sales, Drilling down from quarterly to monthly sales, Slicing the data cube for Q1_2022, Dicing the data cube for Product_A in Q1_2022*

	Product	Region	Time	Sales
i.	Product_A	North	Q1_2022	100
ii.	Product_A	South	Q1_2022	150
iii.	Product_B	North	Q1_2022	200
iv.	Product_B	South	Q1_2022	120
v.	Product_A	North	Q2_2022	120
vi.	Product_A	South	Q2_2022	180
vii.	Product_B	North	Q2_2022	220
viii.	Product_B	South	Q2_2022	130

Figure 1

4+4+4+8=20

Q.6

Write short notes on

a) Hierarchical Methods in clustering

b) Differentiate OLTP and OLAP

c) Correlation Analysis

d) Cluster Affinity Matrix

e) GCS

5X4 = 20