REG.NO.:

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

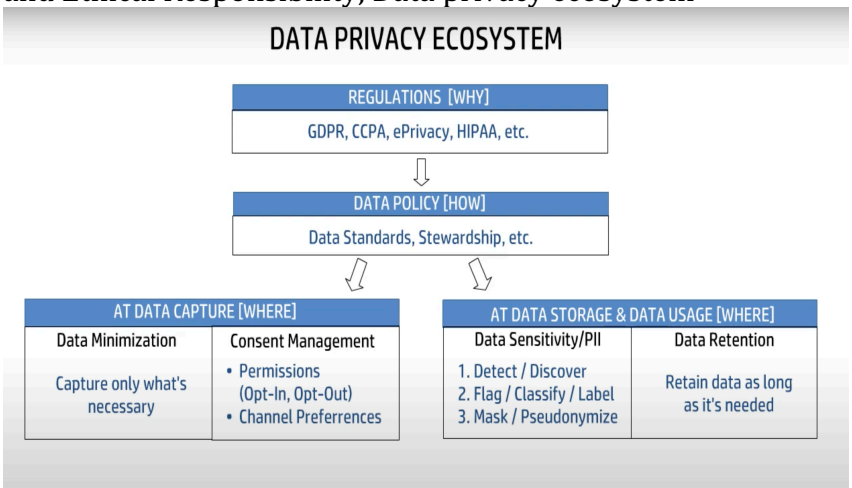
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING (SCOPE)
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2024-2025

SLOT: C2

Programme Name & Branch :B.Tech(CSE)
Course Code and Course Name :BCSE318L
Faculty Name(s) :Common to ALL C2_slots
Class Number(s) : VL2024250506630, VL2024250501609 VL2024250501588,
: VL2024250501573, VL2024250501605, VL2024250501720
: VL2024250501569
Date of Examination :29_01_2025, 2.00pm to 3.30pm
Exam Duration :90 minutes **Maximum Marks: 50**

General instruction(s):

- Answer All Questions

Q. No	Question	M	CO	BL
1.	What is your understanding of data privacy, and why is it important in today's world? Discuss the Data privacy ECO-system. **Key:** **Data privacy** refers to the right of individuals to control how their personal information is collected, used, shared, and stored. **important in today world**: Rise of Digital Interactions, Prevention of Identity Theft, Compliance with Regulations, Trust in Technology and Ethical Responsibility, Data privacy ecosystem 	10	CO1	BL2
2.	A bank is preparing to share its financial transactions dataset with external analysts to study spending patterns and improve fraud detection algorithms. However, since the dataset contains sensitive information (e.g., **transaction id, amounts, customer details, and geolocations etc**.), it is critical to ensure privacy protection through **Statistical Disclosure Control (SDC)** techniques. Identify the different stages SDC approaches, and infer the specific inputs and outputs associated with each stage on above scenario. Key: 1. Why is confidentiality protection needed? 2. What are the key characteristics and use of the data? 3. Disclosure risk definition and assessment.	10	CO1	BL2

	4. Disclosure control methods. 5. Implementation.			
3.	a. Calculate the global risk level for the micro data given below: (5 marks) **ID** | **Age** | **Gender** | **Zip Code** | **Salary** 1 | 25 | Male | 12345 | 50,000 2 | 40 | Female | 12345 | 70,000 3 | 40 | Female | 12346 | 70,000 4 | 25 | Male | 12346 | 50,000 5 | 35 | Female | 12347 | 80,000 b. Convert the above micro data in to indistinguishable using Non-perturbative techniques such as **range** and **suppression** on Quasi identifiers.(5 marks) **Key:** **a: Calculate Individual Risk: (5 marks)** For each record, the risk is 1/Frequency. For this dataset: • All records have a frequency of 1. • Therefore, the individual risk for all records is 1/1=1.0. **Calculate Global Risk Level:** The global risk level is the average of the individual risks: Global Risk Level=Sum of Individual Risks/Total Number of Record Global Risk Level=(1.0+1.0+1.0+1.0+1.0)/5=1.0 A **global risk level of 1.0** indicates that each record is uniquely identifiable, making this dataset highly vulnerable to re-identification. **b: Generalizations on Age and Zip code: (5 marks) Marks can also be awarded if the students use All are "*" in Zip code** **ID** | **Age** | **Gender** | **Zip Code** | **Salary** 1 | 20-30 | Male | 1234* | 50,000 2 | 30-50 | Female | 1234* | 70,000 3 | 30-50 | Female | 1234* | 70,000	10	CO2	BL3

	4	20-30	Male	1234*	50,000			
	5	30-50	Female	1234*	80,000			

4 Apply the following perturbation techniques to the dataset provided below to ensure data privacy:

a. Noise Addition (5 marks)
b. Micro-aggregation (5 marks)

ID	Age	Gender	Diagnosis	Treatment_Cost	Visit_Date
P1	45	Male	Diabetes	1500	10/01/24
P2	29	Female	Hypertension	2000	12/01/24
P3	37	Male	Asthma	1800	14/01/24
P4	50	Female	Heart Disease	2500	15/01/24
P5	33	Male	Migraine	1200	17/01/24

Key:

10 CO2 BL3

a. Noise addition involves slightly altering numerical values in the dataset to maintain privacy while preserving overall trends. Add a random percentage of noise to the Treatment_Cost column. (For eg. Add a noise value within ±10% of the original Treatment_Cost.)

ID	Age	Gender	Diagnosis	Treatment_Cost	Visit_Date
P1	45	Male	Diabetes	1620	10/01/24
P2	29	Female	Hyper-tension	1900	12/01/24
P3	37	Male	Asthma	1836	14/01/24
P4	50	Female	Heart Disease	2375	15/01/24

		P5	33	Male	Migraine	1260	17/01/24			

b. Steps to Apply Micro-Aggregation:

- **Identify the Sensitive Attribute(s)**: In this dataset, let's consider Age and Treatment_Cost as sensitive attributes.
- **Determine the Group Size (k)**: Choose a group size, e.g., k=3.
- **Sort the Data**: Sort the dataset based on the sensitive attributes, in this case **Age**).
- **Cluster Records**: Divide the sorted dataset into groups of size k.
- **Aggregate Within Each Group**: Replace the values of sensitive attributes within each group with the group's average.

1. **Sort the micro data by Age**
2. **Group the Data (k=3)**
   - Group 1: P2, P5, P3
   - Group 2: P1, P4
3. **Calculate Group Averages**
   - **Group 1 Averages**:
     - Age = (29 + 33 + 37) / 3 = 33
     - Treatment_Cost = (2000 + 1200 + 1800) / 3 = 1667
   - **Group 2 Averages**:
     - Age = (45 + 50) / 2 = 47.5 =(47)
     - Treatment_Cost = (1500 + 2500) / 2 = 2000
4. **Replace Values with Group Averages:**

Pat_ID	Age	Gender	Diagnosis	Treatment_Cost	Visit_Date
P2	33	Male	Diabetes	1667	12/01/24
P5	33	Female	Hyper-tension	1667	17/01/24
P3	33	Male	Asthma	1677	14/01/24
P1	47	Female	Heart Disease	2000	10/01/24
P4	47	Male	Migraine	2000	15/01/24

5	Imagine a healthcare organization wants to share patient data with researchers to study disease patterns and improve treatments. The organization needs to remove or mask personal details like names and addresses to protect patient privacy. However, they also want	10	CO2	BL2

	the data to remain accurate enough for researchers to find meaningful insights. How can a healthcare organization share patient data with researchers in a way that protects privacy while keeping the information accurate enough for valuable analysis? **Key:** Healthcare organizations can balance privacy and utility by applying anonymization, synthetic data, secure access, data minimization, obtaining consent, conducting risk assessments, and complying with privacy regulations. Key points alone mentioned above, further details are required from the students:			

<<<THE END>>>