440

REG.NO.: 22BCE10168

Vellore Institute of Technology

SCHOOL COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2024-2025

SLOT: C1 + TC1

Programme Name & Branch : B.Tech. & Computer Science (Core and Specialization)
Course Code and Course Name : BCSE318L & Data Privacy
Faculty Name(s) : Prof. ANBARASI M, Prof. UMADEVI K S, Prof. GLADYS GNANA KIRUBA B, Prof. SUDARSHAN NANDY, Prof. YOGANAND, Prof. THANGARAMYA K
Class Number(s) : VL2024250501576, 1565, 1719, 1580, 1604, 1607
Date of Examination : 29-01-2025
Exam Duration : 90 minutes
Maximum Marks: 50

General instruction(s):

- Answer All Questions

Q. No	Question	M	CO	BL
1.	a) Classify and describe the different types of identifiers for protecting sensitive data from in the given dataset.	5	1	2

Name	Age	Gender	Blood Type	Date of Admi	Insurance Provider	Billing Amount	Medication	Test Results
Bobby Jackson	30	Male	B-	31-01-2024	Blue Cross	18856.28	Paracetamol	Normal
Leslie Terry	62	Male	A+	20-08-2019	Medicare	33643.33	Ibuprofen	Inconclusive
Danny Smith	76	Female	A-	22-09-2022	Aetna	27955.1	Aspirin	Normal
Andrew Watts	28	Female	O+	18-11-2020	Medicare	37909.78	Ibuprofen	Abnormal
Emily Johnson	36	Male	A+	20-12-2023	UnitedHealthcare	48145.11	Ibuprofen	Normal
Edward Edwards	21	Female	AB-	03-11-2020	Medicare	19580.87	Paracetamol	Inconclusive
Christina Martinez	20	Female	A+	28-12-2021	Cigna	45820.46	Paracetamol	Inconclusive

Q. No	Question	M	CO	BL
	b) In a healthcare organization, data privacy is a complex issue with numerous stakeholders. Identify the key players, their roles and responsibilities in detail.	5		
2.	You are asked to analyse the credit risk assessment of the loan applicants for ABC bank based on an extensive data of the loan applicants such as demographic information (age, income, occupation), financial history (credits score, loan history) and behavioural data (spending pattern, transaction history). Create the loan dataset with minimum of 5 records and convert it into anonymized dataset using above constraints and present the privacy gain and utility loss in the anonymized data. Also discuss the importance to balance between utility and privacy.	10	1	2
3.	Explain the various methods that do not supress values in the dataset, but alter the values to limit disclosure risk by creating uncertainty around the true values. Most of the methods are based on principles of matrix masking and also applied to both categorical and continuous datasets.	10	2	2
4.	a) A public health agency wants to release data on COVID-19 vaccination rates by zip code to inform public health initiatives. However, releasing the raw data could potentially reveal sensitive information about individuals living in sparsely populated	5		

zip codes. There is a disclosure risk, if a zip code has a very small population and a unique vaccination rate, it is possible to identify individuals within that zip code, especially if they have other publicly available information (like age or occupation). Use global recoding technique and convert the original data into recoded data.

Hint:

i). Determine zip codes with very low population counts.

ii). Group Zip Codes: Combine these sparsely populated zip codes into broader geographic regions (e.g., "Rural Area A," "Rural Area B").

iii). Calculate the vaccination rates for these newly formed regions by taking average.

Original Data:

Zip Code 12345: Population 50, Vaccination Rate 80%
Zip Code 56789: Population 15, Vaccination Rate 90%
Zip Code 98765: Population 8, Vaccination Rate 75%
Zip Code 76254: Population 40, Vaccination Rate 85%
Zip Code 65738: Population 35, Vaccination Rate 70%
Zip Code 26453: Population 5, Vaccination Rate 65%
Zip Code 18495: Population 10, Vaccination Rate 75%

2 3

b)

Sl.No.	Occupation	City/Town	Gender
1.	Mayer	Chennai	Male
2.	Plumber	Vellore	Male
3.	Teacher	Chennai	Female
4.	Fisherman	Cuddalore	Male
5.	Police	Chennai	Female

Check whether local suppression technique can be applied or not for the above-mentioned dataset. If applied justify the reason. 5

5. a) Consider the following table in which there are two identifiers as ID and Name. Use suitable privacy preserving method to convert original EI into masked EI. Discuss the identifier technique.

Original EI

ID	Name
12345	John
56789	Harry
52131	Hari
76548	Mary
87543	Srini
23131	Chan

5

b) Sensitive datasets (SD) consist of numerical data or strings. Usually, SD are left in their original form to enable the utility of data, but there are some exceptional scenarios where it is necessary to protect SD to a certain extent. With this scenario explain the data protection methods to effectively anonymize the sensitive data for the multidimensional dataset. 5 2 3