

School of Computer Science and Engineering

Winter Semester 2023-24

CONTINUOUS ASSESSMENT TEST – 1

SLOT: G1+TG1

Programme Name: B.Tech- Computer Science

Course Name & Code: Data Privacy & BCSE318L

Class Number (s): VL2023240500995, VL2023240500997

Faculty Name (s): Dr. Geetha Mary A, Dr. Jasmin T Jose

Exam Duration: 90 Min.

Maximum Marks: 50

Answer all Questions

Q. No.	Question	Max. Marks	CO	BL																																																																		
1	A. Compare and contrast Data privacy and Data security with suitable example. (5 Marks) Defn. of both data privacy and data security- 2Marks Explanation- 2Marks Example -1Mark B. Discuss different data protection methods with example (5 Marks) Explanation of Cryptography, Anonymization and Tokenization- 3Marks Example for each- 2Marks	10	CO1	BL1																																																																		
2	Two data sets are given below. Carry out different stages of SDC approach on these two data sets. Show the results of each stage of SDC on these data sets separately. A. Data set1 (4 Marks) <table><tr><th>Time</th><th>Wind</th><th>Rain</th><th>Humidity</th><th>Visibility</th></tr><tr><td>1.00 pm</td><td>15km/hr</td><td>1.2mm</td><td>54%</td><td>16km</td></tr><tr><td>3.00 pm</td><td>15km/hr</td><td>1.0mm</td><td>59%</td><td>16km</td></tr><tr><td>5.00 pm</td><td>15km/hr</td><td>1.1mm</td><td>58%</td><td>16km</td></tr><tr><td>7.00 pm</td><td>14.8km/hr</td><td>0.5mm</td><td>57%</td><td>15.7km</td></tr><tr><td>9.00 pm</td><td>14.8km/hr</td><td>0.3mm</td><td>53%</td><td>15.7km</td></tr></table> Stage 1: Need for privacy protection: As there is no sensitive data in the given data set, no privacy preservation is required. The given data set can be released without applying any privacy preservation techniques. (4Marks) B. Data set2 (6 marks) <table><tr><th>Id</th><th>Name</th><th>Location</th><th>Age</th><th>Gender</th><th>Disease</th></tr><tr><td>1200</td><td>Amos</td><td>Vellore</td><td>61</td><td>M</td><td>flu</td></tr><tr><td>1400</td><td>Pratheek</td><td>Ambur</td><td>29</td><td>M</td><td>ulcer</td></tr><tr><td>1500</td><td>Steffi</td><td>Ranippettu</td><td>22</td><td>F</td><td>gastritis</td></tr><tr><td>2000</td><td>Krithik</td><td>Vellore</td><td>27</td><td>M</td><td>pneumonia</td></tr><tr><td>2600</td><td>Ram</td><td>Vellore</td><td>43</td><td>M</td><td>asthma</td></tr></table>	Time	Wind	Rain	Humidity	Visibility	1.00 pm	15km/hr	1.2mm	54%	16km	3.00 pm	15km/hr	1.0mm	59%	16km	5.00 pm	15km/hr	1.1mm	58%	16km	7.00 pm	14.8km/hr	0.5mm	57%	15.7km	9.00 pm	14.8km/hr	0.3mm	53%	15.7km	Id	Name	Location	Age	Gender	Disease	1200	Amos	Vellore	61	M	flu	1400	Pratheek	Ambur	29	M	ulcer	1500	Steffi	Ranippettu	22	F	gastritis	2000	Krithik	Vellore	27	M	pneumonia	2600	Ram	Vellore	43	M	asthma	10	CO2	BL3
Time	Wind	Rain	Humidity	Visibility																																																																		
1.00 pm	15km/hr	1.2mm	54%	16km																																																																		
3.00 pm	15km/hr	1.0mm	59%	16km																																																																		
5.00 pm	15km/hr	1.1mm	58%	16km																																																																		
7.00 pm	14.8km/hr	0.5mm	57%	15.7km																																																																		
9.00 pm	14.8km/hr	0.3mm	53%	15.7km																																																																		
Id	Name	Location	Age	Gender	Disease																																																																	
1200	Amos	Vellore	61	M	flu																																																																	
1400	Pratheek	Ambur	29	M	ulcer																																																																	
1500	Steffi	Ranippettu	22	F	gastritis																																																																	
2000	Krithik	Vellore	27	M	pneumonia																																																																	
2600	Ram	Vellore	43	M	asthma																																																																	

	<table><tr><td>3000</td><td>Prathyaksha</td><td>Ambur</td><td>52</td><td>M</td><td>flu</td></tr><tr><td>3400</td><td>Anushka</td><td>Alamelupuram</td><td>47</td><td>F</td><td>bronchitis</td></tr><tr><td>5000</td><td>Lathika</td><td>Ranippettu</td><td>30</td><td>F</td><td>Bronchitis</td></tr><tr><td>5700</td><td>Sonam</td><td>Vellore</td><td>36</td><td>M</td><td>Pneumonia</td></tr><tr><td>6100</td><td>Hrithika</td><td>Vellore</td><td>32</td><td>F</td><td>cancer</td></tr><tr><td>6400</td><td>Pooja</td><td>Hozur</td><td>70</td><td>M</td><td>Flu</td></tr><tr><td>7000</td><td>Avinash</td><td>Ambur</td><td>36</td><td>M</td><td>Ulcer</td></tr><tr><td>8600</td><td>George</td><td>hozur</td><td>48</td><td>M</td><td>pneumonia</td></tr></table> 5 stages of SDC -2 Marks Inputs and outputs of each stage based on given data-4 Marks	3000	Prathyaksha	Ambur	52	M	flu	3400	Anushka	Alamelupuram	47	F	bronchitis	5000	Lathika	Ranippettu	30	F	Bronchitis	5700	Sonam	Vellore	36	M	Pneumonia	6100	Hrithika	Vellore	32	F	cancer	6400	Pooja	Hozur	70	M	Flu	7000	Avinash	Ambur	36	M	Ulcer	8600	George	hozur	48	M	pneumonia																																																																			
3000	Prathyaksha	Ambur	52	M	flu																																																																																																															
3400	Anushka	Alamelupuram	47	F	bronchitis																																																																																																															
5000	Lathika	Ranippettu	30	F	Bronchitis																																																																																																															
5700	Sonam	Vellore	36	M	Pneumonia																																																																																																															
6100	Hrithika	Vellore	32	F	cancer																																																																																																															
6400	Pooja	Hozur	70	M	Flu																																																																																																															
7000	Avinash	Ambur	36	M	Ulcer																																																																																																															
8600	George	hozur	48	M	pneumonia																																																																																																															
3	A. “Maintaining the balance between privacy and utility of data is very important”. Elaborate the statement with neat sketches and suitable example. (6 Marks) Explanation- 3 Marks Diagram- Privacy Vs. Utility -2 Marks Example – 1 Mark B. Identify the records which are in high disclosure risk with the help of frequency table based on the combination of 3 Quasi Identifiers (Marital Status, Age and Gender) for the following data set. (4Marks) <table><tr><th>Id</th><th>Marital status</th><th>Location</th><th>Age</th><th>Gender</th><th>Disease</th></tr><tr><td>1400</td><td>M</td><td>Ambur</td><td>29</td><td>M</td><td>Ulcer</td></tr><tr><td>3000</td><td>N</td><td>Ambur</td><td>48</td><td>F</td><td>Flu</td></tr><tr><td>5000</td><td>N</td><td>Hozur</td><td>30</td><td>F</td><td>Bronchitis</td></tr><tr><td>5700</td><td>M</td><td>Vellore</td><td>17</td><td>M</td><td>Pneumonia</td></tr><tr><td>7000</td><td>N</td><td>Ambur</td><td>36</td><td>M</td><td>Ulcer</td></tr><tr><td>8600</td><td>N</td><td>Hozur</td><td>48</td><td>F</td><td>Pneumonia</td></tr><tr><td>2300</td><td>M</td><td>Vellore</td><td>29</td><td>M</td><td>Cancer</td></tr><tr><td>2302</td><td>N</td><td>Ambur</td><td>30</td><td>F</td><td>Flu</td></tr><tr><td>3210</td><td>M</td><td>Vellore</td><td>29</td><td>F</td><td>Ulcer</td></tr><tr><td>2440</td><td>N</td><td>Hozur</td><td>36</td><td>M</td><td>Flu</td></tr><tr><td>8100</td><td>M</td><td>Vellore</td><td>29</td><td>M</td><td>Cancer</td></tr><tr><td>3120</td><td>M</td><td>Ambur</td><td>29</td><td>F</td><td>Ulcer</td></tr><tr><td>2700</td><td>N</td><td>Vellore</td><td>48</td><td>F</td><td>Bronchitis</td></tr></table> Frequency table: 2 Marks <table><tr><th>Marital status</th><th>Age</th><th>Gender</th><th>Frequency</th></tr><tr><td>M</td><td>29</td><td>M</td><td>3</td></tr><tr><td>N</td><td>48</td><td>F</td><td>3</td></tr><tr><td>N</td><td>30</td><td>F</td><td>2</td></tr><tr><td>M</td><td>17</td><td>M</td><td>1</td></tr><tr><td>N</td><td>36</td><td>M</td><td>2</td></tr><tr><td>M</td><td>29</td><td>F</td><td>2</td></tr></table> Identification of record with highest re-identification risk- 1 Marks Explanation- 1 Marks	Id	Marital status	Location	Age	Gender	Disease	1400	M	Ambur	29	M	Ulcer	3000	N	Ambur	48	F	Flu	5000	N	Hozur	30	F	Bronchitis	5700	M	Vellore	17	M	Pneumonia	7000	N	Ambur	36	M	Ulcer	8600	N	Hozur	48	F	Pneumonia	2300	M	Vellore	29	M	Cancer	2302	N	Ambur	30	F	Flu	3210	M	Vellore	29	F	Ulcer	2440	N	Hozur	36	M	Flu	8100	M	Vellore	29	M	Cancer	3120	M	Ambur	29	F	Ulcer	2700	N	Vellore	48	F	Bronchitis	Marital status	Age	Gender	Frequency	M	29	M	3	N	48	F	3	N	30	F	2	M	17	M	1	N	36	M	2	M	29	F	2	10	CO1	BL2
Id	Marital status	Location	Age	Gender	Disease																																																																																																															
1400	M	Ambur	29	M	Ulcer																																																																																																															
3000	N	Ambur	48	F	Flu																																																																																																															
5000	N	Hozur	30	F	Bronchitis																																																																																																															
5700	M	Vellore	17	M	Pneumonia																																																																																																															
7000	N	Ambur	36	M	Ulcer																																																																																																															
8600	N	Hozur	48	F	Pneumonia																																																																																																															
2300	M	Vellore	29	M	Cancer																																																																																																															
2302	N	Ambur	30	F	Flu																																																																																																															
3210	M	Vellore	29	F	Ulcer																																																																																																															
2440	N	Hozur	36	M	Flu																																																																																																															
8100	M	Vellore	29	M	Cancer																																																																																																															
3120	M	Ambur	29	F	Ulcer																																																																																																															
2700	N	Vellore	48	F	Bronchitis																																																																																																															
Marital status	Age	Gender	Frequency																																																																																																																	
M	29	M	3																																																																																																																	
N	48	F	3																																																																																																																	
N	30	F	2																																																																																																																	
M	17	M	1																																																																																																																	
N	36	M	2																																																																																																																	
M	29	F	2																																																																																																																	
4	Consider the data set given below, A. Identify different types of attributes in it and suggest some methods to protect the above mentioned attribute types. (2 marks) EI- Name. ID	10	CO2	BL3																																																																																																																

	Methods: Masking and suppression QI- Location, Age, Gender Methods: Any anonymization/De-identification techniques SD- Income, Disease No transformation on SD, it is left in its original form B. Apply any four perturbative methods for data privacy preservation on following data set and show the releasable data set after applying each method (8Marks) <table><tr><th>Name</th><th>Location</th><th>Income</th><th>Age</th><th>Id</th><th>Gender</th><th>Disease</th></tr><tr><td>Pratheek</td><td>Ambur</td><td>15K</td><td>29</td><td>1400</td><td>M</td><td>Ulcer</td></tr><tr><td>Prathyaksha</td><td>Ambur</td><td>20K</td><td>52</td><td>3000</td><td>M</td><td>Flu</td></tr><tr><td>Lathika</td><td>Hozur</td><td>65K</td><td>30</td><td>5000</td><td>F</td><td>Bronchitis</td></tr><tr><td>Sonam</td><td>Vellore</td><td>15K</td><td>36</td><td>5700</td><td>M</td><td>Pneumonia</td></tr><tr><td>Avinash</td><td>Ambur</td><td>35K</td><td>36</td><td>7000</td><td>M</td><td>Ulcer</td></tr><tr><td>George</td><td>Hozur</td><td>48K</td><td>48</td><td>8600</td><td>M</td><td>Pneumonia</td></tr></table> Any 4 perturbative methods with output- 8 Marks	Name	Location	Income	Age	Id	Gender	Disease	Pratheek	Ambur	15K	29	1400	M	Ulcer	Prathyaksha	Ambur	20K	52	3000	M	Flu	Lathika	Hozur	65K	30	5000	F	Bronchitis	Sonam	Vellore	15K	36	5700	M	Pneumonia	Avinash	Ambur	35K	36	7000	M	Ulcer	George	Hozur	48K	48	8600	M	Pneumonia			
Name	Location	Income	Age	Id	Gender	Disease																																															
Pratheek	Ambur	15K	29	1400	M	Ulcer																																															
Prathyaksha	Ambur	20K	52	3000	M	Flu																																															
Lathika	Hozur	65K	30	5000	F	Bronchitis																																															
Sonam	Vellore	15K	36	5700	M	Pneumonia																																															
Avinash	Ambur	35K	36	7000	M	Ulcer																																															
George	Hozur	48K	48	8600	M	Pneumonia																																															
5.	A. Annotate the trade-off between information loss and utility of data in data privacy and list out the minimum criteria that should be preserved for the acceptable level of information loss in a protected microdata. (5 marks) Explanation- 3 marks Diagram – 1 Mark Minimum criteria -1 Mark - Preserving mean and variance on subsets - Maintaining at-least one distributional characteristics B. ‘Global recoding is better than local recoding’. Give your comment on this statement and prove your answer with suitable example. (5 marks) Correct comment with explanation – 3 marks The given statement is not true always. If privacy is the concern, Global recoding is better. But, if information gain/utility is the concern, local recoding is better. Example for both- 2 marks	10	CO2	BL2																																																	