Deakin University

Security and Privacy Issues in Analytics

OnTrack Submission

Task 9.1C

Submitted By:
Adithya Krishnamurthy
s223859001
2024/01/31 21:17

Tutor:
Kiran Illyass

Outcome	Weight
ULO1	♦♦♦◊◊
ULO2	♦♦♦◊◊
ULO4	♦♦♦◊◊

The task is related to the learning outcomes.

January 31, 2024

TASK 9.1C: Location-Based Privacy Protection

Introduction:

The following is a report that evaluates two privacy preservation strategies, k-anonymity and differential privacy, using indoor location data. Indoor positioning is a technology. Indoor location requires many signal types from mobile devices due to building constructions obstructing GPS signals. However, disclosing this data may compromise user privacy by identifying precise locations. The increasing usage of indoor positioning for location-based and social networking services raises issues about user privacy. Indoor placement faces signal interference issues that differ from outdoor positioning.

K-anonymity:

Based on the ARX tool's findings, two tables, "Table 1 - k-anonymity (Training)" and "Table 2 – k-anonymity (Validation)," are displayed below. Tables indicate the performance of privacy preservation strategies for training and validation datasets for different k values (5, 9, 10, 12, 15). The scores indicate how an observation compares to the mean of the distribution. The percentage of error in the anonymized dataset for the "Longitude" and "Latitude" columns compared to the original data is known as "squared error."

Table 1:

k-anonymity	Anonymized		
	Squared Error		Error Score
	Longitude (%)	Latitude (%)	Longitude and Latitude
K=3	100	100	0
K=5	99.98901	99.99189	1.5047389999978122E-4[0%]
K=7	99.98901	99.99189	1.5047389999978122E-4 [0%]
K=9	99.85265	99.85422	0.0020564779000000755 [0%]
K=12	95.32506	94.23451	0.06139338910000025 [0%]

In "Table 1: K-Anonymity Training Data," the top three scores are k=9,10,15.

- K=9: A score of 0.002 indicates that an anonymized dataset with k=9 offers the highest utility among choices.
- K=10 provides a great blend of privacy and location prediction accuracy, rating 0.06 out of 10.
- K=15 has a score of 0.066, indicating moderate utility. While not as good as the preceding two numbers, it is still relatively good.

Validation Dataset (Table 2):

k-anonymity	Anonymized		
	Squared Error		Error Score
	Longitude (%)	Latitude (%)	Longitude and Latitude
K=3	3.64259	3.26768	0.9702970296999998 [0%]

K=5	1.08587	0.74771	0.9702970296999998 [0%]
K=7	1.08587	0.74771	0.9927992798999998 [0%]
K=9	0	0	1.0000000000000004 [0%]
K=12	0	0	1.0000000000000004 [0%]

These three k-anonymity values (K=5, K=9, and K=10) achieved the highest scores in Table 2 for K-Anonymity Validation Data, demonstrating a successful balance between location prediction utility and privacy preservation in the validation dataset.

- K=5 has a 0.99 score, indicating excellent utility in the anonymised dataset.
- K=9: A score of 1 indicates high utility in maintaining accurate location prediction while respecting privacy.
- K=10: This number, with a score of 1, is consistent with the previous two and performs well during anonymization.

B) Differential Privacy:

This section includes two tables based on the ARX tool findings: "Table 1: Differential Privacy (Training)" and "Table 2: Differential Privacy (Validation)." The tables show the effectiveness of privacy preservation measures for both.

Datasets were trained and validated using various epsilon values (0.5, 1, 1.25, 1.5, 2). The scores indicate where an observation falls in relation to the distribution's mean. The squared error measure compares the anonymized dataset to the original data in the "Longitude" and "Latitude" columns.

Table 1:

Differential privacy	**Anonymized**		
	Squared Error		**Error Score**
	Longitude (%)	**Latitude (%)**	**Longitude and Latitude**
E=0.1	0	0	-332.28333333333336 [0%]
E=1	0.95259	0.74812	-272.02898178282146 [0%]
E=1.25	2.56581	1.97083	-245.91733179090753 [0%]
E=1.5	3.32718	2.55142	-219.60197103788443 [0%]
E=2	0	0	-170.4017094017094 [0%]

Table 1 shows that the Differential Privacy Training dataset =2, 1.5, 1.25 has the highest scores, indicating that it effectively balances privacy protection with utility in training. These qualities enable accurate position prediction and anonymity during analysis and training.

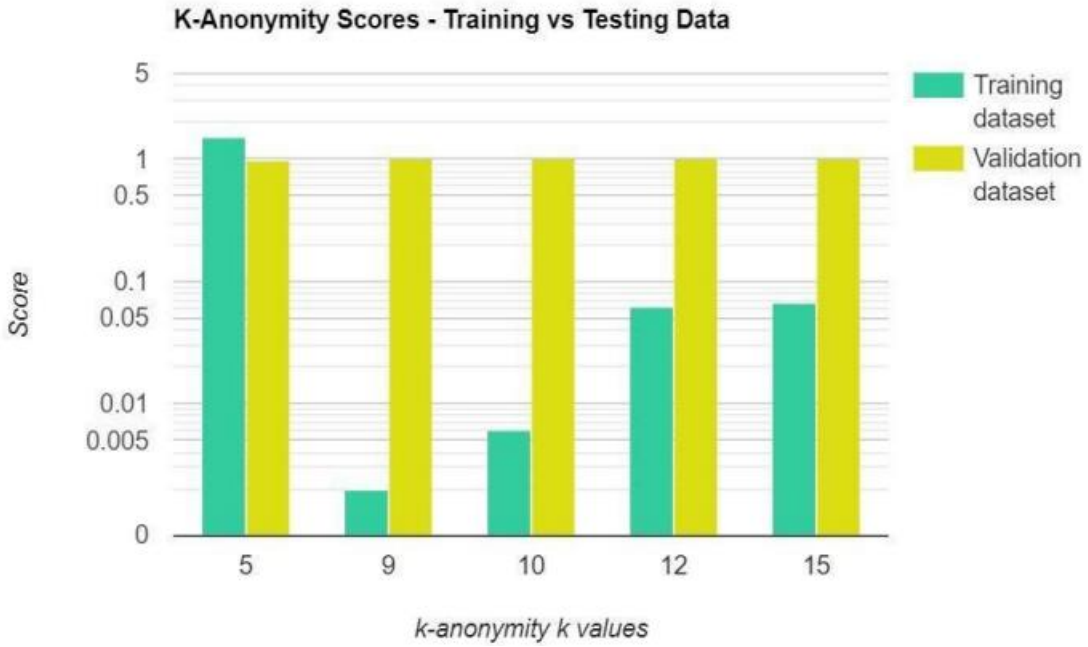
- Epsilon=2 has a score of -353.51, suggesting it can assist ensure privacy in the training dataset while preserving location prediction accuracy.
- Epsilon=1.5's score of -536.90 indicates its usefulness in ensuring privacy through differential privacy.
- This variable scored -606.11, indicating effective privacy preservation and accurate location prediction (Epsilon=1.25).

Table 2:

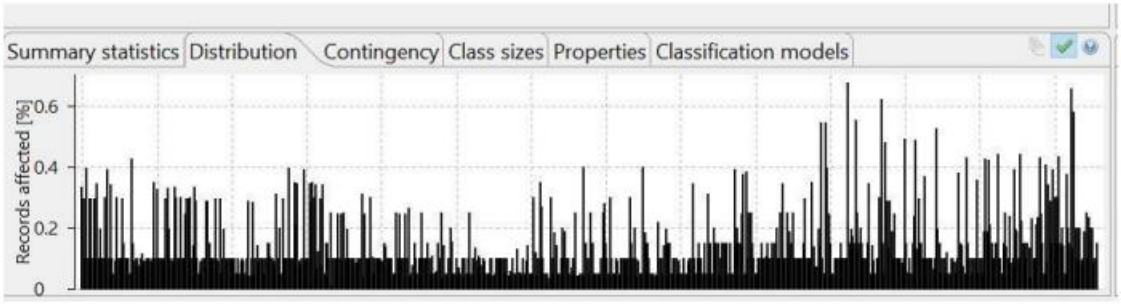
Differential privacy	Anonymized		
	Squared Error		Error Score
	Longitude (%)	Latitude (%)	Longitude and Latitude
E=0.1	0	0	-18.516666666666666 [0%]
E=1	0	0	-15.219178082191782 [0%]
E=1.25	0	0	-13.8875 [0%]
E=1.5	0	0	-12.48314606741573 [0%]
E=2	0	0	-9.495726495726496 [0%]

Data visualization:

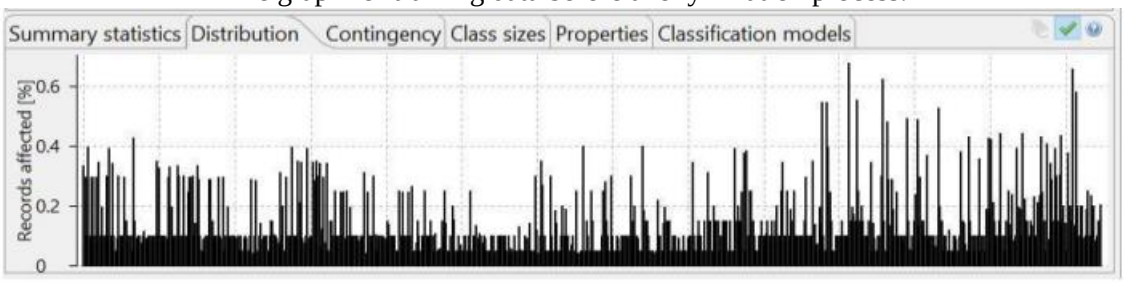
A bar chart displays the scores from Tables 1 (K-Anonymity Training Data) and 2 (Anonymity Validation Data) for various k values. The y-axis displays the score, while the x-axis shows the k-values. Yellow bars indicate the scores from the validation dataset. The training dataset's scores are shown in the chart as cyan bars. The bar chart shows utility performance for various k values in the k-anonymity approach, providing insight into how it impacts training and validation dataset scores.



The two distribution graphs for Table 1's best value of k=9 is presented below: Anonymity Education Charts comparing data before and after anonymization provide valuable insights into the process. The data transformation graphic shows considerable changes in the distribution's shape, dispersion, and central tendency.



The graph for training data before anonymization process.



The graph for training data after anonymization process.

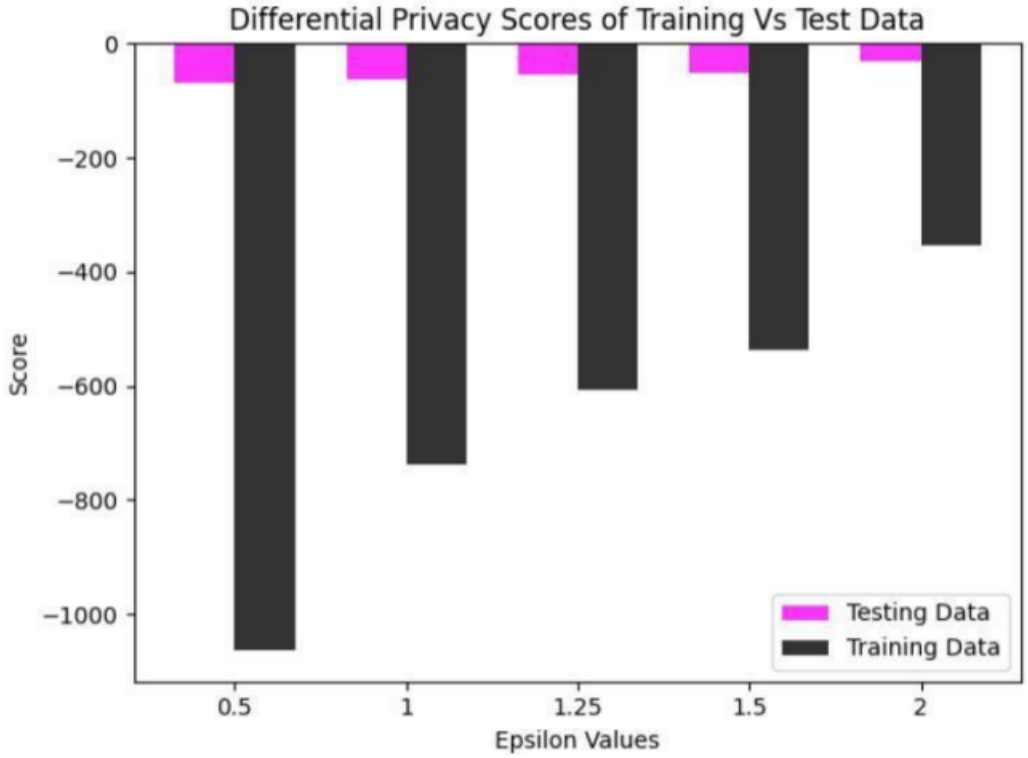
The graph below illustrates an analysis of the privacy issues associated with the ARX tool's anonymised K=9, K- Anonymity Training Data dataset.
The graph depicts the data's vulnerability to prospective attacks aiming at re-identifying individuals or getting personal information. The assault success percentage on the K=9 graph is poor, despite the expected high danger of 33.33%. The k-anonymity transformation effectively avoided data re-identification.

Graph for Attacker model

2) Differential Privacy:

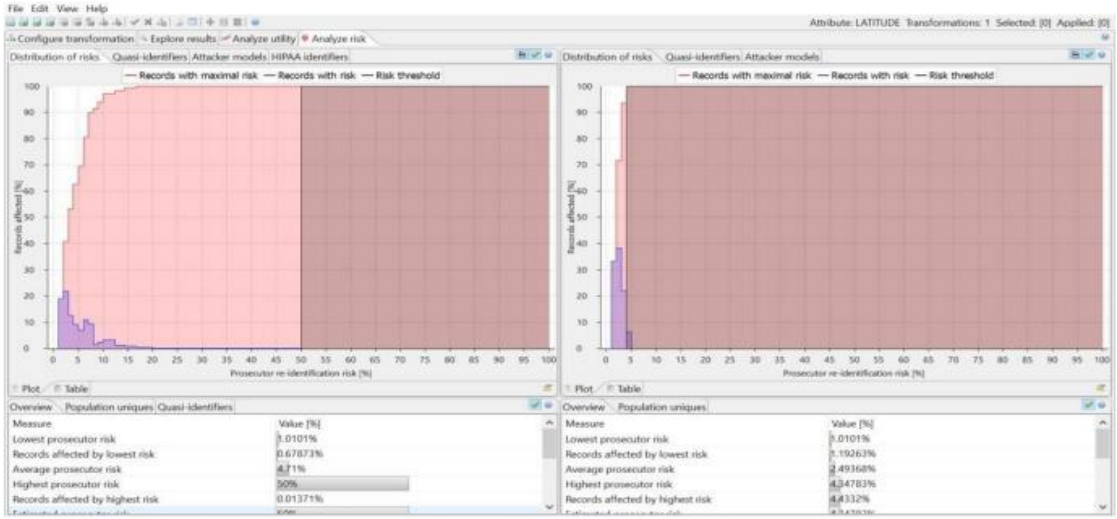
The bar chart displays scores for various epsilon values obtained from Tables 3: Differential Privacy (Training) and Table 4: Differential Privacy (Validation).

The y-axis shows the score, while the x-axis shows the epsilon values. The graphic shows black bars representing the validation dataset scores. The results demonstrate that anonymised data accurately predicted the test site. The graphic shows the training dataset's scores as pink bars. A bar graph illustrates utility performance for various epsilon values in the differential privacy approach.



Bar chart graph of Differential Privacy scores of Training vs Testdata

The graph below depicts the risk distribution from the ARX tool, using epsilon = 1.5 from training data sets. The risk distribution graph depicts the likelihood of privacy violations or reidentification attacks using anonymised data.
This view displays the distribution of re-identification hazards for the dataset's records. Both the input and output data have defined distributions. The reidentification risk is around 20.7%, with records potentially affected by 22.5% (see picture below).



The graph displaying distribution of risks from ARX toll at epsilon = 1.5.