Anonymization of a Dataset with Utility and Risk Analysis

Security and Privacy - Assignment #4
2023/2024

Marta Longo 202207985
Sara Táboas 202205101

Table of Contents

1. Introduction

Data privacy is a critical concern in today's digital age, where vast amounts of personal information are collected, stored, and processed by organizations. Ensuring that this data is protected against unauthorized access and misuse is essential. One effective method for protecting personal data is anonymization, which aims to alter the data in such a way that individual identities cannot be easily discerned.

The objective of this assignment is to apply anonymization techniques to a large dataset using the ARX Data Anonymization Tool. ARX is a powerful open-source software designed to facilitate the anonymization process by providing a variety of privacy models, data transformation methods, and risk analysis tools. This assignment will guide you through the process of anonymizing a dataset to meet specified privacy requirements while maintaining data utility.

2. ARX Software

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of privacy and risk models, methods for transforming data and methods for analyzing the usefulness of output data.

This software allows users to import data from various sources, including CSV files, MS Excel, and SQL databases. It also supports exporting the anonymized data back into these formats, enabling easy integration with other tools and workflows. It also permits the classification attributes into identifying, quasi-identifying, sensitive, and insensitive categories. This classification helps in applying appropriate anonymization techniques to different types of data. It also has many privacy models, with or without the need to have a sensitive attribute and supports various data transformation techniques.

By leveraging ARX's capabilities, users can ensure that their datasets are anonymized effectively, balancing privacy protection with data utility. This comprehensive approach helps organizations comply with privacy regulations and protect individual privacy while maintaining the usefulness of their data for analysis and decision-making

3. Dataset and attributes analysis

In this section, we will classify the attributes in four categories:

1. **Identifying** - These are attributes that can uniquely identify an individual. They are directly associated with the identity of a person.
2. **Quasi-identifying** - Attributes that do not uniquely identify an individual on their own but can be combined with other quasi-identifiers to re-identify individuals. They are critical for linkage attacks where an attacker uses external data sources to identify individuals.
3. **Sensitive** - These attributes contain sensitive information that should be protected to prevent harm or discrimination if disclosed. They often hold personal information about health, finances, or other confidential aspects.
4. **Insensitive** - Insensitive attributes are those that do not contribute to identifying an individual and do not hold sensitive information. They are generally safe to be left unchanged or minimally transformed.

Attributes	**Classification**	**Justification**
sex	Quasi-Identifying (0.006%-distinction; 43.8%-separation)	Sex is a 0.006%-distinction attribute, which means there are few distinct values and a 43.8%-separation attribute, which means 43% of all possible combinations can be separated by sex. Using sex and age we can almost identify the individual (QID). Other examples like sex and occupation.
age	Quasi-Identifying (0.23%-distinction; 97.8%-separation)	Age is a 0.23%-distinction attribute, which means there are some distinct values and a 97.8%-separation attribute, which means 97.8% of all possible combinations can be separated by age. Therefore, age can be considered in the dataset a QID attribute. Combining age with, per example, sex we can almost identify an individual.
race	Quasi-Identifying (0.016%-distinction; 25.1%-separation)	Race is a 0.016%-distinct attribute, which means there are a few distinct values and a 25.1%-separation attribute, which means 25.1% of all possible combinations can be separated by race. Race can be

		considered a QID attribute, due the its combination with other attributes, such as age, that almost allows individual identification.
marital-status	Sensitive (0.023%-distinction; 65.7%-separation)	Marital-status is a 0.023%-distinct attribute, which means there are few distinct values and a 65.7%-separation attribute, which means 65.7% of all possible combinations can be separated by marital-status. This is an attribute that can be considered private/personal information about individuals, which explains its classification as sensitive.
education	Quasi-identifying (0.053%-distinction; 80.7%-separation)	Education is a 0.053%-distinct attribute, which means there are few distinct values and a 80.7%-separation attribute, which means 80.7% of all possible combinations can be separated by education. Through the combination of education with other attributes (per example occupation), we almost achieve individual identification.
native-country	Quasi-Identifying (0.14%-distinction; 16.8%-separation)	Native country is a 0.14%-distinct attribute, which means there are few distinct values, and a 16.8%-separation attribute, which means 16.8% of all possible combinations can be separated by native-country. Through the combination of education with other attributes (per example education), we almost achieve individual identification.
workclass	Quasi-Identifying (0.023%-distinction; 43.8%-separation)	Workclass is a 0.023%-distinct attribute, which means there are few distinct values, and a 43.8%-separation attribute, which means 43.8% of all possible combinations can be separated by workclass. Through the combination of education with other attributes (per example age), we almost achieve individual identification.

occupation	Quasi-Identifying (0.046%-distinction; 89.5%-separation)	Occupation is an 0.046%-distinct attribute, which means there are few distinct values, and a 89.5%-separation attribute, which means 89.5% of all possible combinations can be separated by occupation. Through the combination of education with other attributes (per example age), we almost achieve individual identification.
salary-class	Sensitive (0.007%-distinction; 37.4%-separation)	Salary-class is a 0.007%-distinct attribute, which means there are very few distinct values, and a 37.4%-separation attribute, which means 37.4% of all possible combinations can be separated by salary-class. In spite of not being possible to identify individuals through their salary-class, due to the fact that the attribute is only divided in 2 different labels: “<=50K” and “>50K”, it can be considered private/personal information about individuals, which explains its classification as sensitive.

Fig. 1 - Classification of each attribute

4. Characterization/analysis of the privacy risks of the dataset in original form

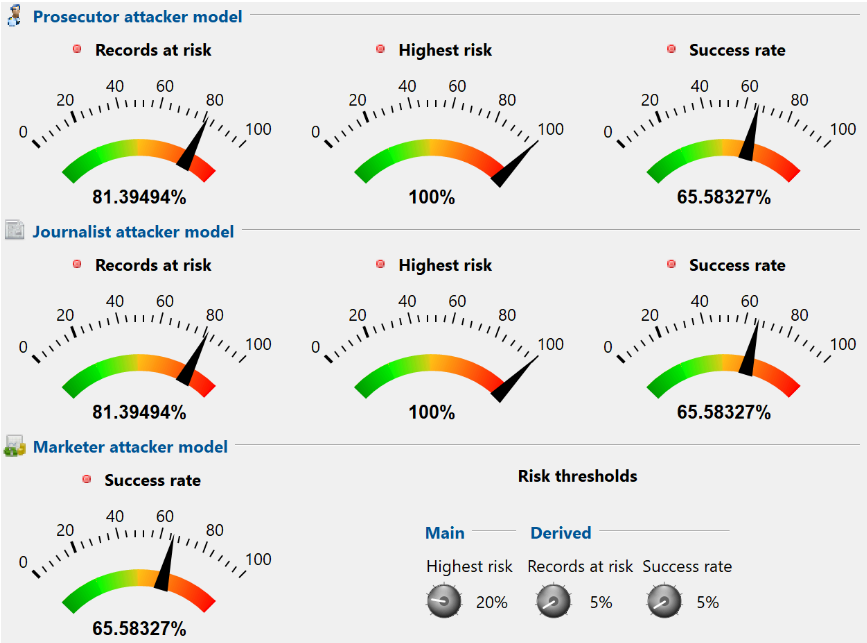
To analyze the dataset in the original format, we started by visualizing the risk levels of the 3 attacker models (prosecutor, journalist and marketer attacker model), provided by the ARX software. The threshold values used for this evaluation are 20%, 5% and 5% for highest risk, records at risk and success rate, respectively.

Prosecutor attacker model: This model targets one specific individual in the anonymized dataset and the adversary knows whether the target individual is in the dataset. Additionally, the model is based on the assumption that the attacker has specific prior knowledge about some individuals within the dataset.
Records at risk represent the proportion of records with risk above the threshold. The high percentage value (81.39%) indicates a significant vulnerability in the anonymized dataset, meaning a large number of records is in risk of being re-identified, compromising the individuals' privacy. The highest risk for a single record reaches the maximum in our scale. Finally, the average probability that a record in the anonymized dataset can be successfully re-identified, known as success rate, is 65.58%, which suggests a significant vulnerability and that more than two-thirds of the records in the dataset can be successfully re-identified by an attacker with prior knowledge.

Journalist attacker model: This model targets a specific individual, however, it is not expected that the attacker possesses background knowledge about membership. The percentage value of records at risk is equally high (81.39%), also indicating a significant vulnerability. The value of the highest risk for a single record is maximum and the success rate is similar to the one verified in the previous attacker model (65.58%).

Marketer attacker model: This model does not target specific individuals, but aims at re-identifying a high number of individuals (group of individuals). An attack can therefore only be considered successful if a larger fraction of the records could be re-identified. The average probability that a record in the anonymized dataset can be successfully re-identified is 65.58%, which suggests a significant vulnerability and that more than two-thirds of the records in the dataset can be successfully re-identified by an attacker with prior knowledge.

These results were expected, since the dataset hasn't had any anonymization technique applied.



Prosecutor attacker model

Records at risk 81.39494%

Highest risk 100%

Success rate 65.58327%

Journalist attacker model

Records at risk 81.39494%

Highest risk 100%

Success rate 65.58327%

Marketer attacker model

Success rate 65.58327%

Risk thresholds

Main

Highest risk 20%

Derived

Records at risk 5%

Success rate 5%

Fig. 2 - Re-identification risk of the original Dataset

5. Analysis of the performance of each privacy model applied to the dataset

- K-Anonymity

K-Anonymity is a privacy model used to protect individuals' identities in a dataset. It ensures that each record is indistinguishable from at least k-1 other records regarding certain identifying attributes, known as quasi-identifiers. This means that for any combination of these quasi-identifiers, there are at least k individuals with the same set of attributes, thereby making it difficult to re-identify any single person within the dataset.
The primary goal of k-anonymity is to prevent privacy breaches while maintaining data utility for analysis.
The value of k is chosen based on several considerations that balance the trade-offs between privacy and data utility.

- L-Diversity

L-Diversity is a privacy model that can be used to protect data against attribute disclosure by ensuring that each sensitive attribute has at least ℓ "well represented" values in each equivalence class. The idea of this model claims sensitive attributes must be “diverse” within each quasi-identifier equivalence class.
The value of ℓ is chosen intending to enhance the level of privacy k-anonymity provides.

Privacy models application:

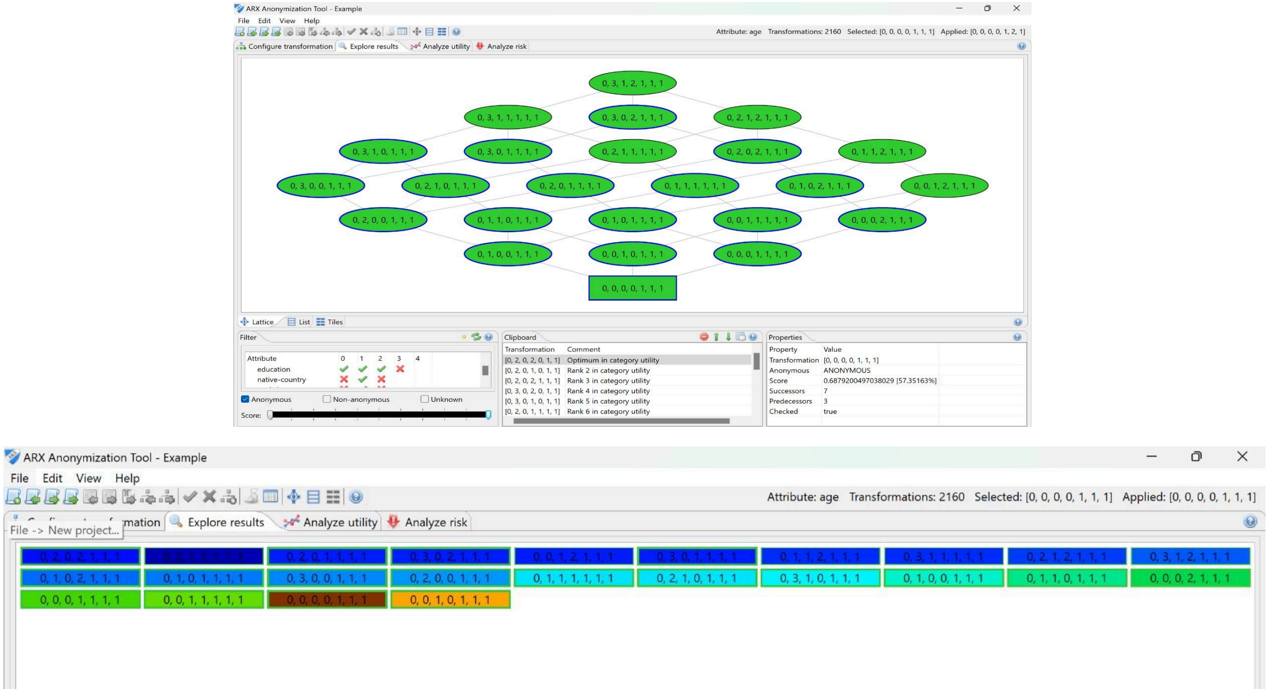
We chose to apply 2 privacy models, one used to Quasi-Identifier attributes (K-Anonymity) and another used to Sensitive attributes (L-Diversity).

- Re-identification risk of the anonymized dataset:

The models’ application has a significant impact on the dataset risks. Comparing the values of the image above with the original dataset risk levels, we can verify that in general all of our risk values decreased in all of the attacker models described in the previous section. In the anonymized dataset, we obtain no records at risk (0%) and the highest risk for a single record does not exceed percentage values of 20%, in both prosecutor and journalist attacker models. The success rate, when measured in the 3 models, is about 7.67%.

These results suggest that applying both 5-Anonymity and 3-Diversity privacy models to the original dataset reduced the privacy risk associated with the records, improving dataset anonymity.

The picture below displays a subset of our solution space, after applying the privacy models and after filtering the hierarchy levels.



Attribute: age Transformations: 2160 Selected: [0, 0, 0, 0, 1, 1, 1] Applied: [0, 0, 0, 0, 1, 2, 1]

Attribute	0	1	2	3	4
education	✓	✓	✓	✗	
native-country	✗	✓	✗		

☑ Anonymous ☐ Non-anonymous ☐ Unknown

Score:

Transformation	Comment
[0, 2, 0, 2, 0, 1, 1]	Optimum in category utility
[0, 2, 0, 1, 0, 1, 1]	Rank 2 in category utility
[0, 2, 0, 2, 1, 1, 1]	Rank 3 in category utility
[0, 3, 0, 2, 0, 1, 1]	Rank 4 in category utility
[0, 3, 0, 1, 0, 1, 1]	Rank 5 in category utility
[0, 2, 0, 1, 1, 1, 1]	Rank 6 in category utility

Property	Value
Transformation	[0, 0, 0, 0, 1, 1, 1]
Anonymous	ANONYMOUS
Score	0.6879200497038029 [57.35163%]
Successors	7
Predecessors	3
Checked	true

Attribute: age Transformations: 2160 Selected: [0, 0, 0, 0, 1, 1, 1] Applied: [0, 0, 0, 0, 1, 1, 1]

0, 2, 0, 2, 1, 1, 1	[illegible]	0, 2, 0, 1, 1, 1, 1	0, 3, 0, 2, 1, 1, 1	0, 0, 1, 2, 1, 1, 1	0, 3, 0, 1, 1, 1, 1	0, 1, 1, 2, 1, 1, 1	0, 3, 1, 1, 1, 1, 1	0, 2, 1, 2, 1, 1, 1	0, 3, 1, 2, 1, 1, 1
0, 1, 0, 2, 1, 1, 1	0, 1, 0, 1, 1, 1, 1	0, 3, 0, 0, 1, 1, 1	0, 2, 0, 0, 1, 1, 1	0, 1, 1, 1, 1, 1, 1	0, 2, 1, 0, 1, 1, 1	0, 3, 1, 0, 1, 1, 1	0, 1, 0, 0, 1, 1, 1	0, 1, 1, 0, 1, 1, 1	0, 0, 0, 2, 1, 1, 1
0, 0, 0, 1, 1, 1, 1	0, 0, 1, 1, 1, 1, 1	0, 0, 0, 0, 1, 1, 1	0, 0, 1, 0, 1, 1, 1						

Fig. 3 and 4 - Solution Space after applying the models

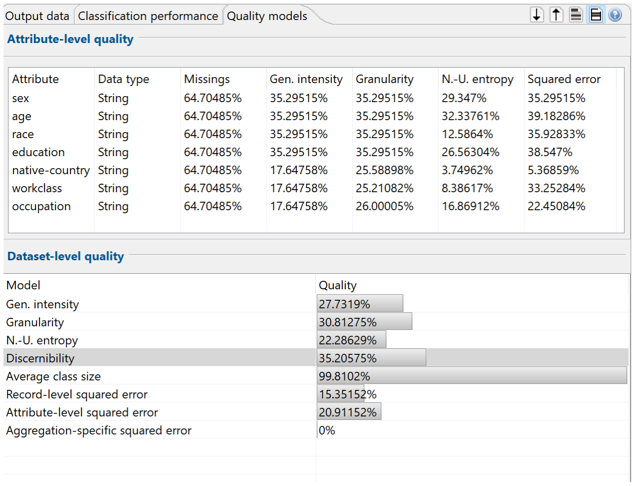
The selected transformation in figure 3 corresponds to the first one with the orange background in figure 4 (also selected). Orange denotes the transformation that is optimal regarding the Information Loss. Considering that the anonymized score is not very high, the data can maintain a certain level of quality of data.

This transformation is described as [0,0,0,0,1,1,1]. These numbers represent the hierarchy levels of each QID attribute, meaning that higher the hierarchy, more generalization.

- Utility level of the anonymized dataset:

As anonymization can lead to information loss, we have to take into account the amount of information that stays available when anonymizing our dataset. This is measured by utility, based on utility metrics.

After applying our models, we can measure the utility level of the anonymized dataset. Discernibility measures the size of groups of indistinguishable records and introduces a penalty for records which have been completely suppressed. Here, we obtained a value of 35.2%, which means a significant portion of the original data's information content has been preserved, allowing for meaningful analysis while still providing a high level of privacy protection. Another metric that can be used to measure utility in our dataset is Average class size. This metric measures the average size of groups of indistinguishable records. Our average class size is 99.8%.



Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
sex	String	64.70485%	35.29515%	35.29515%	29.347%	35.29515%
age	String	64.70485%	35.29515%	35.29515%	32.33761%	39.18286%
race	String	64.70485%	35.29515%	35.29515%	12.5864%	35.92833%
education	String	64.70485%	35.29515%	35.29515%	26.56304%	38.547%
native-country	String	64.70485%	17.64758%	25.58898%	3.74962%	5.36859%
workclass	String	64.70485%	17.64758%	25.21082%	8.38617%	33.25284%
occupation	String	64.70485%	17.64758%	26.00005%	16.86912%	22.45084%

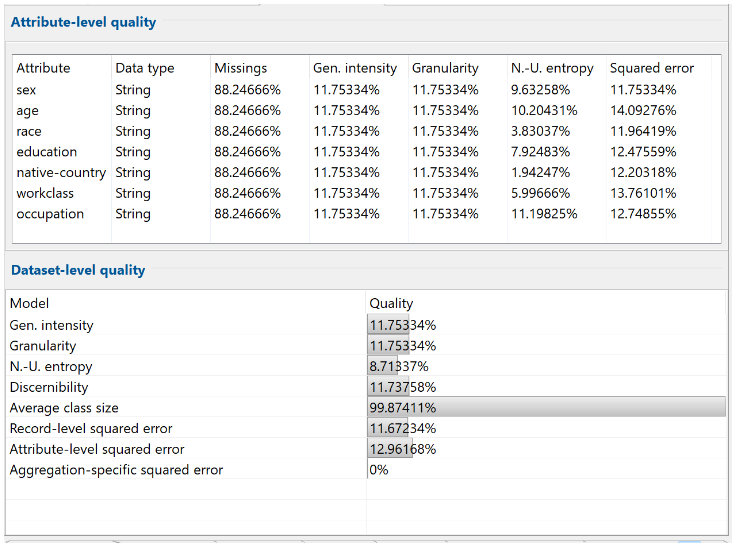
Dataset-level quality

Model	Quality
Gen. intensity	27.7319%
Granularity	30.81275%
N.-U. entropy	22.28629%
Discernibility	35.20575%
Average class size	99.8102%
Record-level squared error	15.35152%
Attribute-level squared error	20.91152%
Aggregation-specific squared error	0%

Fig. 5 - Utility metrics of our anonymized dataset

- The effect of the level of suppression and coding model (favor generalization vs suppression) on the results

When we use coding model to favor suppression over generalization, we obtain lower values for our utility metrics, except for the Average class size that was maintained. This happens because suppression promotes information loss by prioritizing the removal of specific data values rather than broadening them into ranges or categories. This approach leads to a higher degree of information loss since individual data points are more likely to be eliminated from the dataset. As a result, the utility metrics, which measure the usefulness or information content of the data, tend to decrease.



Attribute-level quality

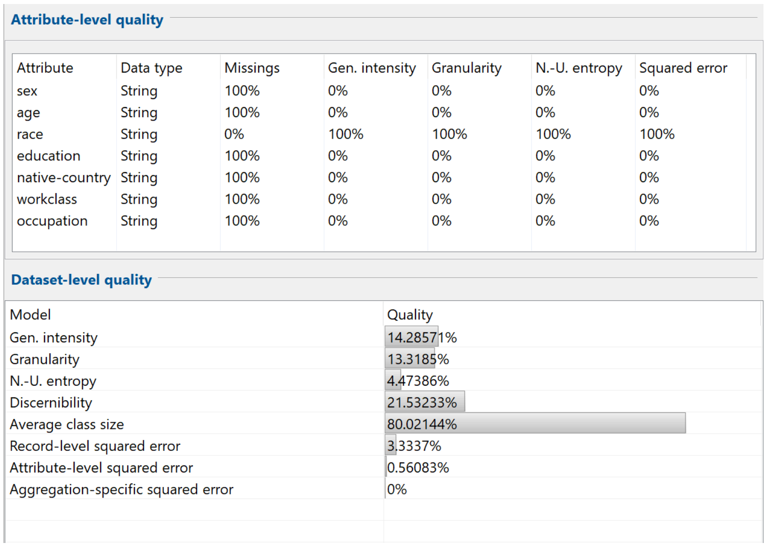
Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
sex	String	88.24666%	11.75334%	11.75334%	9.63258%	11.75334%
age	String	88.24666%	11.75334%	11.75334%	10.20431%	14.09276%
race	String	88.24666%	11.75334%	11.75334%	3.83037%	11.96419%
education	String	88.24666%	11.75334%	11.75334%	7.92483%	12.47559%
native-country	String	88.24666%	11.75334%	11.75334%	1.94247%	12.20318%
workclass	String	88.24666%	11.75334%	11.75334%	5.99666%	13.76101%
occupation	String	88.24666%	11.75334%	11.75334%	11.19825%	12.74855%

Dataset-level quality

Model	Quality
Gen. intensity	11.75334%
Granularity	11.75334%
N.-U. entropy	8.71337%
Discernibility	11.73758%
Average class size	99.87411%
Record-level squared error	11.67234%
Attribute-level squared error	12.96168%
Aggregation-specific squared error	0%

Fig. 6 - Utility measures after favoring suppression

When we use coding model to favor generalization over suppression, we also obtain lower values for the utility metrics. (for example, discernibility value decreases to 11.7%. With generalization, the data values get distributed into more generalized categories or ranges. This results in a decrease in utility metrics due to the loss of detailed information.



Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
sex	String	100%	0%	0%	0%	0%
age	String	100%	0%	0%	0%	0%
race	String	0%	100%	100%	100%	100%
education	String	100%	0%	0%	0%	0%
native-country	String	100%	0%	0%	0%	0%
workclass	String	100%	0%	0%	0%	0%
occupation	String	100%	0%	0%	0%	0%

Dataset-level quality

Model	Quality
Gen. intensity	14.28571%
Granularity	13.3185%
N.-U. entropy	4.47386%
Discernibility	21.53233%
Average class size	80.02144%
Record-level squared error	3.3337%
Attribute-level squared error	0.56083%
Aggregation-specific squared error	0%

Fig. 7 - Utility measures after favoring generalization

Therefore, we can conclude that the most accurate coding model is one that strikes a balance between suppression and generalization. By combining elements of both approaches, this balanced coding model aims to optimize both anonymity and utility metrics, thereby achieving an effective compromise between privacy protection and data usability.

6. Conclusions

In this assignment, we analyzed the anonymation of a dataset with the ARX anonymization tool to balance the trade-offs between data utility and privacy protection. Our analysis consisted in identifying the dataset’s attributes into 4 categories (Identifying, QIDs, Sensitive and Insensitive). Additionally, we applied k-anonymity and l-diversity models to protect individual identities while preserving the dataset's usability for meaningful analysis.

Our initial risk analysis of the original dataset highlighted significant vulnerabilities, with high re-identification risks across all attacker models (prosecutor, journalist, and marketer). This underscored the necessity for robust anonymization strategies to safeguard personal information. By implementing 5-anonymity and 3-diversity models, we observed a significant reduction in privacy risks.

In terms of utility, we found that anonymization inevitably leads to information loss and concluded that a balanced approach combining both suppression and generalization techniques is optimal. This balanced coding model effectively minimizes re-identification risks while maximizing the utility of the anonymized dataset.

In conclusion, the anonymization of datasets is a critical process for ensuring data privacy in today's data-driven world.

7. References

https://arx.deidentifier.org/anonymization-tool/analysis/

https://youtu.be/N8I-sxmMfqQ?feature=shared