U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Anonymization of a Dataset with Utility and Risk Analysis

Rui Coelho, up202106772

Sérgio Coelho, up202107951

Table of Contents

Introduction 2
ARX Software 2
Dataset 3
  Dataset Analysis 3
  Attributes Analysis 3
  Sensitivity and Attacker Incentives 7
Risk Analysis 8
  Prosecution Attacker Model 8
  Journalist Attacker Model 8
  Marketer Attacker Model 8
Privacy Models 9
  K-Anonymity 9
  δ-Disclosure Privacy 10
Data Assessment 10
Application of the Privacy Models 12
  Risk Analysis 12
  Utility Analysis 14
  Suppression vs Generalization 15
Conclusion 16

Introduction

In an era where data-driven insights are increasingly valuable, protecting the privacy of individuals represented in datasets has become a critical concern. Data anonymization offers a mechanism to publish or share datasets in a way that minimizes the risk of re-identification, while retaining the utility necessary for meaningful analysis.

The primary objective of this assignment is to explore the trade-off between data privacy and data utility by applying privacy-preserving transformations to an example dataset provided by the ARX project. The dataset is initially assessed in its original form to identify potential privacy risks. Each attribute is classified according to its role in re-identification, and privacy risks are evaluated using attacker models available in ARX.

Subsequently, two different privacy models are applied to the dataset. The effectiveness of each model is analyzed in terms of risk reduction and utility preservation. We investigate how varying the parameters of these models affect the balance between data utility and re-identification risk.

ARX Software

ARX is an open-source software designed for data anonymization and privacy-preserving data transformations. It supports a wide variety of privacy and risk models and methods to analyze and transform data.

This software allows users to import data from different sources, such as CSV files, MS Excel and SQL Databases. It also supports exporting the anonymizes data back into these formats, enabling easy integration with other tools and workflows.

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

Dataset

The analyzed dataset in this report is the one provided in the official ARX website, which can be found in their downloads page.

Dataset Analysis

The dataset under analysis includes various attributes that describe personal and demographic aspects of individuals. Each column represents a distinct characteristic, and most are easily understood based on their names, so detailed explanations are unnecessary. The included attributes are:

- Sex
- Age
- Race
- Marital Status
- Education
- Native Country
- Work Class
- Occupation
- Salary Class

Due to the nature of the data, many of these attributes have the potential to either directly or indirectly reveal someone's identity. When several quasi-identifiers are combined, they can significantly increase the risk of re-identification.

Altogether, these variables contribute to forming a comprehensive profile of each individual in the dataset. If left unanonymized, this information could be misused, leading to privacy breaches, discrimination, or other forms of personal risk. Therefore, it is necessary to apply privacy-preserving techniques to reduce these risks while still maintaining the dataset's analytical value.

Attributes Analysis

Before proceeding with the risk analysis, it is essential to categorize the dataset's attributes according to their role in privacy exposure. Below is a brief overview of each attribute type and its relevance in the anonymization process:

Identifying Attributes – These attributes have the potential to directly reveal the identity of an individual. Because they pose a high privacy risk, they are typically removed or strongly modified during anonymization to avoid explicit identification.

Quasi-Identifying Attributes (QIDs) – QIDs may not individually identify someone, but when combined with other data points, they can enable re-identification. These attributes are key in anonymization processes and require strategies like generalization or suppression to mitigate linkage risks.

Sensitive Attributes – These are fields that contain confidential or personally sensitive information. Although they do not directly identify someone, their exposure can lead to discrimination or harm. It is important to protect these attributes to ensure ethical data use and legal compliance.

Insensitive Attributes – These attributes hold information that is not likely to compromise an individual's privacy. Since they do not present significant risks, they generally don't require any specific anonymization steps.

Before proceeding with the risk analysis, it is necessary to categorize the dataset's attributes according to their role in potential re-identification. To do this, we rely on two key metrics provided by ARX: distinction and separation. **Distinction** measures the uniqueness of values in an attribute, that is, how likely it is to distinguish individuals based solely on that attribute. **Separation** refers to how effectively an attribute contributes to splitting the dataset into distinct groups, which influences its potential to identify individuals when combined with other data.

Based on these values and the nature of the information each attribute conveys, we classify them into one of four categories: Identifiers, Quasi-Identifiers, Sensitive Attributes, and Insensitive Attributes. The following table summarizes this classification along with a justification grounded in both the statistical metrics and the privacy relevance of the attributes.

Attributes	Classification	Justification
sex	Quasi-Identifying (0.04%-distinction; 40.17%-separation)	Sex has a very low distinction value (0.04%), indicating that it has very few unique values. However, its separation value is relatively high (40.17%), which shows that this attribute contributes significantly to splitting the dataset into different groups. While it cannot identify someone on its own, it becomes useful in combination with other

		quasi-identifiers, justifying its classification as quasi-identifying.
age	Quasi-Identifying (0.08%-distinction; 42.97%-separation)	Age shows a slightly higher distinction (0.08%) due to a wider range of possible values (especially if numerical). Its separation of 42.97% indicates it plays a notable role in dividing individuals into distinct categories. This high separation potential, combined with moderate uniqueness, makes it a quasi-identifier.
race	Quasi-Identifying (0.06%-distinction; 12.97%-separation)	The very low distinction (0.06%) implies that race rarely identifies individuals on its own. This makes sense given that datasets often include only a few race categories (e.g., 3–6), and many people share the same value. As a quasi-identifier, this level of separation justifies its inclusion in generalization strategies: it contributes to identification when combined with other attributes like age.
marital status	Quasi-Identifying (0.04%-distinction; 46.21%-separation)	This attribute has very low uniqueness (0.04%) but a relatively high separation value (46.21%), meaning it can divide the dataset effectively when used with other attributes. Since it does not identify someone directly but increases the

		re-identification risk when combined, it's rightly classified as quasi-identifying.
education	Quasi-Identifying (0.06%-distinction; 50.96%-separation)	Education has low distinction (0.06%) but a high separation score (50.96%). This means that while there aren't many unique values, education helps in segmenting the data into different subgroups. It supports identifying individuals when combined with other quasi-identifiers, thus quasi-identifying.
native country	Quasi-Identifying (0.06%-distinction; 1.57%-separation)	Although native country has moderate distinction (0.06%), its separation is very low (1.57%), meaning it rarely helps distinguish records in this dataset. Still, it may assist in certain combinations for re-identification and typically gets treated as quasi-identifying for conservative anonymization.
work class	Sensitive (0.04%-distinction; 25.15%-separation)	The very low distinction value (0.04%) indicates that work class rarely makes individual records unique. The moderate separation value (25.15%) shows that work class provides a reasonable ability to differentiate between groups.

occupation	Sensitive (0.06%-distinction; 66.62%-separation)	The very low distinction value (0.06%) indicates that occupation alone does not make individual records unique in the dataset. he high separation value (66.62%) shows that occupation is highly effective at distinguishing groups within the data. This means that it provides an important analytical value.
salary class	Sensitive (0.04%-distinction; 44.96%-separation)	Despite its high separation (44.96%), salary class has very low distinction (0.04%) and represents only two categories. It contains personal financial information, which is often considered confidential and could lead to discrimination. It's thus rightly labeled as sensitive.

To sum up, we can divide the attributes in two main groups which are the **sensitive attributes** like occupation, work class and salary class and the **quasi-identifier attributes** which are sex, age, marital status, education, native country, work class and occupation.

Sensitivity and Attacker Incentives

From an attacker's perspective, attributes such as **work class**, **occupation**, and **salary class** are highly valuable targets. While these attributes, for most cases, don't identify individuals on their own, they reveal critical and crucial socioeconomic information that can be exploited for malicious purposes. For example, this data can help attackers infer a person's income level, job stability, or access to financial resources, all of which can be used to facilitate fraud, identity theft, or targeted scams.

Because these attributes increase the exposure of individuals to financial and social risks, they must be treated as **Sensitive** within any privacy-preserving strategy, regardless of their low distinction values.

Risk Analysis

The ARX software evaluates re-identification risk using three distinct attacker models, each simulating a different type of threat scenario. Understanding these models is crucial for selecting appropriate anonymization strategies and ensuring robust privacy protection.

Prosecution Attacker Model

This model assumes that the attacker has prior knowledge about a specific individual and is attempting to locate that person's record in the anonymized dataset. The attack is considered successful if the attacker can narrow down the search to a single or small number of candidate records. This model simulates scenarios where the attacker has a personal interest in a specific target and highlights the risk of record-level re-identification

Journalist Attacker Model

In contrast to the Prosecutor, the Journalist model represents an attacker who does not start with a specific individual in mind. Instead, the attacker explores the dataset to uncover any identities or sensitive patterns that may be newsworthy or revealing. This model is more opportunistic and simulates real-world risks in datasets released to the public, where unexpected disclosure of sensitive information can occur even without targeting individuals directly.

Marketer Attacker Model

The Marketer model assumes the attacker is not interested in individuals but rather in profiling groups or segments of the population for purposes like advertising or market analysis. A re-identification attempt is only considered successful if it applies to a larger portion of the dataset. This model focuses on group-level privacy risks, raising concerns about how datasets can be exploited to draw inferences about categories of people (e.g., by age, income, or location), even if no single person is fully identified.

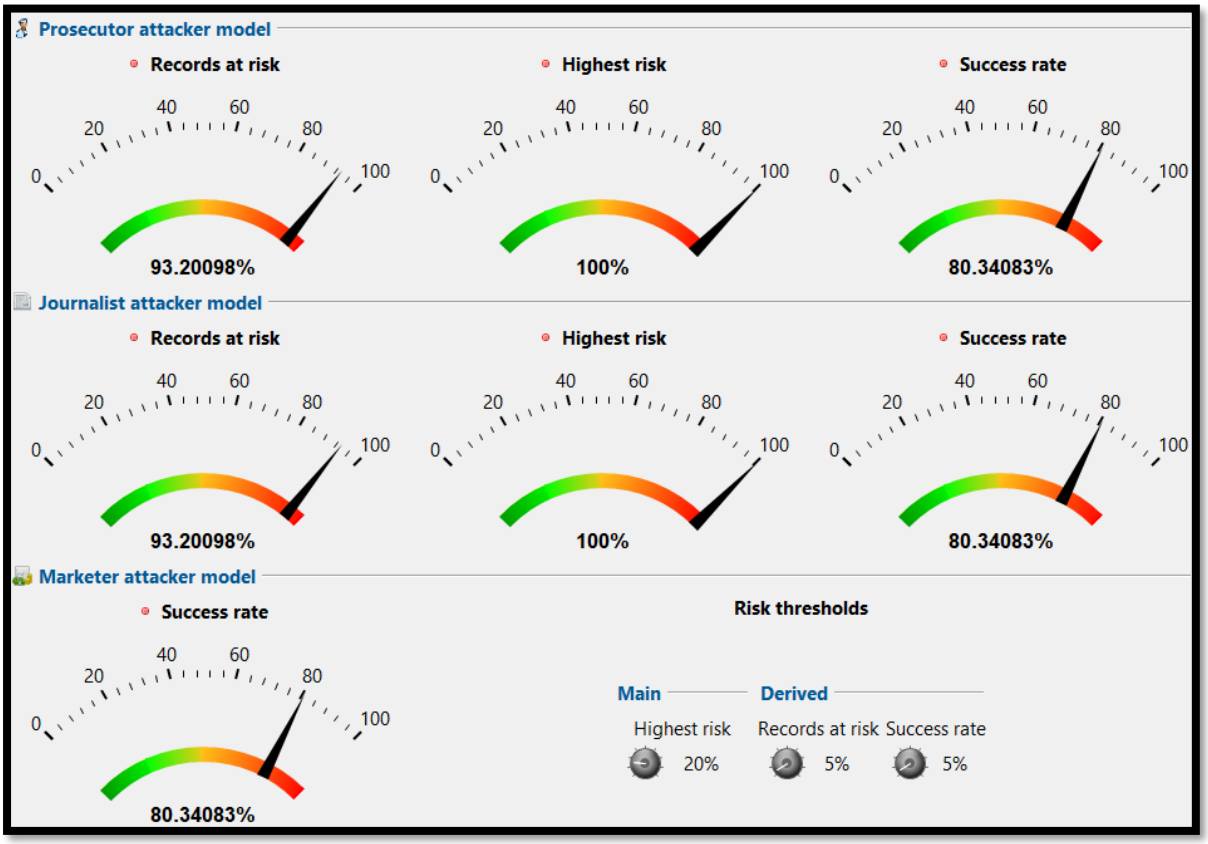


Figure 1 - Attacker Models and its risks before applying the privacy models

Privacy Models

To reduce the re-identification risks identified in the original dataset, we applied two well-established privacy models using ARX**: K-Anonymity** and **Differential Privacy**. These models follow different principles and provide complementary approaches to protecting sensitive data.

K-Anonymity

k-Anonymity is a foundational privacy model designed to protect against re-identification in scenarios like the Prosecutor attacker model. A dataset satisfies k-anonymity if every record is indistinguishable from at least k−1 other records when considering the quasi-identifiers. In other words, each combination of quasi-identifiers must appear in at least k records. These records form equivalence classes, ensuring that no single individual can be uniquely identified within a group. While effective against identity disclosure, k-anonymity does not inherently protect against attribute disclosure or linkage attacks involving background knowledge beyond quasi-identifiers.

δ -Disclosure Privacy

δ -Disclosure Privacy is a privacy model designed to protect datasets against attribute disclosure, focusing on limiting how much information an attacker can infer about sensitive attributes. Similar to t-closeness, it constrains the difference between the distribution of sensitive values within each equivalence class and the distribution in the overall dataset. However, δ -disclosure privacy adopts a multiplicative approach to measure these differences, making it a stricter and more conservative model than t-closeness.

Data Assessment

Before applying the privacy models, we analysed the age distribution in the dataset and considered redistributing the ages into multiple intervals. This approach aimed to enhance privacy while achieving a more homogeneous distribution of the data.

This analysis supported our decision to group the age attribute into intervals, as it revealed significant variation across individual values. By creating age ranges, we improved both data privacy and consistency, laying a better foundation for the application of privacy models.

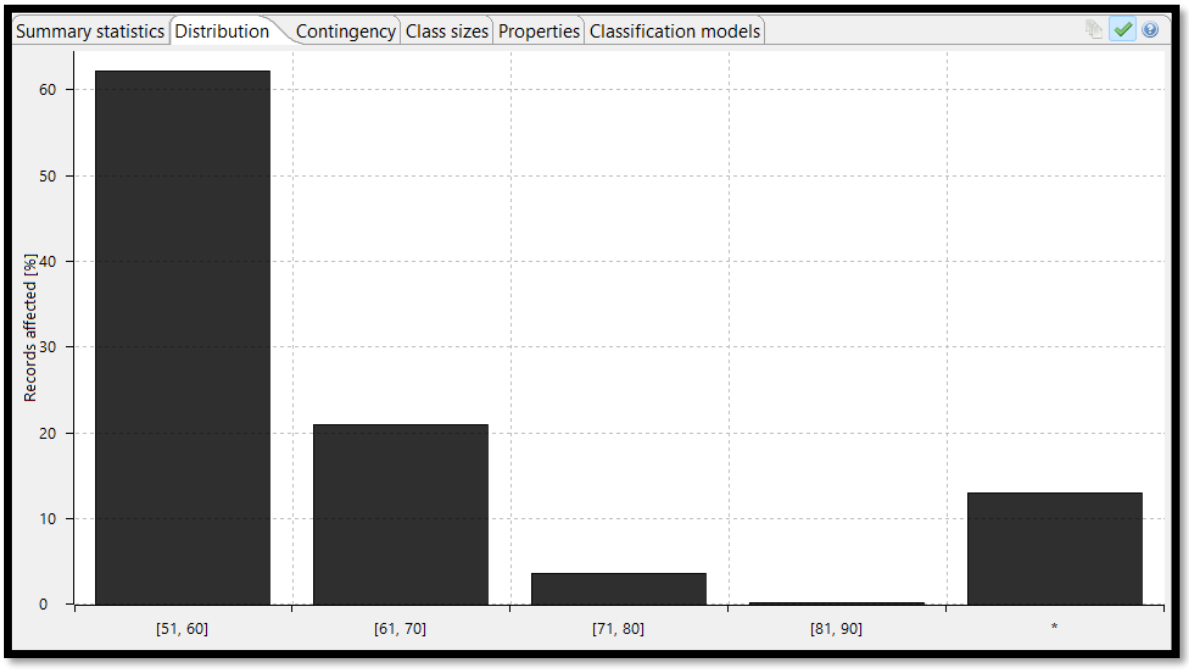


Figure 2 - Age distribution after applying the hierarchy using intervals

From a privacy perspective, we also chose to anonymize the sex attribute in the database. This means that the sex of each individual was not directly disclosed. Instead, individuals were grouped into a group {Male, Female} according to a predefined hierarchy for classification and ordering.

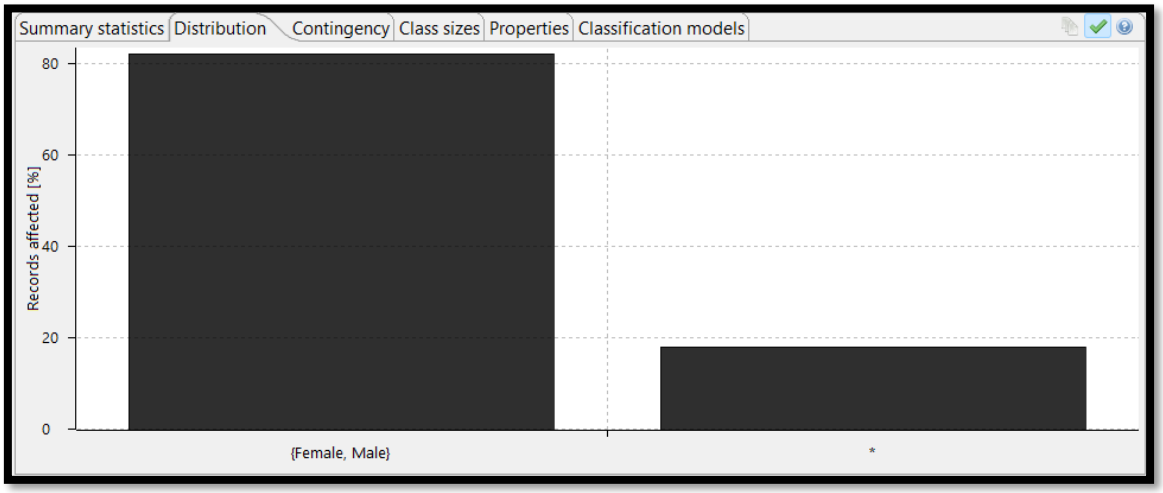


Figure 3 - Distribution of sex after applying the hierarchy for classification and ordering

After having the quasi-identifying and sensitive attributes correctly assigned, we chose to adjust the default weights assigned to the quasi-identifying attributes.

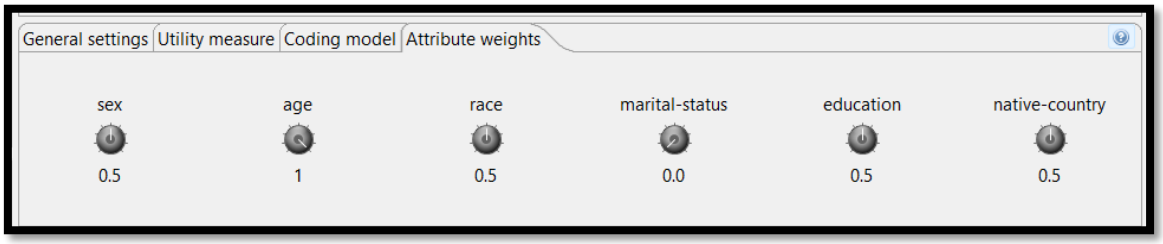


Figure 4 – Quasi-identifying Attributes and its weights

This decision was grounded in the understanding that, statistically, more experienced employees tend to earn higher salaries than recent hires (*According to a 2023 report by Pew Research Center, older workers tend to earn significantly more than younger workers, with median hourly wages rising from $13 in 1987 to $22 in 2022 for workers aged 65 and over,* Pew Reasearch 2023). Recognizing this relationship is crucial because an attacker aiming to extract sensitive information would prioritize attributes that reveal the most valuable insights, such as potential income levels associated with age and experience.

Application of the Privacy Models

After applying the Privacy Models, K-Anonymity for Quasi-identifying values (sex, age, marital-status, education, native-country,race) and δ-Disclosure Privacy for sensitive values (salary-class, occupation and workclass), we reached multiple conclusions and results.

Risk Analysis

The figure below presents the results of the risk analysis using three attacker models.

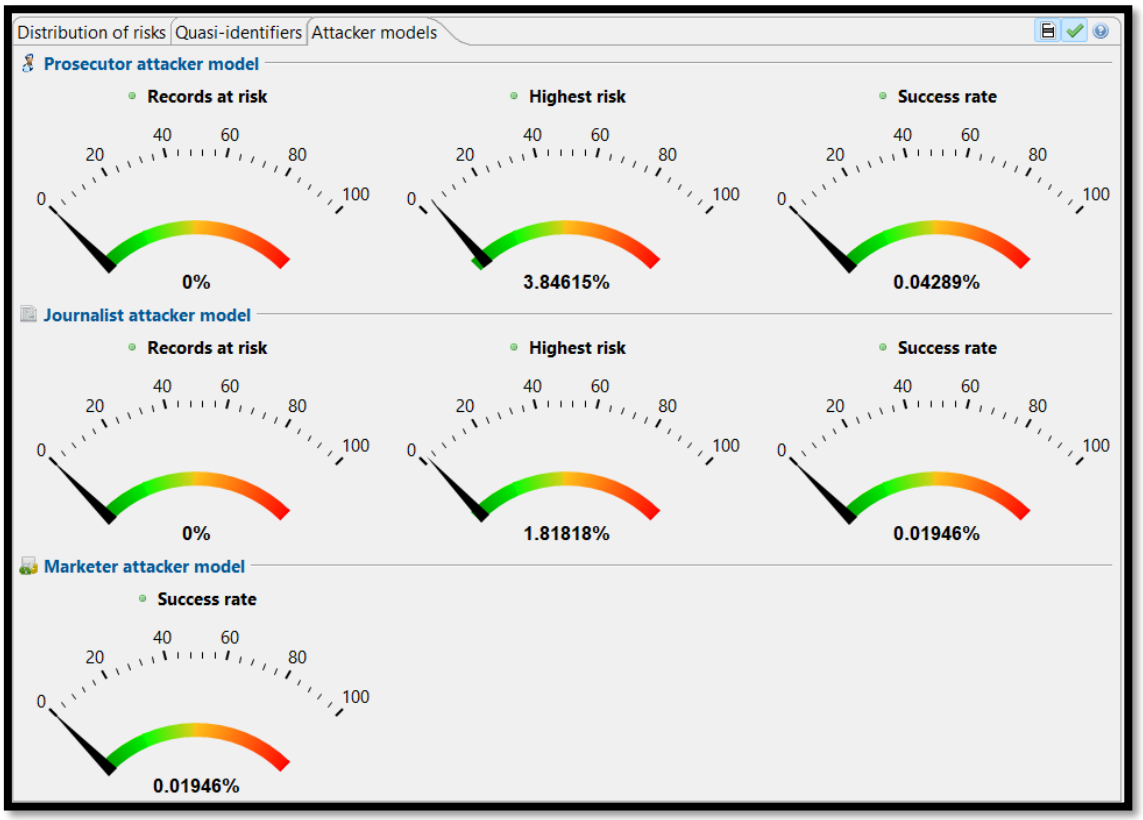
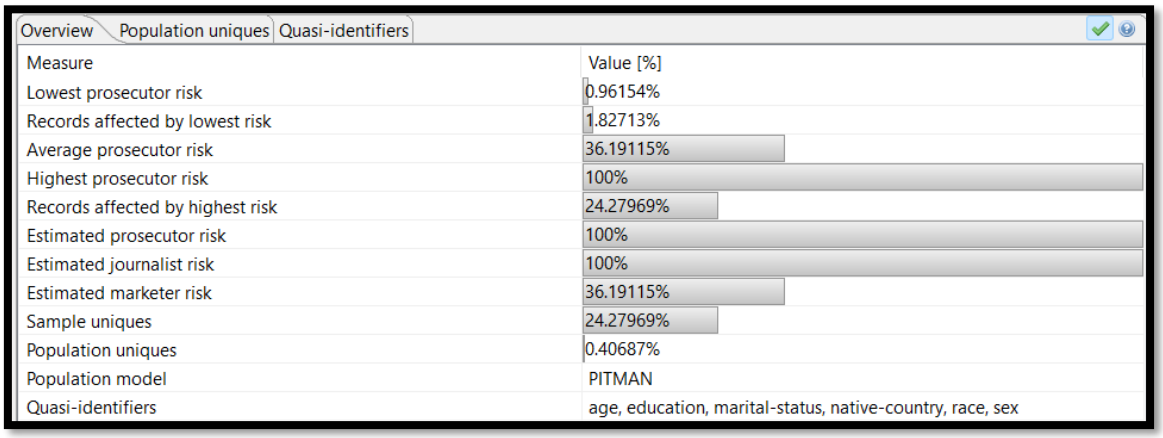


Figure 5 - Attacker models and its risks after applying the privacy models.

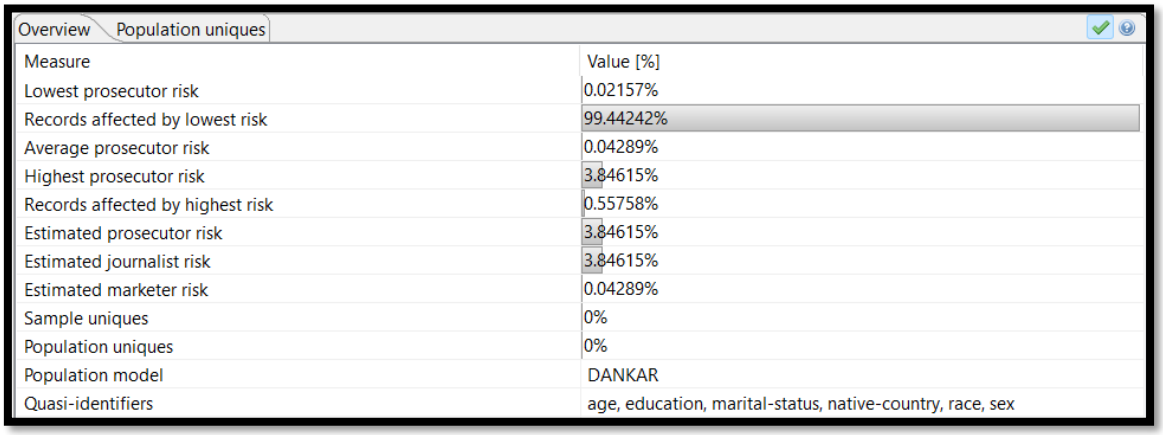
As shown in the image, the risk levels across all attacker models, Prosecutor, Journalist, and Marketer, are extremely low. Specifically, both the Prosecutor and Journalist models report 0% of records at risk, while the highest individual risk remains below 4%, and success rates for all models are negligible (under 0.05%). These results demonstrate that the dataset has been successfully anonymized. To achieve this level of privacy, we had to apply both generalization and suppression techniques. After conducting several tests and analyses, we chose to prioritize suppression over generalization, as it allowed us to better preserve the semantic integrity of the data while still minimizing disclosure risk. Nevertheless, this approach came at the cost of some data precision, and we had to carefully manage the trade-off between privacy and utility throughout the anonymization process.

As we can see in the images below (Figure 6 and Figure 7), after attributing the correct data type, we can compare the differences between the initial dataset and the dataset after applying the transformations.



Measure	Value [%]
Lowest prosecutor risk	0.96154%
Records affected by lowest risk	1.82713%
Average prosecutor risk	36.19115%
Highest prosecutor risk	100%
Records affected by highest risk	24.27969%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	36.19115%
Sample uniques	24.27969%
Population uniques	0.40687%
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, native-country, race, sex

Figure 6 - Overview before applying the transformations



Measure	Value [%]
Lowest prosecutor risk	0.02157%
Records affected by lowest risk	99.44242%
Average prosecutor risk	0.04289%
Highest prosecutor risk	3.84615%
Records affected by highest risk	0.55758%
Estimated prosecutor risk	3.84615%
Estimated journalist risk	3.84615%
Estimated marketer risk	0.04289%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, native-country, race, sex

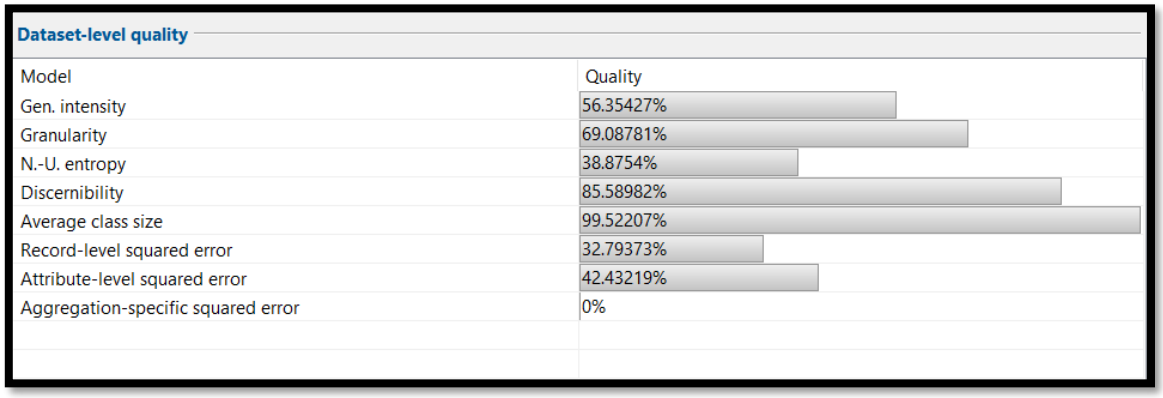
Figure 7 - Overview after applying after transformations

By comparing both overviews, it is clear that the risk percentages, after the transformations were made, are extremely low which means the sensitive information is well protected. However, these changes imply that the dataset's anonymization strategy was adjusted to better protect sensitive information while potentially allowing a slight decrease in data utility, as indicated in the table by increased affected Records.

Utility Analysis

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

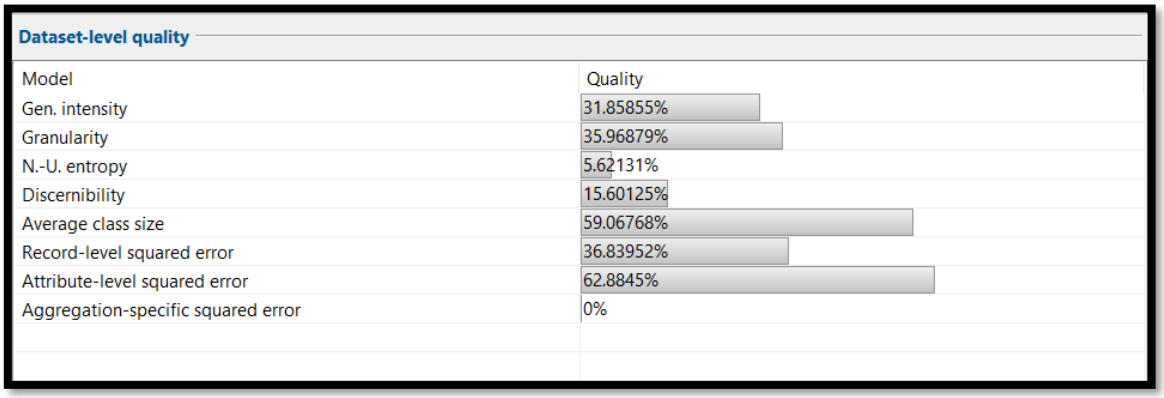
Quality metrics provide a systematic approach to evaluate how well anonymization methods maintain the utility of a dataset while ensuring individual privacy. This analysis will focus on several key metrics including Granularity and Discernibility, for example.



Dataset-level quality

Model	Quality
Gen. intensity	56.35427%
Granularity	69.08781%
N.-U. entropy	38.8754%
Discernibility	85.58982%
Average class size	99.52207%
Record-level squared error	32.79373%
Attribute-level squared error	42.43219%
Aggregation-specific squared error	0%

Figure 8 - Utility metrics before applying the transformations



Dataset-level quality

Model	Quality
Gen. intensity	31.85855%
Granularity	35.96879%
N.-U. entropy	5.62131%
Discernibility	15.60125%
Average class size	59.06768%
Record-level squared error	36.83952%
Attribute-level squared error	62.8845%
Aggregation-specific squared error	0%

Figure 9 - Utility metrics after applying transformations.

Applying anonymization transformations brought some clear and important changes to the dataset's overall quality. Before the transformation (first figure), the dataset had high levels of granularity (69.09%), discernibility (85.59%), and a very large average class size (99.52%), all of which suggested a rich and detailed dataset. At the same time, however, the data already showed notable record-level (32.79%) and attribute-level (42.43%) squared errors, which indicated that some level of information distortion was already present.

After anonymization (second figure), we noticed a drop in several utility metrics: granularity dropped to 35.97%, discernibility to 15.60%, and average class size was reduced to 59.07%. These changes reflect a significant reduction in data detail, which is expected when privacy protections are applied. What stood out, though, was that the record-level squared error actually increased slightly to 36.84%, and attribute-level error rose to 62.88%, a result that might seem surprising at first. This can be explained by our shift from using broad generalization to more focused suppression strategies. While generalization tends to blur data in a more uniform way, suppression can lead to more targeted, but deeper, distortions in specific areas.

In the end, these results show the kind of trade-offs we had to manage — between protecting individuals' privacy and keeping the data useful. Even though we lost some precision, the reduction in generalization intensity (from 56.35% to 31.86%) and the improvement in class size distribution suggest that we reached a thoughtful balance. The final dataset still holds value for analysis while offering much stronger privacy guarantees which was the main goal of this anonymization effort.

Suppression vs Generalization

During the anonymization process, we carried out a series of tests applying various levels of suppression and generalization to assess their respective impacts on privacy protection and data utility. Both techniques offer distinct benefits: generalization transforms attribute values into broader categories or ranges, which helps retain the dataset's analytical structure, whereas suppression removes specific values entirely, providing a stronger barrier against potential re-identification.

Throughout our experiments, we found that while generalization preserved more information, it often led to overly broad equivalence classes, which diminished the dataset's precision. Suppression, in contrast, although more severe in terms of information loss, delivered stronger privacy guarantees, particularly in scenarios involving high-risk attribute combinations.

Given these findings, we made a deliberate choice to favour suppression over generalization. This decision was based on its superior ability to mitigate disclosure risks and its closer alignment with the privacy goals of this assignment. Despite the resulting reduction in data granularity, the suppression-focused approach proved more effective in safeguarding sensitive information while still retaining a sufficient level of utility for analysis.

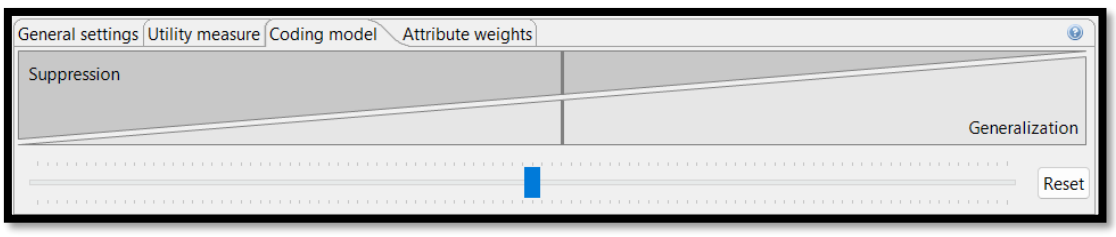


Figure 6 –Adjustment between suppression and generalization

Conclusion

Working on this project allowed us to better understand the real challenges involved in protecting personal data while still keeping it useful for analysis. From the beginning, we focused on classifying each attribute carefully, deciding what could identify someone, what should be treated as sensitive and what posed little risk. This step was key to applying the right privacy techniques.

We tested different privacy models, like k-anonymity and δ-disclosure privacy, and played with various settings to see how changes in suppression and generalization would affect both privacy and utility. In the end, we leaned more on suppression, since it helped us better protect individual privacy in more sensitive parts of the dataset, even though it came with some information loss.

The results showed clear improvements in reducing re-identification risk and maintaining reasonable data quality. Metrics like discernibility and average class size backed up our choices, and ARX's tools helped us make data-driven decisions throughout the process.

One of the biggest takeaways is how small changes, like adjusting weights or suppression limits, can make a big difference. We will definitely explore more privacy models and combinations to find even better results, in the future. Still, this project gave us valuable insights into data anonymization, and showed us how to balance protecting people's privacy with keeping datasets meaningful and usable.