Anonymization of a Dataset with Utility and Risk Analysis

Security and Privacy - Assignment #3
2024/2025

Carlos Ortega- up202303651
David Sá - up202303580
João Morais – up202307077

Index

1. Introduction

1.1. Goal of the assigment

This report presents the anonymization of a dataset using the ARX data anonymization tool. The goal is to evaluate how different privacy models affect both the privacy and utility of the data. Through the classification of attributes, risk analysis, and application of anonymization models, we assess the re-identification risk and measure the trade-off between protecting sensitive information and maintaining data usefulness. The dataset used is the ARX-provided example project, which includes personal and medical attributes representative of typical anonymization challenges.

1.2 Tool used (ARX)

To perform all steps of the assignment, we used ARX, a powerful and open-source data anonymization tool developed by the Technical University of Munich. ARX allows users to import datasets, classify attributes, apply privacy-preserving transformations (such as generalization and suppression), and evaluate privacy risks through attacker models. It also provides metrics to assess the utility of the anonymized data, helping to strike a balance between protecting privacy and maintaining data quality.

1.3 Dataset description

The dataset used is the built-in example project provided by ARX. It contains fictional personal records with attributes such as age, gender, ZIP code, marital status, and disease. These attributes simulate a realistic scenario where some are potentially identifying or quasi-identifying, while others are sensitive or neutral. The dataset is suitable for testing the effects of anonymization models and understanding how certain attributes can increase the risk of re-identification.

2. Attribute Classification

The classification of each attribute in the dataset according to its potential privacy risk is a fundamental step in the anonymization process. In ARX, attributes are assigned to one of four

categories, each playing a specific role in determining the appropriate anonymization strategy. Therefore, we categorized the dataset attributes according to their respective privacy classifications:

1. Identifying Attributes

Attributes that can directly reveal the identity of an individual, such as names, ID numbers, or email addresses. These are typically removed or heavily masked, as they present a direct risk of re-identification.

In the dataset used for this report, there are no identifying attributes such as names, national ID numbers, or other direct personal identifiers

2. Quasi-Identifying Attributes (QIDs)

Attributes that do not identify individuals on their own but may do so when combined with other information. Examples include age, gender, and ZIP code. These attributes are subject to generalization or suppression to prevent re-identification through data linkage.

Atributo	Justificação
sex	Em conjunto com outros atributos, permite distinguir grupos populacionais.
age	Valor contínuo com elevado poder de distinção.
education	O nível educacional ajuda a diferenciar grupos socioeconómicos.

3. Sensitive Attributes

Attributes containing private information that must be protected, such as medical diagnoses, salaries, or political affiliations. The primary objective is to ensure that this information remains confidential, even if a record is linked to an individual.

Atributo	Justificação
occupation	Pode indicar estatuto socioeconómico ou ocupações sensíveis (ex: domésticos).
salary-class	Contém dados económicos diretamente confidenciais (rendimento pessoal).
workclass	Pode refletir vínculos laborais precários ou revelar desemprego, o que é sensível.

4 Insensitive Attributes

Attributes that do not have a significant impact on privacy, such as generic or irrelevant information. These can usually remain unchanged, preserving the utility of the dataset for analysis.

The attributes that we placed in this class could also be considered quasi-identifiers, but to achieve better anonymization results, we decided to classify them as insensitive.

Atributo	Justificação
race	Informação genérica com baixa distintividade no conjunto de dados usado.
native-country	Informação geográfica ampla, pouco associável a indivíduos únicos no contexto analisado.
marital-status	Estado civil é comum, pouco distintivo e não revela informação sensível diretamente.

This classification is critical, as it determines which attributes will be generalized, suppressed, or retained during the anonymization process. Proper assignment of each attribute ensures that privacy risks are minimized while maintaining the analytical value of the data. The ARX interface facilitates this process, enabling organizations to implement effective and tailored privacy protections.

3. Risk Analysis

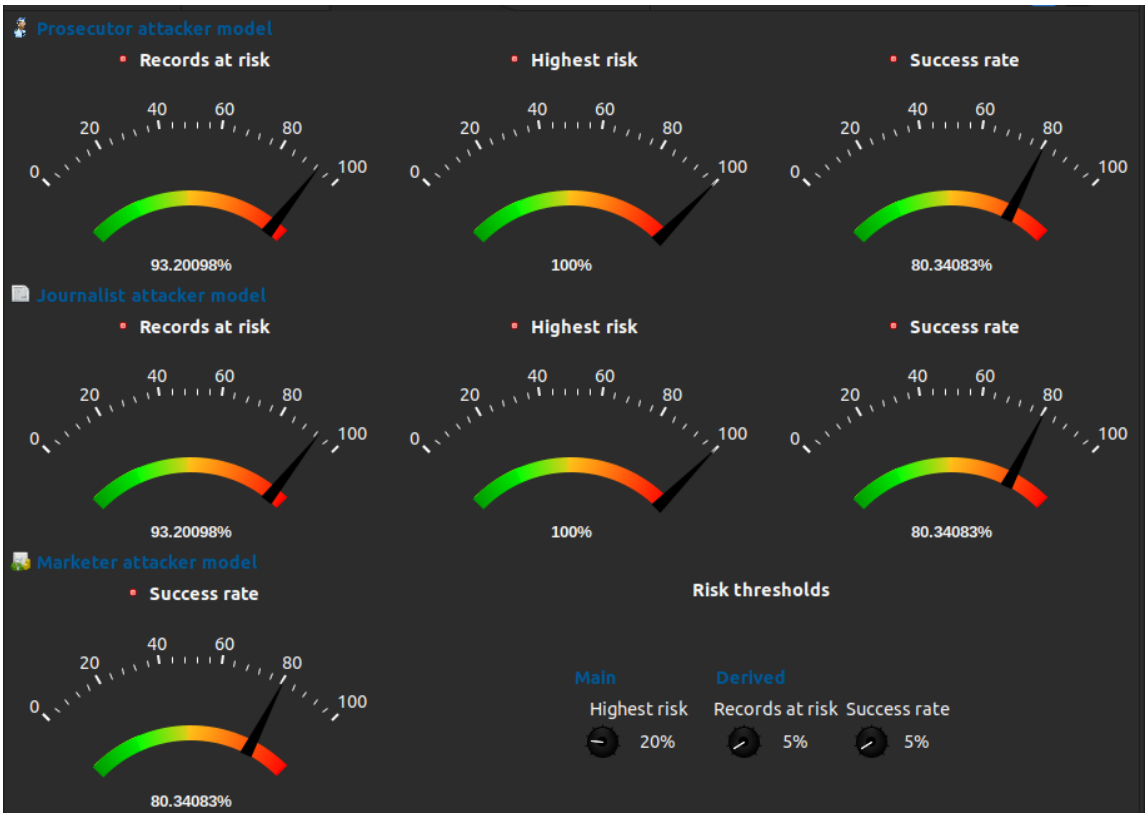


Figure 1: Risk Analysis - initial dataset

To evaluate the privacy risks in the original, non-anonymized dataset, we used the risk analysis tools provided by the ARX software. ARX simulates the behavior of potential attackers through three distinct attacker models, each representing different types of real-world privacy threats:

Prosecutor Attacker Model

This model assumes that the attacker knows that a specific individual is in the dataset. Using partial background knowledge (e.g., age, gender, ZIP code), the attacker tries to match that information with a record. This approach is common in cases where there is a personal or legal motivation to re-identify a particular individual.

- **Records at risk:** 93.20%
- **Highest individual risk:** 100%
- **Success rate:** 80.34%

These values indicate that almost the entire dataset is vulnerable to targeted re-identification, which poses a critical threat to individual privacy.

Journalist Attacker Model

This model simulates a scenario where the attacker does not target anyone specific, but rather looks for any sensitive or newsworthy information that can be linked to individuals. It reflects how a journalist might explore a dataset for potential leaks or ethical violations.

- **Records at risk:** 93.20%
- **Highest individual risk:** 100%
- **Success rate:** 80.34%

Results are nearly identical to the prosecutor model, indicating that even without a specific target, the dataset still enables high-probability re-identification.

Marketer Attacker Model

This model represents an attacker trying to profile groups of people rather than individuals, often for advertising or analysis purposes. It focuses on extracting patterns and trends across demographic categories. A successful attack under this model implies that a large number of records can be re-identified, not just one.

- **Success rate:** 80.34%

Despite being less aggressive than the other models, the marketer model still shows a high probability of successful profiling, which could lead to discriminatory or unethical group-level decisions.

The dataset, in its original state, presents a very high re-identification risk across all attacker models. With more than 93% of records at risk and success rates consistently above 80%, the need

for anonymization is urgent. These results justify the application of robust privacy models in order to ensure compliance with data protection standards and to minimize privacy threats.

4. Application of Privacy Models

After classifying the dataset attributes and analyzing the re-identification risks of the original dataset, we applied privacy models to reduce the exposure of sensitive information while preserving as much data utility as possible. In this assignment, we focused on two widely used models: k-anonymity and l-diversity. These models were implemented using the ARX anonymization tool.

k-Anonymity

k-anonymity ensures that each individual in the dataset is indistinguishable from at least k-1 others with respect to their quasi-identifiers. This is done by grouping records with similar QID values and generalizing or suppressing data so that no combination of QIDs occurs in fewer than k records.

In our case, we applied k = 5, meaning every group of records with the same QID values must contain at least 5 individuals. This prevents an attacker from confidently linking a record to a unique person, even if they have external background knowledge.

l-Diversity

While k-anonymity protects against re-identification, it does not ensure that sensitive attributes are diverse within a group. For example, if all 5 people in a k-anonym ARX-launcher.run ous group have the same disease, the attacker can still learn that disease even if they don't know who is who.

To address this, we applied l-diversity with `l = 2`. This model requires that each group of records sharing the same quasi-identifiers must contain at least `l` different values for the sensitive attribute (in this case, `Salary Class`). This protects against inference attacks — situations where attackers guess sensitive values based on the group a person falls into.

4.1. Data treatment before applying the models

One thing that we noticed on the provided dataset was the there was an overpopulation on the 50-60 age range, with the raw data, the age distribution looked something like this:

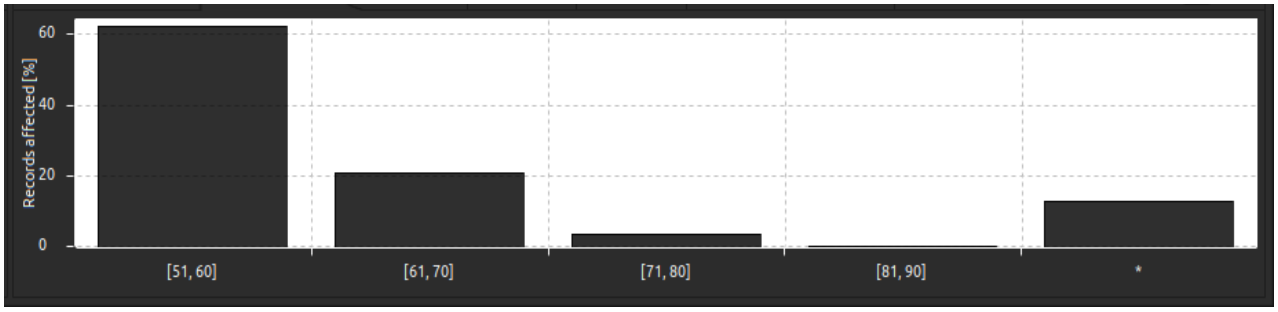


Figure 2: Age distribuiton on initial data

As you can see, most of the data is concentrated on a 10 year span. This makes the other ages easily identifiable and linked with other attributes which can be a big problem.

So we decided to redefine the age intervals to have a more homogeneous spread of the data, making it harder to link people to certain obtained data.

Even though the dataset is really big and a low percentage still is a significant amount of individuals, when linked with other attributes, the number of individuals reduces a lot making it easier to identify.

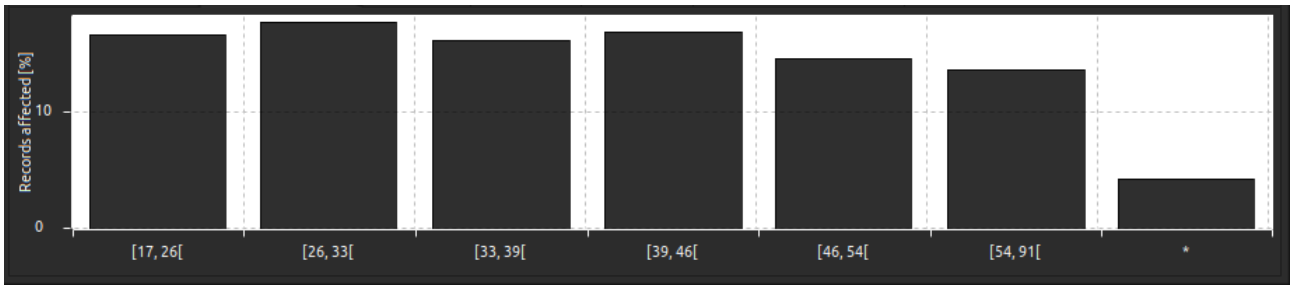


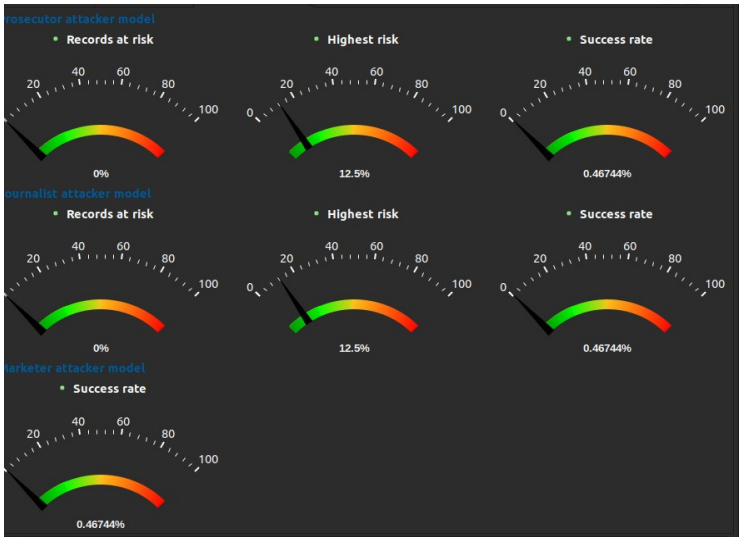
Figure 3: Age distribution after redefine the intervals

With the new generalization hierarchy, the dataset now exhibits a much more balanced distribution, reducing privacy risks while still preserving analytical value.

4.2. Anonymized Datasets

4.2.1. Anonymized Dataset with k-Anonymity (k=5) and l-Diversity (l=2)

The first anonymized dataset was created using k-anonymity with k=5 and l-diversity with l=2, aiming to meet the minimum privacy requirements while preserving as much data utility as possible. This configuration allows each record to be indistinguishable from at least four others and ensures that sensitive attributes have at least two distinct values within each group. Despite being a minimal privacy setup, the dataset showed strong performance in both privacy protection and utility metrics, serving as a solid baseline for comparison with more privacy-focused configurations.



The privacy risk analysis shows that the dataset provides strong protection, even with minimal privacy parameters (k=5, l=2). The highest re-identification risk is 12.5%, which is below the theoretical maximum of 20% for k=5, indicating that equivalence classes are often larger than required. Additionally, no records are fully at risk, and the success rates for attacker models (journalist and marketer) remain extremely low at around 0.47%. These results confirm that the dataset is well protected against both exact and probabilistic reidentification attacks.

The utility analysis of the anonymized dataset, based on quality metrics provided by ARX, indicates that the dataset maintains a high level of analytical value despite the applied privacy constraints. The granularity score of 90.09% and discernibility at 94.11% reflect that quasi-identifiers retained a high degree of detail and that equivalence classes were well-formed, enabling meaningful differentiation between records. Additionally, the attribute-level squared error (89.99%) and record-level squared error (85.15%) suggest that the anonymization introduced minimal distortion at both individual and attribute levels, preserving the dataset's internal consistency. The average class size of 99.35% confirms that group sizes were highly uniform, which helps avoid bias in aggregated analyses. While the non-uniform entropy (64.96%) is moderately lower, it remains acceptable under the applied l=2 constraint, indicating a reasonable level of sensitive attribute diversity. Overall, these results confirm that the anonymization strategy was effective in protecting privacy while ensuring that the dataset remains highly usable for statistical analysis and machine learning tasks.

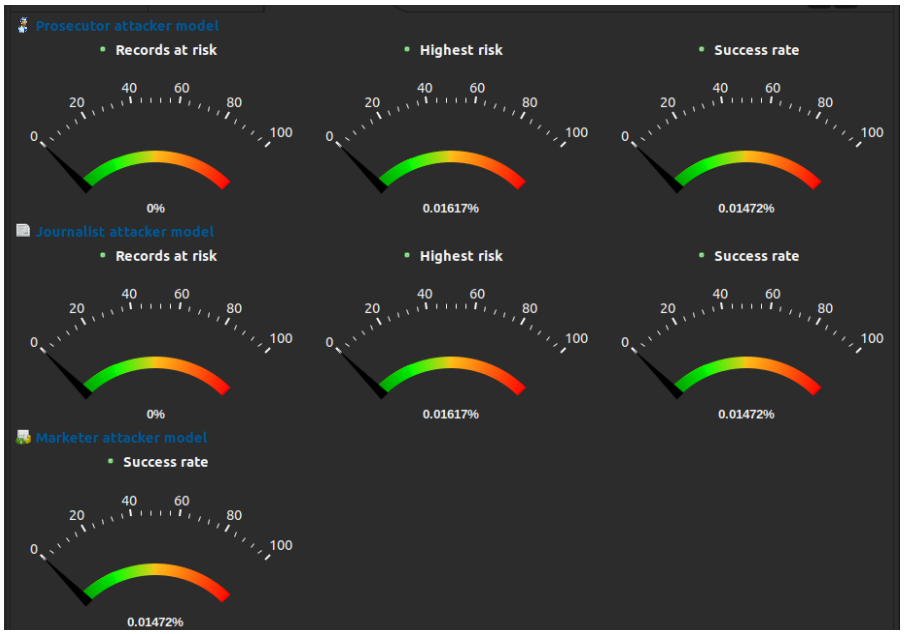
4.2.2. Anonymized Dataset with k-Anonymity (k=15) and l-Diversity (l=2)

The second anonymized version of the dataset was generated using a more privacy-focused configuration, with k-anonymity set to 15 and l-diversity set to 2. This configuration ensures that each record is indistinguishable from at least 14 others, and that each group contains at least two distinct values of the sensitive attribute. As a result, the highest re-identification risk dropped to just 0.08065%, with 0% of records at risk according to all attacker models. The success rate for realistic attackers also remained extremely low (0.03979%), confirming the robustness of this model. This

version clearly prioritizes privacy protection, serving as a strong contrast to more utility-oriented configurations, and highlighting the trade-off between anonymization strength and data usability.

4.2.3 Anonymized Dataset with k-Anonymity (k=5) and T-Closeness(0.2)

In this third configuration, the dataset was anonymized using a combination of k-anonymity with k=5 and t-closeness with t=0.2, aiming to enhance protection against attribute disclosure by ensuring that the distribution of sensitive attributes in each group closely matches the overall distribution in the dataset. This model imposes stricter privacy constraints than k-anonymity or l-diversity alone, providing stronger safeguards against both reidentification and inference attacks. The goal was to assess how well t-closeness could reinforce privacy without excessively degrading data utility.



While the privacy risk metrics show extremely strong results — with a highest re-identification risk of only 0.016% and 0% of records at risk, the quality metrics reveal substantial loss of utility. Generalization intensity (33.78%) and granularity (32.86%) suggest moderate structural distortion, but key indicators like non-uniform entropy (20.95%), record-level error (21.55%), and especially aggregation-specific error (26.04%) highlight significant degradation in data quality. Additionally, the attribute “education” was completely suppressed, which further reduced analytical value. These results reflect a typical side effect of t-closeness: while it strengthens privacy by protecting against attribute inference, it often does so at the cost of losing sensitive attributes and damaging both record-level and aggregate data utility.

4.3.Comparative Analysis of Privacy Models and Their Impact on Data Utility

Across the three anonymization configurations, clear differences emerged in the balance between privacy and data utility. The first model (k=5, l=2) prioritized utility, achieving high-quality scores (discernibility > 94%, attribute-level error 90%) while keeping the highest risk at 12.5% ,the upper limit allowed for k=5. The second configuration (k=15, l=2) significantly reduced the highest risk to 0.06%, improving privacy without excessively harming utility (generalization intensity 59%, aggregation-specific error 79%). Finally, the third model (k=5, t=0.2) achieved the strongest privacy (highest risk 0.016%, 0% of records at risk), but at the cost of major utility loss, with several metrics falling below 30% and one attribute (education) being completely suppressed. These results confirm that tighter privacy constraints, especially t-closeness, come at the cost of analytic usefulness, and that the choice of model must depend on whether privacy or utility is the primary concern for the intended use of the data.

Model	k-Anonymity	l-Diversity	t-Closeness	Highest Risk	Overall Utility	Key Observations
Model 1	5	2	–	12.5%	High (~90%)	Good utility, but higher re-identification risk (theoretical max).
Model 2	15	2	–	0.06%	High (~70%)	Strong privacy with minimal loss of utility.
Model 3	5	–	0.2	0.016%	Low (~40%)	Maximum privacy, but significant utility loss and attribute suppression.

5. Impact of Varying Privacy Model Parameters on Risk and Utility(In dataset with(k=15 and l=2)

To better understand the trade-off between privacy and utility, we analyzed the effect of varying specific privacy model parameters. The ARX tool allows fine-tuning of several key parameters that directly influence the anonymization process and the resulting data quality. In this section, we focus on analyzing the impact of changing the suppression limit on both privacy risk and data utility.

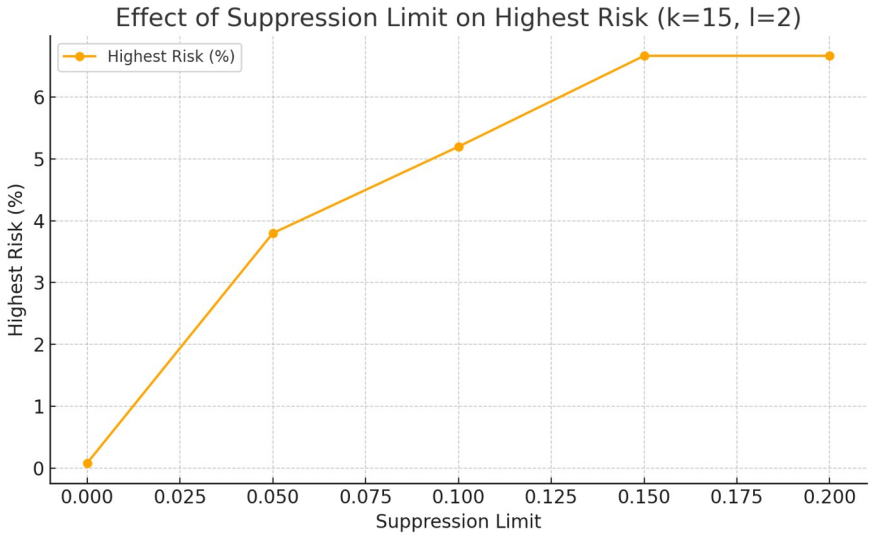
Table 1 summarizes the effect of increasing the suppression limit on both privacy risk and utility metrics. The values confirm that while suppression may help meet stricter privacy constraints, it also leads to greater information loss and decreased analytical reliability. The best configuration must therefore balance these opposing effects.

Table 1:

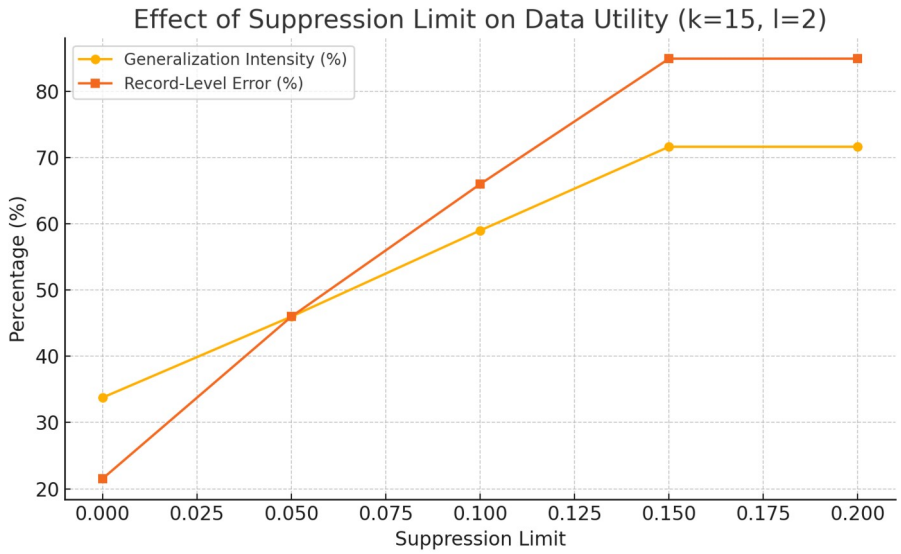
Suppression Limit	Highest Risk (%)	Generalization Intensity (%)	Record-Level Error (%)
0.00	0.08065	33.78	21.55
0.05	3.80	46.00	46.00
0.10	5.20	59.00	66.00

Suppression Limit	Highest Risk (%)	Generalization Intensity (%)	Record-Level Error (%)
0.15	6.67	71.63	84.93
0.20	6.67	71.63	84.93

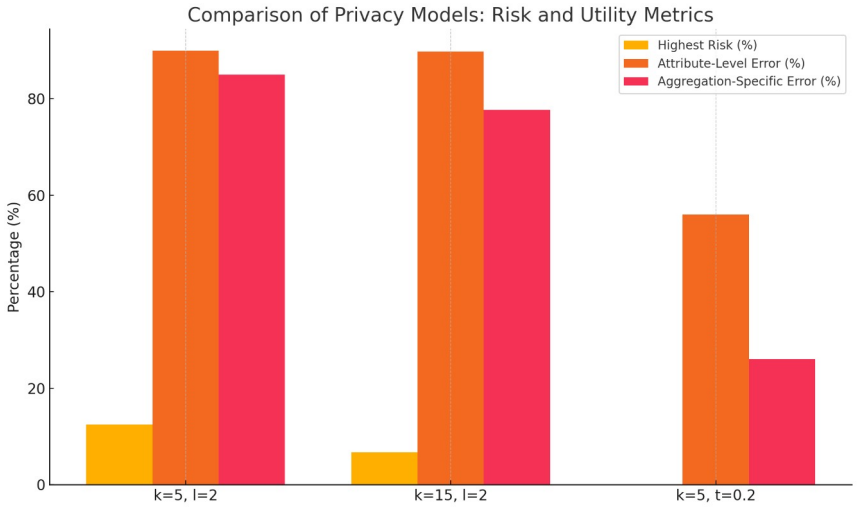
Interestingly, setting the suppression limit to 0.20 did not lead to any noticeable difference compared to configurations with 0.15 or even 0%. This confirms that the ARX tool only applies suppression when strictly necessary. In this case, the dataset already satisfies the privacy constraints (k=15, l=2) through generalization alone, and no records needed to be removed. As a result, the privacy risk and utility metrics remain consistent, regardless of the suppression threshold.



The plot shows that the highest re-identification risk remains extremely low (0.08065%) even with a suppression limit of 0%. Increasing the suppression limit has no further impact on risk, indicating that generalization alone was sufficient to meet the privacy constraints.



This plot illustrates a steady increase in generalization intensity and record-level error as the suppression limit increases. It confirms that allowing more suppression leads to greater distortion, reducing the overall utility of the dataset.



Comparison of Privacy Models on Risk and Utility Metrics. Model 1 (**k=5, l=2**) shows the highest re-identification risk but excellent utility. Model 2 (**k=15, l=2**) provides a strong privacy-utility balance. Model 3 (**k=5, t=0.2**) achieves the lowest risk but at the cost of significantly reduced utility.

The analysis of suppression limit variation under the configuration **`k=15`** and **`l=2`** revealed a surprising but important result: setting the suppression limit to 0% produced better outcomes than allowing higher suppression thresholds. Even with no suppression, ARX achieved a very low highest risk (0.08065%), with 0% of records at risk, and maintained strong utility metrics. This confirms that the dataset structure aligns well with the chosen privacy model, enabling effective anonymization purely through generalization.
Allowing suppression (15% or 20%) did not improve privacy and instead introduced unnecessary distortion. This reinforces the importance of testing and tuning parameters, sometimes, less is more when the data is naturally suitable for the applied model.

6. General Conclusion of the Report

After testing and evaluating three anonymization models, it becomes evident that Model 2, using *k-anonymity with k=15* and *l-diversity with l=2*, offers the most balanced trade-off between privacy and utility. It significantly reduces the highest re-identification risk (to 0.06%) while preserving high data quality, with low attribute distortion and strong aggregation performance. This makes it well-suited for use cases that require both analytical depth and solid privacy guarantees.

t-Closeness, tested in Model 3, provides the strongest level of privacy, with virtually no re-identification risk and full protection against attribute disclosure. However, it comes at a substantial cost: lower utility metrics, high suppression, and in this case, complete loss of a sensitive attribute.

Therefore, t-closeness is best reserved for highly sensitive datasets or public data releases where the risk of attribute inference must be minimized, and some loss in utility is acceptable.

This project reinforces the importance of adjusting anonymization parameters to the nature and structure of the dataset. Applying strong models blindly can lead to full data suppression or unusable output. Datasets with limited size, low diversity in sensitive attributes, or high uniqueness in quasi-identifiers often require a more conservative configuration to avoid total loss of information.

Limitations Observed

- Attribute imbalance, which made it difficult to satisfy t-closeness;
- The small size of the dataset, which limited the possible groupings under stricter privacy models;
- Fixed generalization hierarchies in ARX, which sometimes forced overly aggressive generalization.

Suggestions for Improvement

Future improvements could include:

- Designing custom generalization hierarchies to better reflect the semantics of the data,
- Applying local recoding to isolate and anonymize outliers without distorting the entire dataset,
- And performing pre-processing to reduce uniqueness, such as binning values or merging rare categories, before applying privacy models.