Anonymization of a Dataset with Utility and Risk Analysis

Guilherme Oliveira – up202204987
Mário Minhava - up202206190

Index

1 Introduction

In the digital age, ensuring the privacy of individuals in large datasets is crucial, especially given the risks associated with re-identifying individuals from anonymized data. This project utilizes the ARX De-identification Tool to apply advanced anonymization techniques to a dataset provided by the tool's developers. Our objective is to classify dataset attributes into Identifying, Quasi-Identifiers (QIDs), Sensitive, and Insensitive categories, assess the original data's privacy risks, and apply multiple privacy models to enhance data protection while maintaining utility.

According to the European Union's data protection laws, in particular the General Data Protection Regulation (GDPR), anonymous data is "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable".

The process includes an in-depth analysis of the re-identification risks, the effectiveness of different anonymization strategies, and iterative adjustments based on the results. This report will outline the methods used, justify the choices made, and evaluate the outcomes, aiming to strike a balance between robust privacy protection and the retention of data utility for analytical purposes.

1.1 ARX

ARX is a comprehensive open-source tool designed for data anonymization and privacy-preserving data transformations. It enables users to apply a range of anonymization techniques on personal data sets to protect individuals' privacy while maintaining the utility of the data for analysis. ARX supports multiple privacy models, including k-anonymity, l-diversity, and t-closeness and many others, allowing for flexible adaptation to various privacy requirements. Additionally, it offers functionalities for risk assessment, data transformation, and data import/export in different formats. The tool is user-friendly, equipped with a graphical interface that facilitates the management of data anonymization projects, making it accessible for both technical and non-technical users.

2 Dataset Characteristics

To start the anonymization of the data, we chose the provided dataset which can be found in the ARX Website

2.1 Dataset analysis

The dataset provided is full of personal information of a said person, each column represents an attribute or info of that individual (there will be no explanation of the column meaning as they are pretty self explanatory) for example:

- Sex

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

The problem with this dataset is that it contains a range of attributes that can compromise an individual's privacy due to their potential to uniquely identify someone or reveal sensitive personal information.

Collectively, these attributes can create a detailed profile of individuals. Without proper anonymization, datasets containing such information can expose individuals to risks of re-identification and potential misuse of personal data, such as identity theft, discrimination, or other forms of personal harm. Thus, it is essential to apply robust data protection measures to safeguard privacy effectively.

2.2 Classification of the data

Before initializing the risk analysis we first have to classify the attributes present in the dataset, first we will give a brief explanation of the classes which the data will be placed in.

Identifiers Attributes - These are attributes that can directly identify an individual. These are often removed or heavily altered during the anonymization process to prevent direct identification.

Quasi – Identifier Attributes - Quasi-identifiers are attributes that may not uniquely identify an individual on their own but can do so when combined with other quasi-identifiers. They require careful handling in anonymization processes such as generalization or suppression to prevent linkage attacks that could re-identify individuals.

Sensitive Attributes - These are attributes that contain information that is considered confidential or that could cause harm or discrimination if associated with an individual. Protecting these attributes is crucial for maintaining individual privacy and compliance with legal standards.

Insensitive Attributes - These are attributes that are considered to not significantly impact an individual's privacy if disclosed. These attributes usually do not require special handling in the anonymization process.

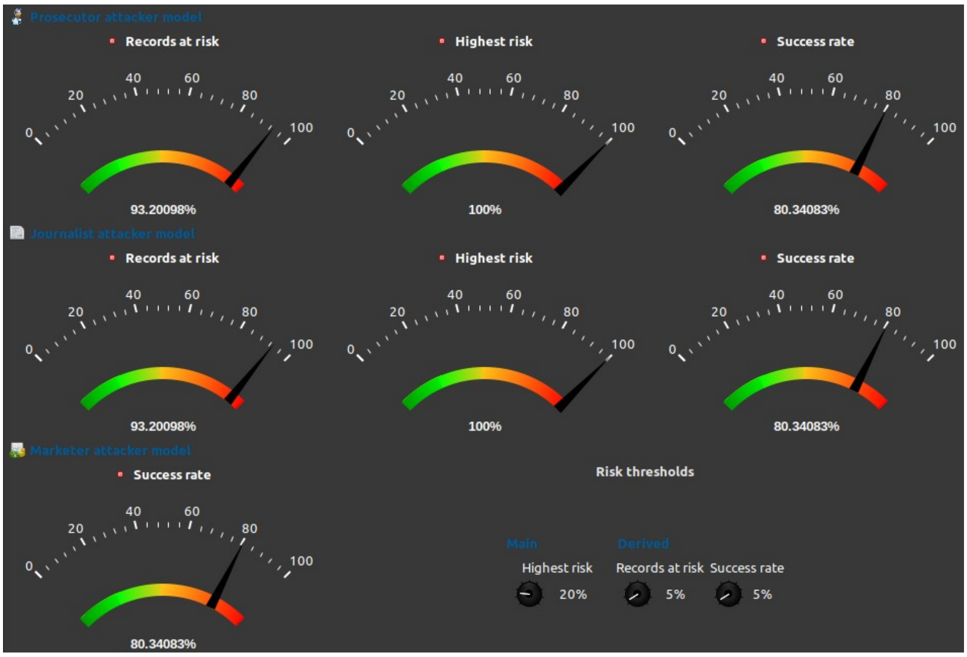
Sex	Considered a **quasi-identifier**, sex can be combined with other attributes to narrow down the identity of an individual within a dataset.
Age	As a **quasi-identifier**, age can be particularly revealing when combined with other demographic data, making it easier to identify specific individuals, especially in smaller or more uniform datasets.
Race	This is a **sensitive attribute** that can lead to privacy issues and discrimination if mishandled. Race information, when linked with other identifiers, can uniquely identify individuals in certain demographics. Contributing to biased decisions in

	many real world problems.
Martial Status	As a **quasi-identifier** this attribute can be used in conjunction with other data points to more precisely identify individuals and infer additional personal information.
Education	Education level can also act as a **quasi-identifier**, especially when correlated with occupation or salary class, potentially revealing more about an individual's background.
Native Country	This can be considered a **sensitive attribute** for the same reasons as race.
Work class	This attribute, which indicates the type of employment sector is a **quasi-identifier**. It provides socioeconomic context about an individual that, when linked with other data, can help narrow down individual identities.
Occupation	Typically treated as a **quasi-identifier**, occupation details can be quite specific and, particularly in combination with other attributes like salary class and work class, can significantly narrow the scope for identifying an individual.
Salary Class	This is a s**ensitive attribute** because it deals directly with an individual's income level, which is considered personal and confidential information. Disclosure of salary information can lead to privacy breaches and discrimination. Although in the provided dataset the values for the salary are only 2 (<50k; >50k), we still consider this to be sensitive data, would be more important if the dataset had more values for the salary.

Quasi – Identifier Attributes: Sex; Age; Martial Status; Education; Work class; Occupation

Sensitive Attributes: Race; Native Country; Salary Class

3 Risk Analysis



Risk thresholds

Main

Highest risk 20%

Derived

Records at risk 5%

Success rate 5%

Figure 1: Risk Analysis - initial dataset

The ARX software provides us with information about 3 types of risks, these models are:

Prosecutor attacker model - This model assumes the attacker aims to identify a specific individual in an anonymized dataset. Using known information about their target, the attacker tries to match this data with records in the dataset. This model is often considered in scenarios where the attacker has a direct and personal interest in identifying the individual.

Journalist attacker model - In this model, the attacker (similar to a journalist) searches for interesting or newsworthy information within the dataset **but does not target a specific individual initially**. The goal is to uncover identities or sensitive details through available data, highlighting the need for robust anonymization to prevent unintended disclosures.

Marketer attacker model - Here, the focus is on profiling groups rather than individuals, primarily for marketing purposes. The attacker uses the dataset to understand characteristics and preferences of demographic groups. This model concerns itself with the broader implications o. An attack can therefore only be considered successful if a larger fraction of the records could be re-identified.

As you can see in the fig. 1 the initial risk are very high, this is obviously do to the data being in their purest form, they haven't gone through the process of anonymization

4 Privacy Models

Privacy models are frameworks designed to protect personal data in datasets by ensuring that the information cannot be exploited to compromise individual privacy. These models provide systematic approaches to reduce the risk of re-identification of individuals from anonymized data.

4.0.1 K-Anonymity

This model ensures that each record in a dataset is indistinguishable from at least k−1 other records concerning certain identifying attributes. By doing so, k-anonymity protects against re-identification in the dataset by making each person's data blend in with at least k−1 others.

4.0.2 δ-Disclosure privacy

The model is a more nuanced approach to assessing and managing the risk of sensitive information disclosure in anonymized datasets. Unlike more general models such as k-anonymity or differential privacy, Delta-Disclosure Privacy focuses specifically on quantifying the risk that sensitive attribute values can be inferred with a high degree of accuracy due to the anonymization process itself.

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:

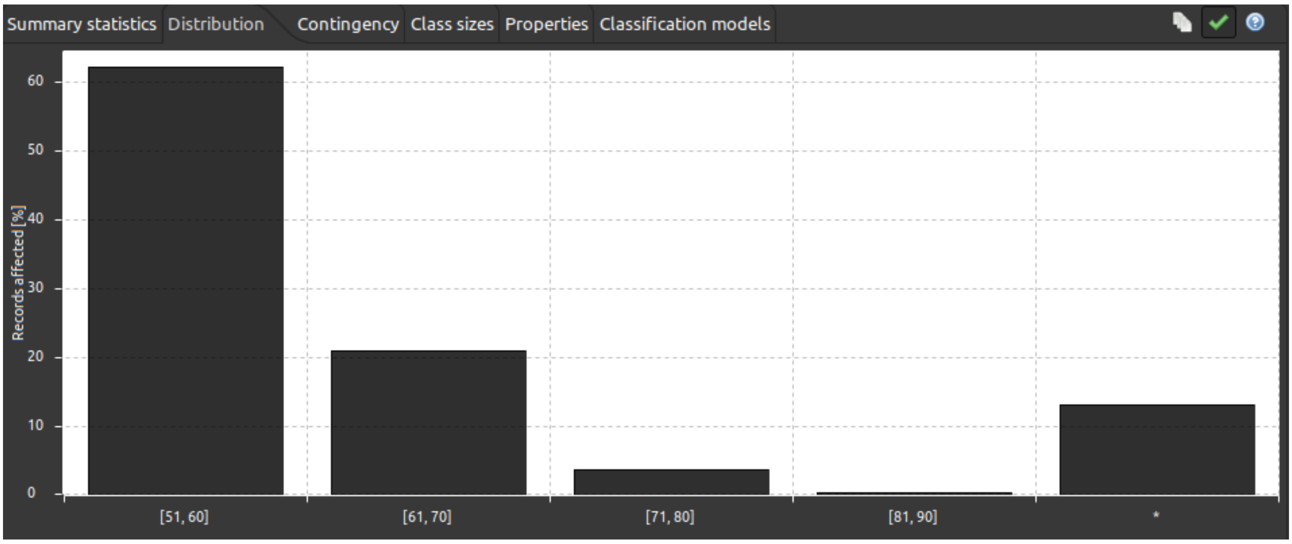


Figure 2: Age distribuiton on initial data

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.

4.2 Result Analysis

After applying the transformations and the privacy models we obtained the following results:

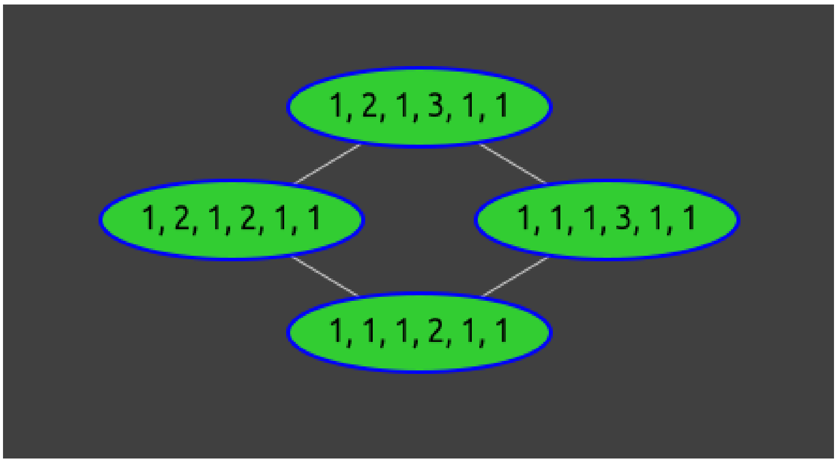


Figure 3: Applied Transformations

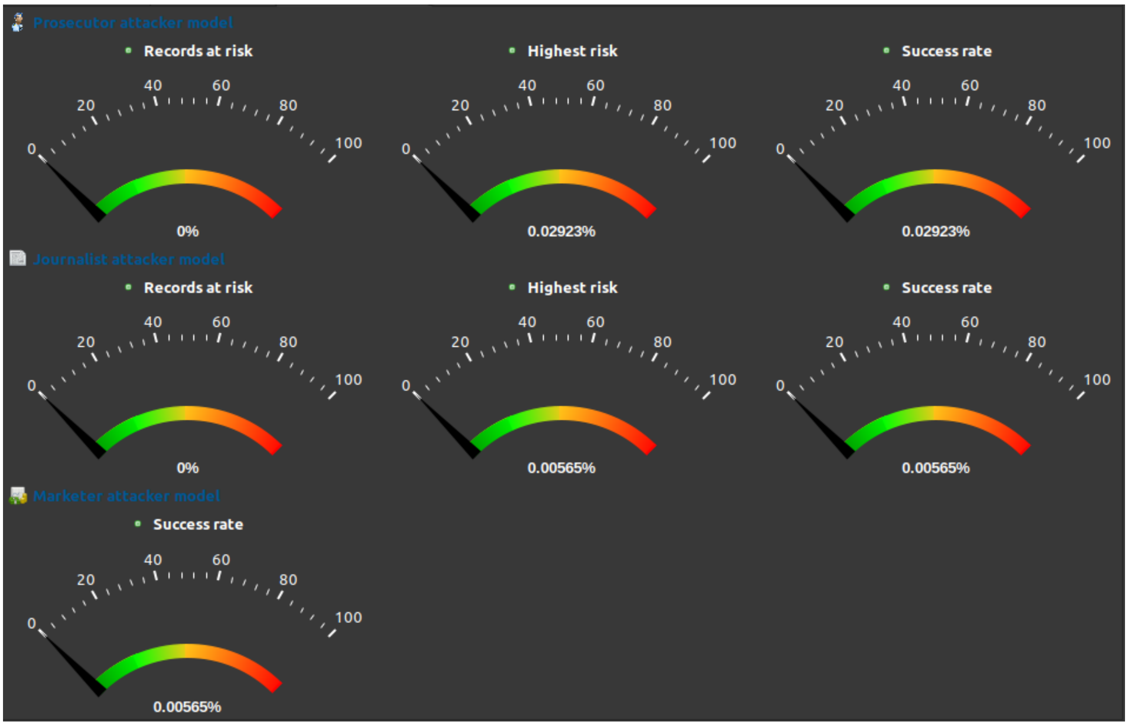
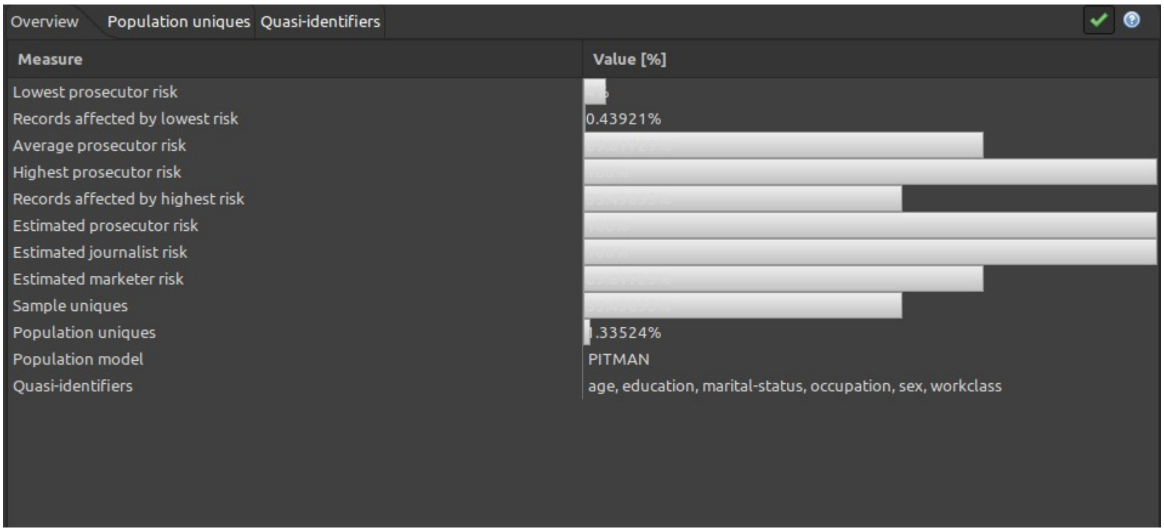


Figure 4: Risk Analysis - transformed dataset

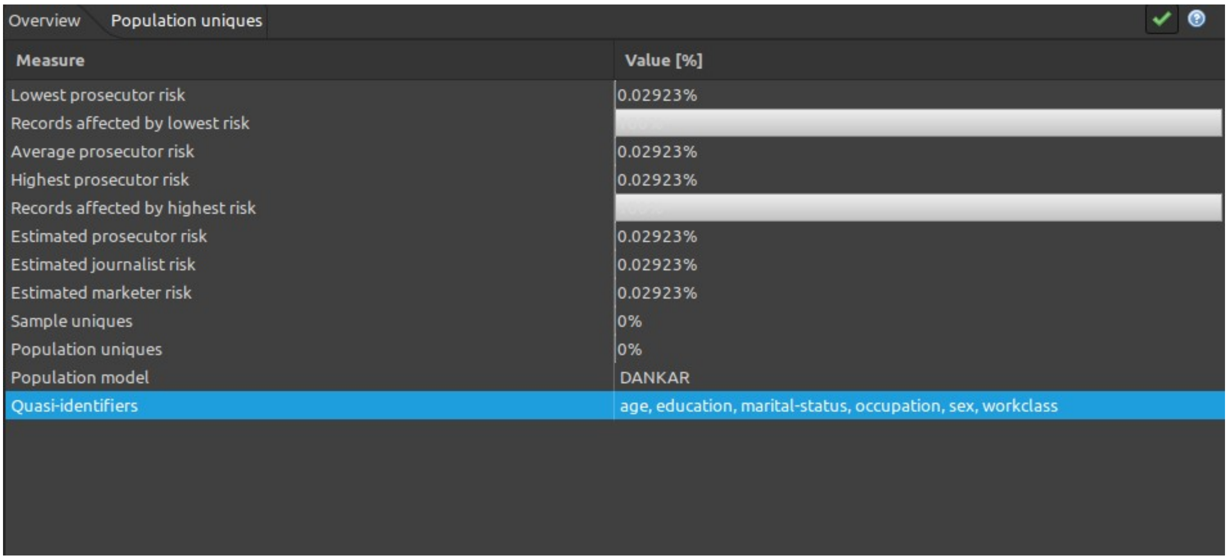
The dataset was successfully anonymized, although it is to notice that by doing so we compromise the utility of the data. So we had to balance the trade off between anonymization and utility and we reached this result.

As you can see in fig.5 and fig.6 we compare the differences between the initial dataset and the datset after the transformations



Measure	Value [%]
Lowest prosecutor risk	[illegible]
Records affected by lowest risk	0.43921%
Average prosecutor risk	[illegible]
Highest prosecutor risk	100%
Records affected by highest risk	[illegible]
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	[illegible]
Sample uniques	[illegible]
Population uniques	1.33524%
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figure 5: Before



Overview | Population uniques

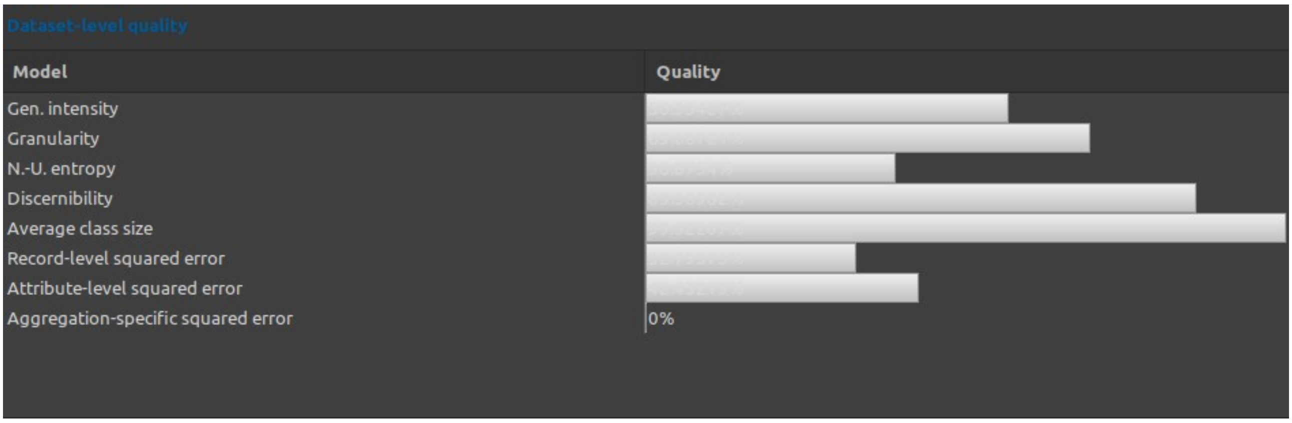
Measure	Value [%]
Lowest prosecutor risk	0.02923%
Records affected by lowest risk	100%
Average prosecutor risk	0.02923%
Highest prosecutor risk	0.02923%
Records affected by highest risk	100%
Estimated prosecutor risk	0.02923%
Estimated journalist risk	0.02923%
Estimated marketer risk	0.02923%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figure 6: After

It is evident that there is a significant improvement in the privacy risk metrics of the dataset between the initial and subsequent anonymization processes. These changes imply that the dataset's anonymization strategy was adjusted to better protect sensitive information while potentially allowing for a slight decrease in data utility, as indicated by increased risks. This approach enhances privacy by making it more challenging to link quasi-identifiers to specific individuals, thus improving compliance with privacy standards and reducing the vulnerability to data breaches.

4.2.1 Quality Metrics Analysis

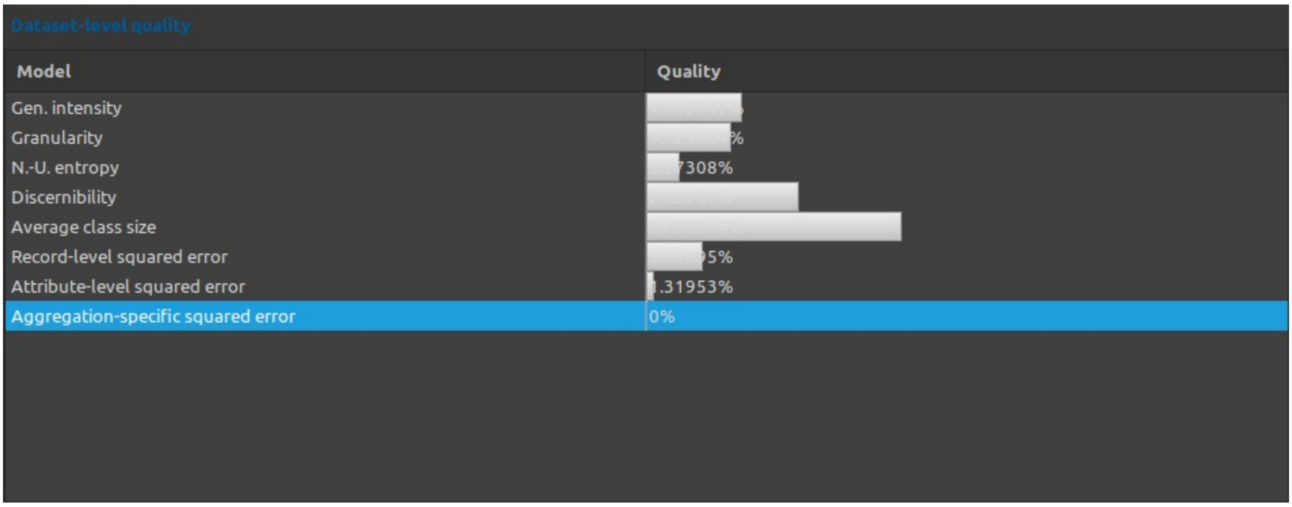
Quality metrics provide a systematic approach to evaluate how well anonymization methods maintain the utility of a dataset while ensuring individual privacy. This analysis will focus on several key metrics, including Generalization Intensity, Granularity, Entropy etc... To understand the trade-offs between data utility and privacy. By examining these metrics before and after applying anonymization techniques, we can discern the effectiveness of these methods and optimize our approach to both protect sensitive information and preserve the value of the data for analysis.



Dataset-level quality

Model	Quality
Gen. intensity	[illegible]
Granularity	[illegible]
N.-U. entropy	[illegible]
Discernibility	[illegible]
Average class size	[illegible]
Record-level squared error	[illegible]
Attribute-level squared error	[illegible]
Aggregation-specific squared error	0%

Figure 7: Quality Metrics - Before



Dataset-level quality

Model	Quality
Gen. intensity	[illegible]
Granularity	[illegible]%
N.-U. entropy	[illegible]7308%
Discernibility	[illegible]
Average class size	[illegible]
Record-level squared error	[illegible]5%
Attribute-level squared error	1.31953%
Aggregation-specific squared error	0%

Figure 8: Quality Analysis - After

The improvements between the initial and subsequent metrics primarily focus on enhancing privacy without a significant loss of data utility. The reduction in discernibility and attribute-level squared error, combined with a slight reduction in generalization intensity, indicates a more sophisticated approach to anonymization. This strategy likely involves better tuning of the anonymization techniques, such as optimized generalization or suppression, to preserve more original data characteristics while still protecting privacy. The increase in average class size also suggests that the data now contains larger, more uniform groups, which complicates efforts to pinpoint individual records, thereby improving privacy. These adjustments suggest a move towards an anonymization approach that more effectively balances privacy protection with the retention of useful data attributes.

4.3 Level of Suppression and Coding Model

When considering the effects of the level of suppression and the coding model on the results of anonymization, it's essential to understand how each strategy impacts both the privacy and utility of the data.

Suppression

Suppression involves removing or masking certain data elements to protect sensitive information. This method is effective in reducing the risk of disclosure but can significantly impact data utility.

Effects on Results:

- **Privacy Increase**: Suppression enhances privacy by completely removing the data points that could potentially identify individuals.
- **Data Loss**: The main drawback is a loss of data utility because suppressed data cannot be used for analysis, leading to gaps in the information available for decision-making.
- **Bias Introduction**: Over-suppression can introduce bias into the dataset, as it might disproportionately affect certain records or attributes, skewing analysis results.

Generalization

Generalization involves broadening the specificity of data, such as modifying age from exact numbers to ranges (e.g., 30-40 instead of 33). This reduces the identifiability of data while retaining more information than suppression.

Effects on Results:

- **Balanced Privacy and Utility**: Generalization typically strikes a better balance between privacy protection and data utility, as it still allows for analysis within the generalized categories.
- **Reduced Precision**: While generalization retains more information than suppression, it reduces the precision of the data, which can affect analyses that rely on detailed attribute values.
- **Scalability**: Generalization is more scalable to larger datasets, as it doesn't remove data but rather adjusts its format, maintaining a richer dataset for analysis.

In conclusion, the choice and degree of suppression versus generalization have profound effects on the balance between protecting privacy and maintaining the utility of the dataset. These strategies need to be carefully calibrated based on the sensitivity of the data, regulatory requirements, and the specific analytical needs of the dataset's users.

5 Conclusion and Final Remarks

In conclusion, this project has successfully navigated the complex balance between safeguarding personal privacy and maintaining the utility of the dataset through the application of various anonymization techniques. By classifying attributes into quasi-identifiers and sensitive categories, and employing different privacy models such as k-anonymity and δ-Disclosure privacy, we have significantly reduced the risk of re-identification. The subsequent analyses using privacy and quality metrics have illustrated that while techniques like generalization and suppression alter the data, they do so in a way that strategically maximizes privacy without rendering the data useless.

The improvements in privacy metrics, particularly in reducing discernibility and enhancing the average class size, demonstrate our success in creating a more secure dataset. Moreover, the adjustments in suppression and generalization have shown to be effective in optimizing both the privacy and utility of the data. As we move forward, the insights gained from this project can guide further refinements in our approach to data anonymization, ensuring that we continue to protect individual privacy while supporting valuable data-driven insights.

Things to learn from the making of this project are, that we should play more with the many changes that could have been performed in the dataset and try to learn more in depth all the privacy models that he could have used, and combining them to have the best result possible.