

1 2 9 0



FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE
COIMBRA

Anonymization of Datasets with Privacy, Utility and Risk Analysis

Segurança e Privacidade • 2025/2026

Relatório

Elementos do Grupo

Daniel Pereira • 2021237092 • uc2021237092@student.uc.pt

Vasco Fernandes • 2025186614 • uc2025186614@student.uc.pt

Índice

<i>Índice</i>	2
1. Introdução e objetivos.....	3
2. Escolha do dataset.....	4
3. Descrição e caracterização do dataset.....	4
4. Distinção e separação.....	6
4.1. Definição.....	6
4.2. QIDs isolados	6
4.3. Combinação de QIDs.....	7
5. Análise e Configuração da Anonimização	8
5.1. Estratégia	8
5.2. Requisitos de privacidade e utilidade	8
5.3. Hierarquização.....	9
5.4. Atribuição de pesos	11
5.5. Limite de supressão.....	12
5.6. Supressão e Generalização.....	12
6. Modelos de Privacidade	13
6.1. k -Anonymity	13
6.2. ℓ -Diversity.....	14
7. Análise do impacto dos Modelos de Privacidade.....	15
7.1. Análise do Risco	15
7.1.1. Quasi-identifiers	15
7.2. Análise da Utilidade.....	19
7.2.2. Análise geral da utilidade do dataset.....	20
8. Conclusão	21
9. Referências	22

1. Introdução e objetivos

O objetivo principal deste trabalho é realizar um processo de anonimização sobre um dataset, preservando a privacidade dos indivíduos representados e garantindo que o conjunto de dados anonimizado mantém utilidade suficiente para fins analíticos.

De forma mais específica, o projeto visa:

1. Selecionar e preparar um conjunto de dados adequado, assegurando qualidade e diversidade de atributos para a aplicação de diferentes modelos de anonimização.
2. Classificar e caracterizar os atributos de acordo com a sua natureza (identificadores, quasi-identificadores, sensíveis e não sensíveis).
3. Aplicar e comparar diferentes modelos de privacidade, nomeadamente k-anonymity, ℓ -diversity e t-closeness [4], [5], avaliando a sua eficácia e impacto.
4. Analisar o equilíbrio entre privacidade e utilidade, estudando métricas de risco de reidentificação e de perda de informação.
5. Apresentar conclusões e recomendações sobre o processo de anonimização, destacando boas práticas para a publicação segura de dados pessoais.

Após esta contextualização inicial, procedeu-se à seleção, importação e caracterização do conjunto de dados, de modo a preparar a sua utilização na ferramenta ARX.

2. Escolha do *dataset*

Para a realização deste trabalho, recorremos ao dataset ***Adult Income*** disponível em **Kaggle** e originalmente proveniente do *UCI Machine Learning Repository*.

Procedemos a um **pré-processamento dos dados**, no qual:

- foram removidas todas as linhas que continham valores em falta (NaN);
- o conjunto foi limitado às primeiras 5000 entradas, em conformidade com as recomendações do projeto e para garantir a compatibilidade e eficiência de processamento no *ARX*;
- adicionamos as colunas *person_id* e *email*;
- adicionamos mais variedade à coluna *income*.

A seleção deste dataset justificou-se pela sua **heterogeneidade de atributos**, englobando variáveis **categóricas e numéricas** que possibilitam uma análise mais abrangente. Sendo composto por **dados reais e moderadamente sensíveis**, o conjunto oferece um cenário apropriado para **estudar e avaliar o equilíbrio entre privacidade, risco de reidentificação e utilidade dos dados**.

3. Descrição e caracterização do dataset

Cada atributo foi analisado e classificado de acordo com a sua função no processo de anonimização, seguindo as categorias definidas pela ferramenta *ARX*:

- **Identificadores diretos**: permitem identificar inequivocamente uma pessoa e devem ser removidos;
- **Quasi-identificadores** (QID): podem permitir a identificação de um indivíduo quando combinados com outros atributos QID;
- **Atributos sensíveis**: contêm informação privada que deve ser protegida;
- **Atributos não sensíveis**: não afetam a privacidade e podem ser mantidos sem alterações.

Atributo	Descrição	Classificação	Observação
<i>person_id</i>	Identificador único de cada indivíduo	Identificador direto	Identifica inequivocamente cada pessoa
<i>age</i>	Idade do indivíduo	Quasi-identificador	Pode contribuir para identificar uma pessoa quando combinado com outros atributos
<i>email</i>	Endereço de email	Identificador direto	Permite identificar diretamente o indivíduo
<i>workclass</i>	Tipo de emprego	Quasi-identificador	Ajuda a distinguir grupos socioeconómicos
<i>education-num</i>	Nível de escolaridade	Quasi-identificador	Permite agrupar indivíduos por nível educacional
<i>marital-status</i>	Estado civil	Quasi-identificador	Informação pessoal que pode distinguir indivíduos
<i>occupation</i>	Profissão	Quasi-identificador	Elevado potencial de identificação
<i>race</i>	Raça ou grupo étnico	Quasi-identificador	Informação demográfica usada para análises estatísticas
<i>gender</i>	Sexo	Quasi-identificador	Frequentemente usado em combinação com a idade e profissão
<i>hours-per-week</i>	Horas médias de trabalho por semana	Não sensível	Não contém informação sensível, baixo risco de reidentificação
<i>native-country</i>	País de origem	Quasi-identificador	Pode ajudar a distinguir indivíduos de origens menos frequentes
<i>income</i>	Rendimento anual	Sensível	Informação económica confidencial

4. Distinção e separação

4.1. Definição

Antes de proceder à análise dos valores de **Distinção** e de **Separção** de cada QID, bem como de alguns conjuntos de QIDs, importa clarificar o significado dessas métricas no contexto da avaliação de risco de reidentificação.

Distinção: avalia a capacidade de um atributo, ou de um conjunto de atributos, de identificar inequivocamente os registos de um dataset. Em termos formais, relaciona-se com o conceito de *identifying key* — um subconjunto de atributos que permitem distinguir cada registo dos restantes [6].

- **Quanto maior for a Distinção**, maior é o **número de valores únicos** existentes num atributo ou numa combinação de atributos

Separção: mede o grau em que as combinações de valores dos atributos distinguem os registos entre si.

- **Quanto maior for a separação**, maior é a **capacidade combinatória dos atributos em distinguir registos entre si**. Isto significa que um maior número de pares de registos apresenta **valores diferentes** em pelo menos um dos atributos analisados, o que **aumenta a possibilidade de identificar indivíduos através da combinação dessas variáveis**.

4.2. QIDs isolados

Após a análise da tabela apresentada na **Figura 1**, observa-se que a maioria dos *quasi-identificadores* (*QIDs*) apresenta **baixo poder distintivo quando considerados isoladamente**, evidenciado pelos valores reduzidos de **Distinção**. No entanto, alguns atributos, mesmo sozinhos, demonstram **níveis de separação significativamente elevados**, o que indica que, **quando combinados entre si**, têm uma **maior capacidade de distinguir registos** e, consequentemente, **de aumentar o risco de reidentificação**.

Embora os atributos ***email*** e ***person_id*** não sejam *quasi-identificadores*, foram **incluídos na tabela para referência**, dado o seu impacto direto no risco de identificação. Como era esperado, ambos apresentam **valores máximos de Distinção e Separção** (100%), uma vez que **possuem valores únicos para cada registo**, permitindo a **identificação inequívoca de cada indivíduo**.

Quasi-identifier	Distinctness	Separation
gender	0.04%	44.21524%
race	0.1%	25.17013%
workclass	0.14%	45.15051%
marital-status	0.14%	66.16267%
occupation	0.28%	89.46863%
educational-num	0.32%	81.14529%
native-country	0.8%	15.59532%
age	1.36%	97.83618%
email	100%	100%
person_id	100%	100%

Figura 1

4.3. Combinação de QIDs

A análise das combinações de *quasi-identificadores* (Figura 2)— variando de conjuntos de 2 a 6 QIDs — evidencia uma **tendência crescente de risco de reidentificação** à medida que aumenta o número de variáveis combinadas. Em todos os conjuntos analisados, **os valores de Separação ultrapassam 99%**, o que demonstra uma elevada capacidade de separação entre indivíduos.

Os valores de **Distinção** evoluem de forma progressiva, partindo de 13% nas combinações duplas mais críticas, até ultrapassar 80% nos maiores conjuntos.

Este comportamento confirma o risco existente de reidentificação quando se combinam vários QIDs.

Quasi-identifier	Distinctness	Separation
age, educational-num	13.22%	99.5498%
age, occupation	13.56%	99.75331%
age, marital-status, occupation	30.7%	99.87161%
age, educational-num, occupation	44.52%	99.93469%
age, workclass, educational-num, occupation	59.72%	99.959%
age, educational-num, marital-status, occupation	62.08%	99.96418%
age, workclass, educational-num, marital-status, occupation, gender	78.5%	99.98416%
age, workclass, educational-num, marital-status, occupation, race, gender	82.2%	99.98743%

Figura 2

5. Análise e Configuração da Anonimização

5.1. Estratégia

Antes de executar qualquer tipo de técnica de anonimização, é essencial definir a orientação estratégica do processo, ou seja, qual o objetivo principal a privilegiar:

- Opção 1: **maximizar a proteção de privacidade**
- Opção 2: **preservar a utilidade dos dados**
- Opção 3: **compromisso equilibrado entre privacidade e utilidade**

Para este *dataset* optámos pela **opção 3**, procurando alcançar um compromisso equilibrado entre **privacidade e utilidade**. A escolha desta opção justifica-se pelo carácter socioeconómico do dataset, que contém informações reais sobre indivíduos e atributos potencialmente sensíveis. Nestes casos, uma abordagem que privilegie apenas a privacidade poderia levar a uma perda excessiva de detalhe, comprometendo a utilidade analítica dos dados. Por outro lado, priorizar apenas a utilidade aumentaria o risco de reidentificação. Assim procurámos **assegurar simultaneamente as duas características**.

5.2. Requisitos de privacidade e utilidade

Antes de avançarmos para o processo de anonimização, foram definidos limiares quantitativos destinados a orientar a configuração do *ARX* e **avaliar se os resultados cumprem os requisitos mínimos de privacidade e utilidade**.

Estes limiares funcionaram como **critérios de aceitação** para a validação das técnicas aplicadas.

- **Risco máximo (*Highest risk*): $\leq 15\%$**

Este valor define o limite máximo de exposição no pior cenário, garantindo que nenhum registo individual apresente um risco de reidentificação superior a 15%.

- **Risco médio (*Average risk*): $\leq 4\%$**

O risco médio reflete o nível global de vulnerabilidade do dataset. Um valor abaixo de 4% assegura que a probabilidade média de reidentificação permaneça muito baixa.

- **Utilidade (*Information utility*): $\geq 90\%$ (equivalente a *information loss* $\leq 10\%$)**

Este requisito assegura que o processo de anonimização preserve a maior parte do valor analítico dos dados, permitindo análises estatísticas fiáveis e mantendo a integridade das distribuições e correlações entre os atributos.

5.3. Hierarquização

A hierarquização dos atributos é uma etapa essencial no processo de anonimização, pois define os diferentes níveis de generalização possíveis para cada variável. Através destas hierarquias, o *ARX* consegue substituir valores específicos por categorias ou intervalos mais amplos, reduzindo o risco de reidentificação enquanto procura manter a utilidade dos dados [7] .

As hierarquias definidas foram as seguintes:

- **age – intervalos etários**, distribuídos por **quatro níveis hierárquicos**. No **nível 0** mantêm-se os valores originais de idade; no **nível 1** as idades são agrupadas em **faixas de 13 anos**; no **nível 2** as faixas passam a **26 anos**; e no **nível 3** é aplicada a **generalização total**, englobando toda a amplitude etária do dataset — dos **17 aos 91 anos**.
- **workclass** – os diferentes tipos de emprego foram agregados em categorias ainda mais gerais, como ***Private*, *Self-employed*, *Government* e *Other***, e no nível 3 foram completamente suprimidos.
- **marital-status** – os estados civis originais foram agrupados, no nível 1, em três categorias: ***Married*, *Single* e *Previously-married*** (viúvo, divorciado, separado). No nível 2 foram completamente suprimidos.
- **race** – as categorias originais foram agregadas em grupos demográficos mais amplos, refletindo apenas **grandes divisões étnicas** (*White*, *Black*, *Other*). No nível 2 foram completamente suprimidos.
- **gender** – como apenas apresenta duas categorias, ambas convergiram para {Female, Male} .
- **native-country** – foi criada uma hierarquia geográfica de três níveis (**País → Continente → Mundo**). Esta hierarquia permite a generalização progressiva da origem dos indivíduos, agrupando os países em regiões geográficas como *Europa*, *América do Norte*, *Ásia* e *América do Sul*.

- **educational-num** – sendo um atributo numérico, foi generalizado por **intervalos de valores**, aumentando progressivamente a amplitude desses intervalos entre o nível 1 ao nível 4. Pode ser observada na figura 3.

Level-0	Level-1	Level-2	Level-3	Level-4
1	[1, 3[[1, 5[[1, 13[[1, 17[
2	[1, 3[[1, 5[[1, 13[[1, 17[
3	[3, 5[[1, 5[[1, 13[[1, 17[
4	[3, 5[[1, 5[[1, 13[[1, 17[
5	[5, 7[[5, 9[[1, 13[[1, 17[
6	[5, 7[[5, 9[[1, 13[[1, 17[
7	[7, 9[[5, 9[[1, 13[[1, 17[
8	[7, 9[[5, 9[[1, 13[[1, 17[
9	[9, 11[[9, 13[[1, 13[[1, 17[
10	[9, 11[[9, 13[[1, 13[[1, 17[
11	[11, 13[[9, 13[[1, 13[[1, 17[
12	[11, 13[[9, 13[[1, 13[[1, 17[
13	[13, 15[[13, 17[[13, 17[[1, 17[
14	[13, 15[[13, 17[[13, 17[[1, 17[
15	[15, 17[[13, 17[[13, 17[[1, 17[
16	[15, 17[[13, 17[[13, 17[[1, 17[

Figura 3

- **occupation** – agrupam-se progressivamente as profissões em **categorias mais amplas**, reduzindo o nível de detalhe para aumentar o grau de anonimização. Esta política de generalização está representada na figura seguinte.

Level-0	Level-1	Level-2	Level-3
Adm-clerical	Office	White-collar	*
Armed-Forces	Military	Public-sector	*
Craft-repair	Manual	Blue-collar	*
Exec-managerial	Management	White-collar	*
Farming-fishing	Agriculture	Blue-collar	*
Handlers-cleaners	Manual	Blue-collar	*
Machine-op-inspect	Manual	Blue-collar	*
Other-service	Service	Blue-collar	*
Priv-house-serv	Service	Blue-collar	*
Prof-specialty	Professional	White-collar	*
Protective-serv	Security	Public-sector	*
Sales	Business	White-collar	*
Tech-support	Technical	White-collar	*
Transport-moving	Transport	Blue-collar	*

Figura 4

5.4. Atribuição de pesos

Com o objetivo de garantir uma maior qualidade dos dados anonimizados, foram atribuídos pesos específicos a cada *quasi-identifying*. No ARX, um QID com **peso mais elevado é tratado como mais relevante**, levando o sistema a preservar **maior detalhe desse atributo** durante o processo de generalização. Os pesos foram definidos com base na influência que cada QID apresenta sobre o atributo sensível *income*.

QID	Peso	Justificação
<i>age</i>	0.60	A idade tem influência significativa no rendimento, visto que o <i>income</i> tende a aumentar com a experiência profissional.
<i>workclass</i>	0.50	O tipo de trabalho apresenta relevância na determinação do rendimento. Contudo, observa-se uma variação significativa do <i>income</i> dentro de cada categoria de <i>workclass</i> . Deste modo optou-se por atribuir um peso intermédio.
<i>educational-num</i>	0.85	É um dos fatores mais fortemente relacionados com o rendimento — maior educação tende a associar-se a maior <i>income</i> .
<i>marital-status</i>	0.35	Embora possa influenciar o <i>income</i> , o impacto é bastante menor comparado com o <i>educational-num</i> .
<i>occupation</i>	0.70	Altamente relevante para o rendimento, já que diferentes ocupações correspondem a diferentes faixas salariais.
<i>race</i>	0.10	A raça apresenta uma correlação muito fraca com o rendimento neste conjunto de dados e, por isso, foi-lhe atribuída um baixo peso.
<i>gender</i>	0.15	Embora existam diferenças médias de rendimento entre géneros, o atributo não apresenta grande relevância para a análise de rendimento.
<i>native-country</i>	0.30	O país de origem tem algum impacto indireto sobre o rendimento, mas como a maioria dos registos pertence a um único país a sua utilidade é reduzida.

5.5. Limite de supressão

O limite de supressão foi definido em **4%**, permitindo que até essa fração dos registos seja removida caso não possam ser generalizados de forma a cumprir as restrições impostas pelos modelos de privacidade que irão ser aplicados.

Uma percentagem inferior poderia limitar a capacidade de atingir os requisitos de anonimização, enquanto valores superiores resultariam numa redução considerável da qualidade e representatividade dos dados.

Considerando que o dataset em estudo não contém informação altamente sensível, mas sim dados socioeconómicos de natureza geral, enquadra-se na categoria de alta utilidade (1%–5%), na qual é aceitável aplicar limites de supressão reduzidos.

Desta maneira, o limite fixado nos 4% revelou-se suficiente para permitir a anonimização eficaz do dataset, assegurando também a preservação da sua utilidade.

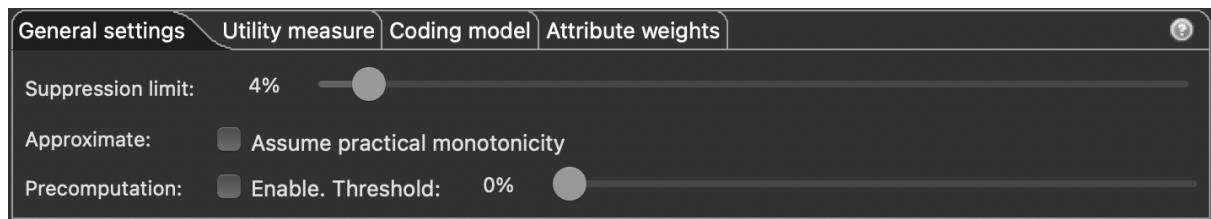


Figura 5

5.6. Supressão e Generalização

A relação entre Supressão e Generalização foi ajustada para **0.70**, passando o ARX a dar maior prioridade à generalização do que à supressão. Esta configuração orienta o processo de anonimização a preservar o maior número possível de registos, aplicando níveis de generalização mais amplos em vez de eliminar entradas individuais.



Figura 6

6. Modelos de Privacidade

Um **modelo de privacidade** estabelece o **conjunto de regras e restrições aplicadas aos dados** para reduzir o **risco de reidentificação** através dos QIDs e **proteger atributos sensíveis**, como o *Income*, assegurando também a utilidade dos dados.

Nesta fase, foram testados vários modelos de privacidade com o propósito de avaliar o seu desempenho face às características específicas deste *dataset* e aos requisitos de privacidade e utilidade estabelecidos anteriormente no ponto 5 deste documento.

Para a anonimização deste *dataset* foram selecionados os modelos de privacidade ***k*-Anonymity** e **ℓ -Diversity**.

6.1. ***k*-Anonymity**

De acordo com a informação que consta no site oficial do software ARX, este modelo de privacidade, amplamente conhecido, tem como objetivo proteger os conjuntos de dados contra a reidentificação, de acordo com ***prosecutor model***. Um *dataset* é considerado *k-anonymous* quando cada registo não pode ser distinguido de, pelo menos, outros $k-1$ registos no que diz respeito aos *quasi-identificadores*. Cada grupo de registos indistinguíveis forma uma *equivalence class* [1].

Para avaliar o impacto do modelo *k-Anonymity* no equilíbrio entre privacidade e utilidade, foram realizados testes com diferentes valores de k , variando entre configurações mais permissivas e mais restritivas.

- Com $k = 4$, os valores de risco máximo de reidentificação, tanto no modelo *Prosecutor* como no *Journalist*, mantinham-se elevados (cerca de 25%), o que não se enquadra nos requisitos estabelecidos anteriormente.
- Por outro lado, ao aumentar para $k = 10$, verificou-se uma redução significativa do risco, mas também um aumento acentuado nas métricas de perda de informação, ultrapassando o nível considerado aceitável para a utilidade dos dados.
- Assim, foi definido $k = 7$ como o valor ideal, uma vez que proporciona um equilíbrio adequado entre privacidade e utilidade, assegurando uma redução substancial do risco de reidentificação sem comprometer de forma significativa a qualidade analítica do dataset.

Os resultados detalhados dos *Attacker Models* (*Prosecutor*, *Journalist* e *Marketer*) serão discutidos mais adiante.

6.2. ℓ -Diversity

O modelo de privacidade ℓ -Diversity deve ser aplicado em complemento ao k -Anonymity quando existe o risco de que uma excessiva homogeneidade nos valores de um **atributo sensível**, em combinação com outros QIDs, possa conduzir a uma perda de privacidade [2].

Este modelo estabelece que cada classe de equivalência formada pelo k -Anonymity deve conter pelo menos ℓ valores distintos para o atributo sensível, assegurando uma diversidade mínima de informação dentro de cada classe. Desta forma, mesmo que um atacante consiga identificar o grupo a que um indivíduo pertence, a probabilidade de inferir o seu valor sensível é reduzida.

No caso deste dataset, o atributo sensível **income** apresenta apenas **5 valores distintos**, o que limita naturalmente a aplicação do modelo ℓ -Diversity. Contudo, optou-se por definir $\ell = 3$, após a realização de testes comparativos com diferentes valores de ℓ (2, 3, 4 e 5).

Os resultados mostraram que:

- Com $\ell = 2$, o risco de reidentificação manteve-se relativamente elevado e a diversidade dos grupos de equivalência foi insuficiente para garantir proteção adequada;
- Com $\ell = 4$ e $\ell = 5$, o risco reduziu ligeiramente, mas à custa de um aumento significativo de *information loss*, devido à necessidade de maior generalização;
- Com $\ell = 3$, observou-se o melhor equilíbrio entre privacidade e utilidade, apresentando o nível mais baixo de *information loss* e um risco de reidentificação já dentro dos limites definidos.

Deste modo, $\ell = 3$ foi considerado o valor mais adequado para este conjunto de dados, assegurando diversidade suficiente nos valores sensíveis e preservando a utilidade do dataset anonimizado.

7. Análise do impacto dos Modelos de Privacidade

7.1. Análise do Risco

7.1.1. Quasi-identifiers

- **Antes da anonimização** (painel esquerdo da *Figura 7*)

Antes da aplicação dos modelos de privacidade, os valores de **Distinção** eram muito elevados, **atingindo valores na casa dos 80%**, o que indica que a maioria das **combinações de QIDs** gerava registos praticamente únicos.

Este cenário traduzia um alto risco de reidentificação, já que um atacante com acesso parcial a atributos como idade, educação ou país de origem poderia muito facilmente identificar um indivíduo no dataset.

A métrica **Separação** apresentava também valores muito elevados, **próximos dos 100%**, revelando que as combinações de atributos separavam eficazmente os registos.

- **Depois da anonimização** (painel direito da *Figura 7*)

Após a aplicação dos modelos de privacidade, observa-se uma redução abrupta da **Distinção**, em que os valores **nunca ultrapassam 1.4%**.

Esta diminuição acentuada da **Distinção** reflete o efeito do *k*-Anonymity, que torna os registos indistinguíveis dentro de classes de equivalência.

Contudo, a **Separação** manteve-se praticamente inalterada. Isto é explicado porque esta métrica mede a capacidade estrutural das combinações de atributos para diferenciar registos, e não é fortemente afetada pela generalização ou pela supressão. Uma **Separação elevada** é desejável quando acompanhada de uma **Distinção reduzida**, pois indica que a **anonimização foi eficaz** (eliminando registos únicos) sem destruir a **estrutura informativa** do dataset. Este equilíbrio reflete um **bom compromisso entre privacidade e utilidade**, assegurando que o risco de reidentificação é mínimo enquanto a qualidade analítica dos dados se mantém.

Distribution of risks	Quasi-identifiers	Attacker models	HIPAA identifiers	Distinction	Separation	
Quasi-identifier						
age, workclass, marital-status, race, gender, native-country	38.98%	99.740693%				
age, educational-num, marital-status, race, gender, native-country	46.98%	99.89664%				
age, workclass, educational-num, race, gender, native-country	48.32%	99.89368%				
age, marital-status, occupation, race, gender, native-country	51.66%	99.93497%				
age, workclass, occupation, race, gender, native-country	51.98%	99.93548%				
age, workclass, educational-num, marital-status, race, native-country	53.7%	99.90196%				
age, workclass, educational-num, marital-status, gender, native-country	55.42%	99.92346%				
age, workclass, educational-num, marital-status, race, gender	57.74%	99.93244%				
age, workclass, marital-status, occupation, race, native-country	58.96%	99.94621%				
age, workclass, marital-status, occupation, gender, native-country	59.7%	99.95293%				
age, workclass, marital-status, occupation, race, gender	61.48%	99.96652%				
age, educational-num, occupation, race, gender, native-country	64.98%	99.97%				
age, workclass, educational-num, occupation, race, native-country	69.04%	99.97204%				
age, educational-num, marital-status, occupation, race, native-country	70.58%	99.97489%				
age, workclass, educational-num, occupation, gender, native-country	71.48%	99.97647%				
age, educational-num, marital-status, occupation, gender, native-country	71.76%	99.97808%				
age, educational-num, marital-status, occupation, race, gender	73.82%	99.98027%				
age, workclass, educational-num, occupation, race, gender	74.1%	99.9794%				
age, workclass, educational-num, marital-status, occupation, native-country	75.88%	99.97964%				
age, workclass, educational-num, marital-status, occupation, race	77.92%	99.98212%				
age, workclass, educational-num, marital-status, occupation, gender	78.5%	99.98416%				
workclass, educational-num, marital-status, occupation, race, gender, native-co...	37.24%	99.70742%				
age, workclass, educational-num, marital-status, race, gender, native-country	60.52%	99.93816%				
age, workclass, marital-status, occupation, race, gender, native-country	64.76%	99.96115%				
age, educational-num, marital-status, occupation, race, gender, native-country	75.58%	99.98183%				
age, workclass, educational-num, occupation, race, gender, native-country	75.98%	99.98112%				
age, workclass, educational-num, marital-status, occupation, race, native-country	79.52%	99.98364%				
age, workclass, educational-num, marital-status, occupation, gender, native-cou...	80.56%	99.98591%				
age, workclass, educational-num, marital-status, occupation, race, gender	82.2%	99.98743%				
age, workclass, educational-num, marital-status, occupation, race, gender, nativ...	83.4%	99.9884%				

Figura 7

7.1.2. Attacker models

- Prosecutor attacker model

No modelo **Prosecutor**, o atacante conhece uma pessoa específica e dispõe de informação parcial sobre a mesma (ex.: idade, profissão). O seu objetivo é confirmar a presença do indivíduo no conjunto de dados e, caso exista correspondência, descobrir o valor de um **atributo sensível** (ex.: *income*) [8].

- Journalist attacker model

No modelo **Journalist**, o atacante não persegue uma pessoa em específico, nem sabe necessariamente que um indivíduo concreto está presente no *dataset*. Em vez disso, procura encontrar algum registo no *dataset* que possa ser reidentificado. O atacante usa técnicas de *linkage*¹ para encontrar correspondências [8].

- Marketer attacker model

No modelo **Marketer**, o atacante tenta reidentificar o maior número possível de pessoas. Ao contrário dos outros modelos, o foco está na escala do ataque, em vez de casos individuais [8].

¹ Cruzamento de dados de diferentes fontes com o objetivo de identificar indivíduos.

○ Antes da anonimização

A Figura 8 apresenta a avaliação do risco de reidentificação segundo os 3 modelos de atacante disponibilizados pelo ARX.

Os resultados obtidos evidenciam um nível de **risco extremamente elevado** em todos os cenários analisados. Nos modelos **Prosecutor** e **Journalist**, o risco máximo atinge 100%, e quase toda a totalidade dos registo apresenta alguma vulnerabilidade, indicando que **quase todos os registos poderiam ser reidentificados** por um atacante que possuísse um **conhecimento parcial sobre um indivíduo**.

O modelo **Marketer** apresentou um **sucess rate de 83.4%**, demonstrando que até um atacante genérico, focado em reidentificar o maior número de pessoas, teria uma elevada probabilidade de sucesso.

Além disso, cerca de 73% dos registo foram classificados como **sample uniques** no conjunto de dados, o que significa que essas combinações de QIDs são **únicas dentro da amostra** e, portanto, **altamente suscetíveis a reidentificação** caso um atacante disponha de alguma informação.

Estes resultados confirmam que o dataset, na sua forma original, **não garante qualquer nível de anonimato**, apresentando **riscos de reidentificação incompatíveis com os princípios de proteção de dados pessoais** definidos pelo **Regulamento (UE) 2016/679 — Regulamento Geral sobre a Proteção de Dados (RGPD)** [3].

○ Depois da anonimização

Após o processo de anonimização, observou-se uma **redução drástica do risco de reidentificação em todos os attacker models**. Essa redução pode ser visualizada na Figura 9.

No **Marketer**, o **sucess rate** caiu de 83.4% para apenas 1.34%. Nos modelos **Prosecutor** e **Journalist**, o risco máximo estimado situa-se agora em cerca de 14%, e a percentagem de registo vulneráveis é residual.

Esses valores são aceitáveis face à média-baixa sensibilidade do dataset.

Neste momento, a probabilidade de um atacante identificar corretamente um indivíduo específico é bastante reduzida.



Figura 8



Figura 9

7.2. Análise da Utilidade

7.2.1. *Attribute-level quality*

A tabela da Figura 10 evidencia a qualidade dos atributos após o processo de anonimização, considerando métricas como ***Generalization Intensity***, ***Granularity***, ***Non-Uniform Entropy*** e ***Squared error***.

Observa-se que os atributos *age*, *workclass*, *education-num* e *occupation* apresentam valores elevados de **granularidade** (superiores a 60%), o que indica que mantêm uma boa capacidade informativa e continuam adequados para análises estatísticas.

Destaca-se o facto de o atributo ***education-num*** ter uma **elevada intensidade de generalização** (72.435%) evidenciando um nível significativo de abstração dos valores originais. Apesar da elevada intensidade de generalização observada no atributo *educational-num*, a sua **granularidade manteve-se elevada**. Este comportamento ocorre porque o **processo de anonimização agrupou valores semelhantes em categorias mais amplas**, reduzindo o risco de reidentificação, **mas preservou as diferenças estatísticas relevantes entre grupos educativos**. Assim o atributo continua a possuir alta utilidade analítica.

Os atributos ***marital-status*, *race* e *gender*** apresentam **100% de valores ausentes**. Este resultado pode dever-se ao **peso reduzido atribuído a estes QIDs**, o que faz com que o *ARX* não os considere. Além disso, estes atributos tendem a **aumentar o risco do dataset**, sem acrescentarem valor significativo em termos de utilidade analítica, uma vez que possuem baixa correlação com o atributo *income*.

Attribute-level quality						
Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
<i>age</i>	String	3.42%	32.19333%	60.6591%	17.61752%	85.01823%
<i>workclass</i>	String	3.42%	84.17%	90.16667%	72.93182%	79.01896%
<i>educational-...</i>	String	3.42%	72.435%	90.14133%	65.2312%	95.03923%
<i>marital-status</i>	String	100%	0%	0%	0%	0%
<i>occupation</i>	String	3.42%	32.19333%	60.72154%	31.57036%	9.02518%
<i>race</i>	String	100%	0%	0%	0%	0%
<i>gender</i>	String	100%	0%	0%	0%	0%
<i>native-country</i>	String	3.42%	32.19333%	2.4145%	0%	0%

Figura 10

7.2.2. Análise geral da utilidade do dataset

○ Antes da anonimização

Antes da aplicação dos modelos de anonimização, o conjunto de dados mantinha 100% da sua utilidade informacional, uma vez que todos os atributos estavam disponíveis no seu formato original e com total granularidade.

Contudo, essa condição resultava num risco de reidentificação muito elevado, tornando o dataset inadequado para partilha ou uso em contextos que exigem conformidade com normas de proteção de dados.

○ Depois da anonimização

Após o processo de anonimização, o ARX calculou uma perda média de informação (information loss) de apenas 1,30% [8].

Este valor é médio-baixo, indicando que o processo de anonimização preservou grande parte da estrutura e a utilidade do conjunto de dados original.

A baixa perda de informação deve-se ao equilíbrio alcançado entre:

- Generalização (0.70), que agregou valores apenas quando necessário para atingir o nível de anonimato exigido;
- Supressão (0.30), evitando a eliminação de regtos e, consequentemente, reduzindo a distorção dos resultados analíticos.

Summary statistics		Distribution	Contingency	Class sizes	Properties	Classification models			
Property	Value								
Score	0.14156247632259422 [1.30854%]								
Successors	5								
Predecessors	7								
Transformation	[2, 1, 2, 2, 2, 0, 2]								
▼ Anonymity	k-anonymity								
k	7								
▼ Anonymity	Distinct l-diversity								
l									

O valor de 1,30% de *loss* demonstra que a utilidade analítica do dataset foi de forma mantida, mesmo após garantir elevados níveis de privacidade.

8. Conclusão

A análise conjunta do risco e da utilidade evidencia que o processo de anonimização implementado foi altamente eficaz, reduzindo o risco médio de reidentificação de valores superiores a 80% para apenas 1,34%, enquanto a perda de informação se manteve mínima (1,30%).

Estes resultados demonstram que o método aplicado, baseado nos modelos k -Anonymity e ℓ -Diversity, atingiu plenamente o objetivo proposto de garantir a privacidade dos dados sem comprometer a sua utilidade analítica.

Assim, pode concluir-se que foi alcançado um equilíbrio ótimo entre proteção da privacidade e preservação da qualidade da informação, validando a eficácia da estratégia de anonimização adotada.

Estes resultados demonstram que é possível implementar técnicas de anonimização eficazes, conciliando a proteção da privacidade com a manutenção do valor informativo dos dados, contribuindo para práticas de gestão de dados mais seguras e conformes com o RGPD.

9. Referências

- [1] ARX Data Anonymization Tool, “Privacy criteria overview.” [Online]. Disponível em: <https://arx.deidentifier.org/overview/privacy-criteria/> [Acedido a 16 de outubro de 2025].
- [2] Data Anonymization Methods, “l-Diversity” [Online]. Disponível em: <https://help.sap.com/docs/hana-cloud-database/sap-hana-cloud-sap-hana-database-data-anonymization-guide/l-diversity/> [Acedido a 17 de outubro de 2025].
- [3] European Union, Regulamento (EU) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016 [Online]. Disponível em: <https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679>. [Acedido a 15 de outubro de 2025].
- [4] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002. [Online]. Disponível em: <https://doi.org/10.1142/S0218488502001648> [Acedido em: 17 de outubro de 2025].
- [5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy Beyond k-Anonymity,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.
- [6] J. Domingo-Ferrer and V. Torra, “A Critique of k-Anonymity and Some of Its Enhancements,” in *Proceedings of the IEEE International Conference on Availability, Reliability and Security (ARES)*, 2008. [Online]. Disponível em: <https://doi.org/10.1109/ARES.2008.32> [Acedido em: 17 de outubro de 2025].
- [7] ARX Documentation, ‘Data Transformation and Attribute Weighting,’ [Online]. Disponível em: <https://arx.deidentifier.org/development/data-transformations/> [Acedido em: 17 de outubro de 2025].
- [8] ARX User Guide, ‘Risk and Utility Analysis,’ [Online]. Disponível em: <https://arx.deidentifier.org/development/risk-analysis/> .[Acedido a 17 de outubro de 2025].