

# Anonimização de Dataset com Análise de Risco e Utilidade

CURSO: IACD

DISCIPLINA: SEGURANÇA E PRIVACIDADE

DATA: MAIO DE 2025

GRUPO: ANA SOFIA QUINTERO, CATARINA ABRANTES E  
LILIANA SILVA

# Índice

<b>1. Introdução.....</b>	<b>2</b>
<b>2. Descrição do Dataset e Classificação dos Atributos.....</b>	<b>3</b>
2.1 Descrição do Dataset.....	3
2.2 Classificação dos Atributos.....	3
<b>3. Análise de Risco do Dataset Original.....</b>	<b>5</b>
3.1 Métricas Gerais de Risco (Average, Highest Risk, % Sample Uniques).....	5
3.2 Modelos de Ataque (Prosecutor, Journalist, Marketer) .....	5
<b>4. Preparação e Modelos de Anonimização.....</b>	<b>7</b>
4.1 Preparação do Dataset e Hierarquias de Pré-processamento .....	7
4.2 Modelo 1: k-Anonymity (k=25) + l-Diversity (l=2) .....	8
4.2.1 Estratégia e Parâmetros .....	8
4.2.2 Análise de Risco Após Anonimização.....	8
4.2.3 Métricas de Utilidade .....	9
4.2.4 Configuração de Pesos e Generalizações.....	11
<b>5. Modelo 2: k-Anonymity (k=25) + t-Closeness (t=0.2).....</b>	<b>12</b>
5.1 Estratégia e Parâmetros .....	12
5.2 – Análise de Risco Após Anonimização .....	12
5.3 Métricas de Utilidade .....	13
<b>6. Avaliação da Variação de Parâmetros dos Modelos.....</b>	<b>14</b>
6.1 Limite de Supressão vs Utilidade (Discernibility) .....	15
6.2 Valor de k vs Utilidade dos Dados.....	16
6.3 Valor de k vs Risco de Reidentificação.....	17
6.4 Conclusão da Avaliação Paramétrica .....	18
<b>7. Comparação Final entre os Modelos.....</b>	<b>19</b>
<b>8. Conclusão Final.....</b>	<b>21</b>

## 1. Introdução

O presente trabalho tem como objetivo realizar a anonimização de um conjunto de dados de maior dimensão, utilizando a ferramenta ARX. A anonimização tem como finalidade a proteção da privacidade individual, reduzindo o risco de reidentificação, enquanto se procura preservar a utilidade dos dados para análise. Foram aplicados diferentes modelos de privacidade, nomeadamente o *k-anonymity* e o *l-diversity*, sendo avaliado o seu impacto tanto no risco de reidentificação como na utilidade dos dados. Adicionalmente, foram estudadas variações de parâmetros como o limite de supressão e os valores de *k*.

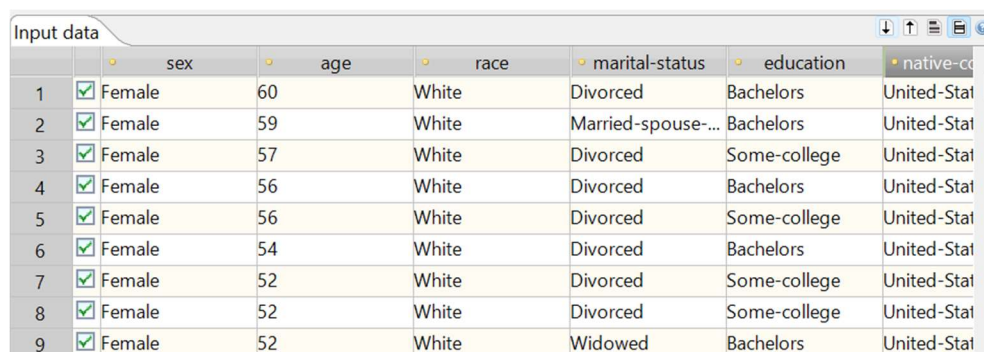
Para uma avaliação abrangente, os modelos foram analisados quanto ao risco de reidentificação, às métricas de utilidade e ao impacto da variação de parâmetros, permitindo assim uma comparação detalhada entre diferentes abordagens de anonimização.

## 2. Descrição do Dataset e Classificação dos Atributos

### 2.1 Descrição do Dataset

O dataset utilizado neste trabalho foi obtido do site ARX Deidentifirer, que fornece projetos exemplos (o ficheiro *exemple.deid*) para este tipo de testes de anonimização. Trata-se de um conjunto de dados com informações demográficas e socioeconómicas, que incluem atributos como idade, sexo, nível de educação, ocupação, entre outros categorizados com respeito o nível de sensibilidade.

Este tipo de dados é particularmente sensível, uma vez que a presença de atributos potencialmente identificáveis — conhecidos como *quasi-identificadores* — permite a possibilidade de reidentificação de indivíduos, sobretudo quando cruzado com bases de dados externas. A elevada diversidade e granularidade dos valores registados tornam este dataset um caso adequado para a aplicação de técnicas de anonimização baseadas em *k-anonymity*, *l-diversity* e *t-closeness*.



	sex	age	race	marital-status	education	native-co
1	Female	60	White	Divorced	Bachelors	United-Stat
2	Female	59	White	Married-spouse...	Bachelors	United-Stat
3	Female	57	White	Divorced	Some-college	United-Stat
4	Female	56	White	Divorced	Bachelors	United-Stat
5	Female	56	White	Divorced	Some-college	United-Stat
6	Female	54	White	Divorced	Bachelors	United-Stat
7	Female	52	White	Divorced	Some-college	United-Stat
8	Female	52	White	Divorced	Some-college	United-Stat
9	Female	52	White	Widowed	Bachelors	United-Stat

Figura 1 - interface do ARX com o head do dataset original carregado

### 2.2 Classificação dos Atributos

A classificação dos atributos foi realizada com base na análise das métricas **distinction** e **separation**, disponíveis na ferramenta ARX, no separador *Attribute Metadata*. Estas métricas são essenciais para identificar os atributos com maior potencial de reidentificação.

- **Distinction** quantifica a proporção de valores únicos de um atributo, refletindo a sua capacidade de distinguir registos individuais no conjunto de dados.
- **Separation** indica a proporção de valores que, isoladamente, permitem separar inequivocamente os registos, apontando assim para um elevado risco de identificação.

Distribution of risks			Quasi-identifiers	Attacker models	HIPAA identifiers
Quasi-identifier			Distinction	Separation	
sex			0.03514%	41.26559%	
salary-class			0.03514%	44.80569%	
race			0.08784%	21.52448%	
marital-status			0.10541%	56.81889%	
workclass			0.12298%	57.92789%	
occupation			0.22839%	89.04276%	
education			0.2811%	82.05174%	
age			0.6676%	94.38234%	
native-country			0.70274%	15.25396%	

Figura 2 – Valores de distinction e separation dos atributos no dataset original

A tabela seguinte resume a classificação dos atributos, com base nestas métricas e respetiva análise contextual:

Atributo	Classificação	Distinction	Separation	Justificação
Salary-class	Sensível	Média	Média	Informação financeira sensível, protegida legalmente
Sex	QID	Baixa	Moderada	Contribui para reidentificação quando combinada com outros atributos
Race	QID	Baixa	Moderada	Potencial de reidentificação combinada com outros atributos
Marital-status	QID	Baixa	Moderada	Informação pessoal identificável em conjunto
Workclass	QID	Alta	Baixa	Variedade elevada de categorias, risco moderado
Native-country	QID	Alta	Alta	Países raros aumentam unicidade dos registos
Occupation	QID	Alta	Alta	Profissões específicas são fortes identificadores
Education	QID	Alta	Alta	Elevada granularidade, disponível publicamente
Age	QID	Muito Alta	Muito Alta	Valor contínuo e específico, elevado risco de identificação

Esta classificação demonstra que todos os atributos, à exceção do sensível, funcionam como quasi-identificadores (QIDs), dado o seu potencial de reidentificação, especialmente quando combinados entre si. O atributo **age**, em particular, apresenta os valores mais críticos, justificando a necessidade de **forte generalização** nos modelos a aplicar.

### 3. Análise de Risco do Dataset Original

#### 3.1 Métricas Gerais de Risco (Average, Highest Risk, % Sample Uniques)

Antes da aplicação de qualquer técnica de anonimização, foi realizada uma análise de risco global sobre o dataset original. Os principais indicadores obtidos através da ferramenta ARX evidenciam um cenário preocupante:

Overview Population uniques Quasi-identifiers		
Measure	Value [%]	
Lowest prosecutor risk	4.54545%	
Records affected by lowest risk	0.38651%	
Average prosecutor risk	75%	
Highest prosecutor risk	100%	
Records affected by highest risk	62.86015%	
Estimated prosecutor risk	100%	
Estimated journalist risk	100%	
Estimated marketer risk	75%	
Sample uniques	62.86015%	
Population uniques	3.55193%	
Population model	PITMAN	
Quasi-identifiers	age, education, marital-status, native-country, occupation, race, sex, workclass	

Figura 3 - Análise de dados e respectivos riscos dos modelos de atacante

Estes valores indicam que uma larga proporção dos registos é única dentro da amostra, o que, quando combinado com atributos classificáveis, como quasi-identificadores, representa um risco real de reidentificação.

#### 3.2 Modelos de Ataque (Prosecutor, Journalist, Marketer)

A ferramenta **ARX** permite simular cenários de risco com diferentes **perfis de atacante**, cada um representando distintos níveis de conhecimento prévio sobre os dados. No presente trabalho, foram analisados os três **modelos clássicos**:

- **Promotor (Prosecutor Attacker Model):**

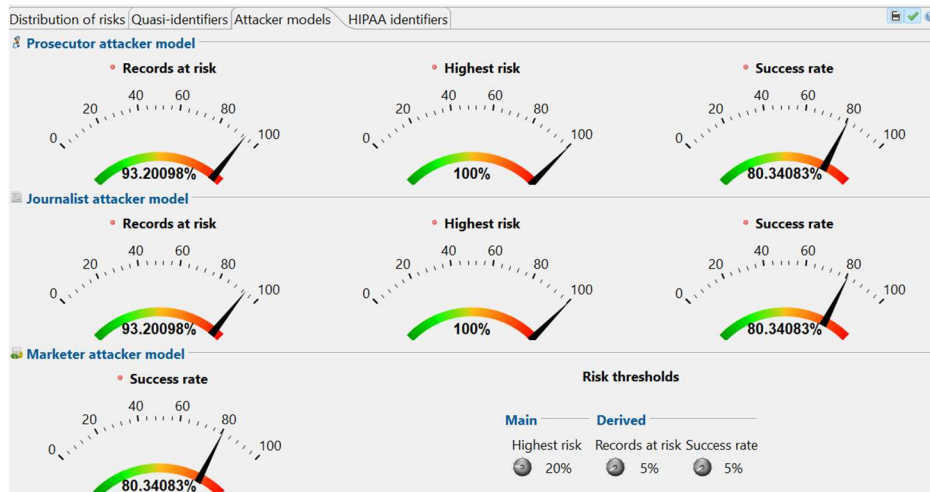
O atacante possui conhecimento detalhado sobre um indivíduo específico (por exemplo, idade, sexo, ocupação) e tenta reidentificá-lo nos dados anonimizados. A anonimização deve ser suficientemente robusta para impedir essa reidentificação, mesmo perante um nível elevado de informação prévia.

- **Jornalista (Journalist Attacker Model):**

O atacante não tem um alvo concreto, mas procura descobrir factos relevantes ou sensíveis através de ligações simples entre atributos públicos. A anonimização deve proteger os dados contra este tipo de inferência baseada em correlações superficiais.

- **Profissional de Marketing (Marketer Attacker Model):**

Este atacante foca-se na análise de grupos e tendências, em vez de indivíduos. O seu objetivo é identificar padrões para efeitos de segmentação. A anonimização deve garantir a proteção de detalhes individuais, mantendo simultaneamente a utilidade estatística dos dados agregados.



*Figura 4 - Risco de cada modelo de atacante no dataset original*

Em todos os modelos de atacante — Promotor, Jornalista e Profissional de Marketing — a proporção de registos em risco é extremamente elevada (93,2%), e o risco máximo de reidentificação atinge os 100%. A taxa de sucesso dos atacantes é também elevada (80,34%), o que demonstra que, independentemente do nível de conhecimento prévio ou das motivações do atacante, existe uma elevada probabilidade de reidentificar corretamente os indivíduos presentes no conjunto de dados.

Este resultado era expectável, tendo em conta que os dados se encontram ainda no seu formato original, sem qualquer técnica de anonimização aplicada. Um risco máximo de 100% implica que existem combinações de atributos capazes de identificar inequivocamente determinados indivíduos. Adicionalmente, a média de risco é extremamente elevada, com cerca de 3 em cada 4 registos a apresentarem uma probabilidade significativa de reidentificação.

Estes dados justificam, de forma inequívoca, a necessidade de aplicar modelos de anonimização que reduzam a granularidade dos atributos e mitiguem o risco de exposição individual.

## 4. Preparação e Modelos de Anonimização

### 4.1 Preparação do Dataset e Hierarquias de Pré-processamento

Antes de aplicar os modelos de anonimização, foi necessário preparar o dataset e definir hierarquias de generalização para os atributos quasi-identificadores. Estas hierarquias são essenciais para permitir uma redução controlada da granularidade e garantir a consistência dos dados anonimizados.

Para a anonimização de dados, especialmente dos quasi-identificadores (QID), foram desenvolvidas hierarquias que permitem substituir os valores originais por outros mais generalizados. O ARX oferece a possibilidade de construir hierarquias com diferentes níveis de anonimização dos dados, considerando o domínio do atributo.

Procedemos à modificação das hierarquias, nomeadamente dos atributos age e sex. Para os outros serão mantidas as que foram implementadas automaticamente, já que não afetam os resultados.

As hierarquias definidas para o atributo sex podem ser relativamente simples ou limitadas, mas são por vezes necessárias para assegurar uma maior generalização e garantir a privacidade.

Level-0	Level-1
Female	{Female, Male}
Male	{Female, Male}

Figura 5 - Hierarquia Sex com um nível

As hierarquias definidas para o atributo age permitem generalizá-lo em intervalos mais amplos, reduzindo assim a unicidade dos dados e o risco de identificação.

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	
17	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
18	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
19	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
20	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
21	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
22	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
23	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
24	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*

Figura 6 - Hierarquias com intervalos de 5 níveis



## 4.2 Modelo 1: k-Anonymity (k=25) + l-Diversity (l=2)

### 4.2.1 Estratégia e Parâmetros

Para o primeiro modelo de anonimização foi aplicada uma combinação de **k-anonymity** com  $k = 25$  e **l-diversity** com  $l = 2$ .

**k-anonymity (k = 25):** Garante que para cada combinação de quasi-identificadores, existam pelo menos 25 registos iguais, tornando difícil isolar indivíduos.

**l-diversity (l = 2):** Assegura que, dentro de cada grupo de k registos, existam pelo menos 2 valores distintos no atributo sensível, evitando inferências sobre dados confidenciais.

### 4.2.2 Análise de Risco Após Anonimização

Após a aplicação do Modelo 1, que combina **k-anonymity (k=25)** com **l-diversity (l=2)**, observou-se uma **redução significativa nos níveis de risco de reidentificação**.



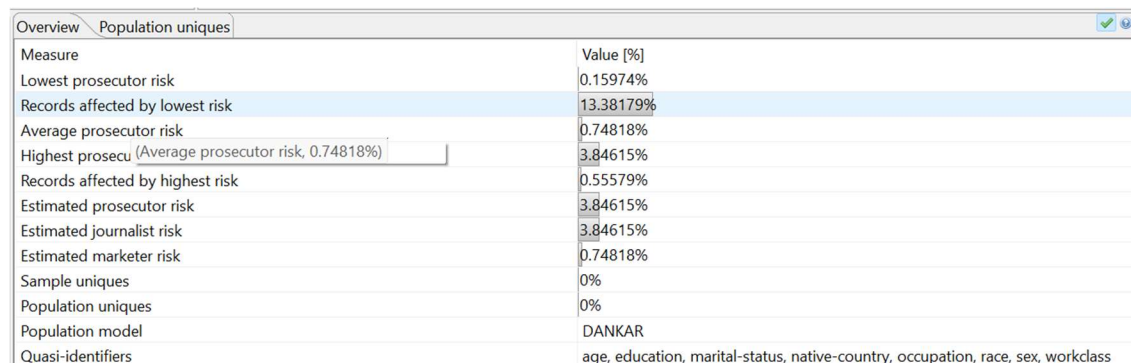
Figura 7 - Risco de cada atacante após a aplicação do modelo 1

Verifica-se uma **redução clara na taxa média de sucesso dos atacantes**:

- **Records at risk: 0%**
- **Highest risk: 3.84% (Prosecutor), 1.75% (Journalist)**
- **Success rate: ~0.16-0.74%**

**Indicadores drasticamente reduzidos → alta eficácia do modelo.** Além disso, todos os modelos reportam **0% de registos em risco**, o que confirma a eficácia da generalização em eliminar casos diretamente identificáveis.

Observa-se uma redução superior a 90% nos indicadores médios de risco e nas taxas de sucesso de ataque em todos os modelos de atacante (Prosecutor, Journalist, Marketer). Apesar do “highest risk” residual, de 3,84%, as taxas de sucesso reais de ataque caíram para valores entre 0,16% e 0,74%, o que demonstra que os dados anonimizados apresentam um perfil de risco mínimo.



Measure	Value [%]
Lowest prosecutor risk	0.15974%
Records affected by lowest risk	13.38179%
Average prosecutor risk	0.74818%
Highest prosecu (Average prosecutor risk, 0.74818%)	3.84615%
Records affected by highest risk	0.55579%
Estimated prosecutor risk	3.84615%
Estimated journalist risk	3.84615%
Estimated marketer risk	0.74818%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, native-country, occupation, race, sex, workclass

*Figura 8 - Estatística de risco e unicidade após aplicação do Modelo 1*

A Figura 8 complementa a análise ao mostrar os indicadores detalhados de risco. Os principais destaques incluem:

- **Average prosecutor risk:** 0.74818%
- **Sample uniques:** 0%
- **Population uniques:** 0%
- **Registos afetados por risco elevado:** 0%

Estes valores demonstram que não existe qualquer registo diretamente identificável ou único, e que a generalização aplicada foi suficiente para eliminar a unicidade.

Em conjunto, estas métricas confirmam que o Modelo 1 atingiu o seu objetivo: proteger eficazmente a identidade dos indivíduos sem comprometer a integridade dos dados, reduzindo a probabilidade de reidentificação a níveis residuais e assegurando diversidade no atributo sensível salary-class.

## 4.2.3 Métricas de Utilidade

Após a aplicação do Modelo 1, foram avaliadas diversas métricas de utilidade da ferramenta ARX. As métricas apresentadas permitem quantificar o impacto da anonimização na **granularidade e qualidade informacional dos dados**.

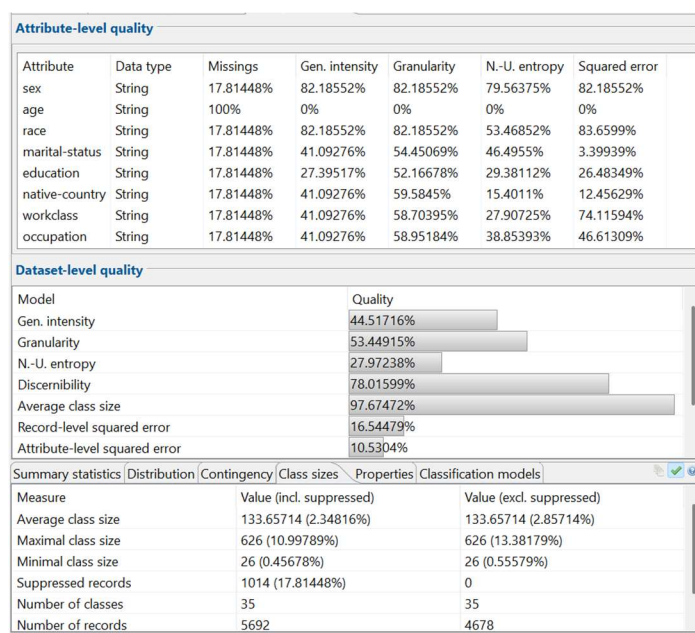


Figura 9 - Métricas de utilidade dos dados após aplicação do Modelo 1

A análise da Figura 9 revela os seguintes pontos principais:

- **Granularidade:** 53.44915%

Indica que quase metade da granularidade original foi preservada, mantendo os dados relativamente específicos.

- **Non-uniform Entropy (Entropia não-uniforme):** 27.97238%

Esta métrica representa a penalização por agrupamento; valores elevados significam menos granularidade, mas o valor aqui obtido permanece aceitável

- **Discernibility:** 78.01599%

Esta métrica representa a penalização associada ao agrupamento; valores elevados correspondem a menor granularidade, mas o valor aqui obtido mantém-se dentro de limites aceitáveis.

- **Erro quadrático médio por registo:** 16.54479%

Representa o desvio médio entre os valores originais e os dados anonimizados, situando-se dentro de uma margem considerada aceitável.

- **Tamanho médio das classes (Average class size):** 97.67472%

Mostra que os registos foram agrupados em classes amplas, o que assegura o cumprimento da condição de k-anonymity com robustez.

Estes resultados confirmam que, apesar da aplicação de dois modelos de proteção (k e l), a utilidade dos dados foi preservada de forma satisfatória, permitindo a realização de análises estatísticas significativas. A perda de granularidade era expectável, mas encontra-se claramente controlada e justificada pelo ganho em privacidade.

#### 4.2.4 Configuração de Pesos e Generalizações

Na avaliação da utilidade, foi adotada uma estratégia personalizada de ponderação dos atributos, refletindo a sua **importância relativa no contexto da análise**. Esta configuração teve como objetivo **preservar com maior rigor a qualidade de certos atributos** considerados mais relevantes para futuras análises estatísticas.

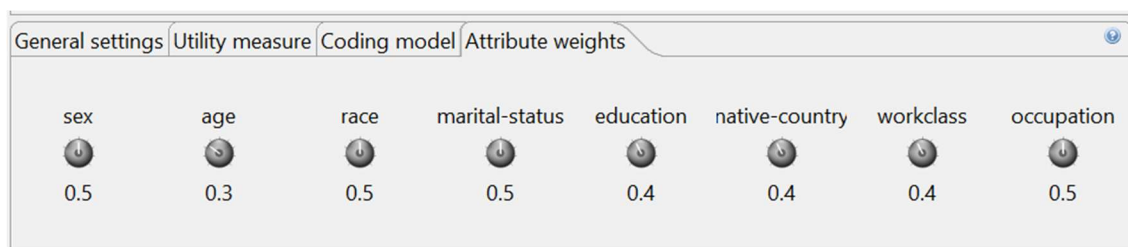


Figura 10 - Configuração dos pesos atribuídos no modelo de utilidade

A imagem mostra os pesos atribuídos a cada atributo durante a avaliação da utilidade, destacando-se:

- Atributos com **maior peso (0.5)**: *sex*, *race* e *occupation* – escolhidos pelo seu forte poder de distinção e interesse analítico.
- Atributos com **peso intermédio (0.4)**: *education*, *native-country*, *workclass* – relevantes, mas com menor sensibilidade.
- Atributos com **peso mais baixo (0.3)**: *age* – devido à sua elevada variabilidade e à necessidade de generalização mais agressiva.

Esta configuração permitiu um **compromisso eficaz entre anonimização e utilidade**, priorizando a preservação da informação em atributos mais críticos e aceitando maior degradação naqueles menos relevantes para as análises pretendidas.

Adicionalmente, as **hierarquias de generalização** previamente definidas foram mantidas neste modelo. O atributo *age*, por exemplo, foi generalizado em **intervalos de 10 anos**. Já *education* e *occupation* foram agrupados em **categorias funcionais**, garantindo diversidade sem comprometer totalmente a capacidade analítica.

Em suma, o Modelo 1 demonstrou ser eficaz na redução dos riscos de reidentificação, mantendo uma utilidade global aceitável. A combinação de *k-anonymity* e *l-diversity*, aliada a generalizações bem calibradas e ponderações estratégicas, permitiu um equilíbrio sólido entre privacidade e qualidade dos dados.

## 5. Modelo 2: k-Anonymity (k=25) + t-Closeness (t=0.2)

### 5.1 Estratégia e Parâmetros

Neste segundo modelo, foi aplicada a técnica de **k-anonymity** com  $k = 25$ , combinada com **t-closeness** com  $t = 0.2$ . Esta configuração visa não apenas garantir que cada registro é indistinguível de, pelo menos, 24 outros (k-anonymity), mas também que **a distribuição do atributo sensível dentro de cada grupo de equivalência não diverge significativamente da distribuição global** (t-closeness).

O atributo sensível utilizado foi novamente *salary-class*. Para cumprir as exigências de t-closeness, foram necessárias **generalizações mais agressivas** em atributos como *age*, *education*, *occupation* e *workclass*. A ferramenta ARX ajustou automaticamente as transformações de modo a respeitar o limiar de  $t = 0.2$ , com vista a **reduzir o risco de inferência estatística** por parte de um atacante.

Esta configuração é particularmente importante em contextos com dados sensíveis, onde não basta garantir anonimato — é essencial **proteger também contra ataques por aproximação ou distribuição**.

### 5.2 – Análise de Risco Após Anonimização

Após a aplicação do Modelo 2, que combina **k-anonymity (k=25)** com **t-closeness (t=0.2)**, os valores máximos de risco mantiveram-se semelhantes aos do Modelo 1. No entanto, observou-se uma **ligeira subida no risco médio de reidentificação**, especialmente no modelo *Prosecutor*, onde o **Average Risk** aumentou para **0.7799%**.



Figura 11 - Riscos por tipo de atacante após aplicação do Modelo 2

Como apresentado na Figura 11:

- O **Highest Risk** manteve-se em **3.84615%** no modelo *Prosecutor* e **1.75439%** no modelo *Journalist*.
- A **taxa de sucesso dos atacantes** foi de **0.7799%** no *Prosecutor* e **0.17071%** nos modelos *Journalist* e *Marketer*.
- Todos os modelos apresentam **0% de registos em risco**, o que indica que **nenhum registo está diretamente identificável** após a anonimização.

Para complementar esta análise, foi também avaliado o painel de resumo estatístico com os indicadores globais de risco.

Overview	Population uniques	
Measure		Value [%]
Lowest prosecutor risk		0.22222%
Records affected by lowest risk		19.4974%
Average prosecutor risk		0.7799%
Highest prosecutor risk		3.84615%
Records affected by highest risk		1.12652%
Estimated prosecutor risk		3.84615%
Estimated journalist risk		3.84615%
Estimated marketer risk		0.7799%
Sample uniques		0%
Population uniques		0%
Population model		DANKAR
Quasi-identifiers		age, education, marital-status, native-country, occupation, race, sex, workclass

Figura 12 - Estatística de risco e unicidade após aplicação do Modelo 2

A Figura 12 mostra:

- **Average prosecutor risk:** 0.7799%
- **Highest prosecutor risk:** 3.84615%
- **Registos afetados por risco elevado:** 1.12652%
- **Registos afetados pelo risco mais baixo:** 19.4974%
- **Sample uniques e population uniques:** 0%
- **Population model:** DANKAR

Estes resultados comprovam que o Modelo 2 mantém o **anonimato a nível estrutural (k-anonymity)** e introduz uma camada adicional de **proteção estatística contra inferência (t-closeness)**, sendo especialmente eficaz em cenários sensíveis. O aumento ligeiro no risco médio é compensado pela **eliminação total da unicidade** e pela **preservação da coerência estatística** do atributo sensível *salary-class*.

## 5.3 Métricas de Utilidade

As métricas de utilidade após a aplicação do Modelo 2 revelam um **impacto mais severo na granularidade e qualidade dos dados**, como seria de esperar com a imposição de t-closeness.

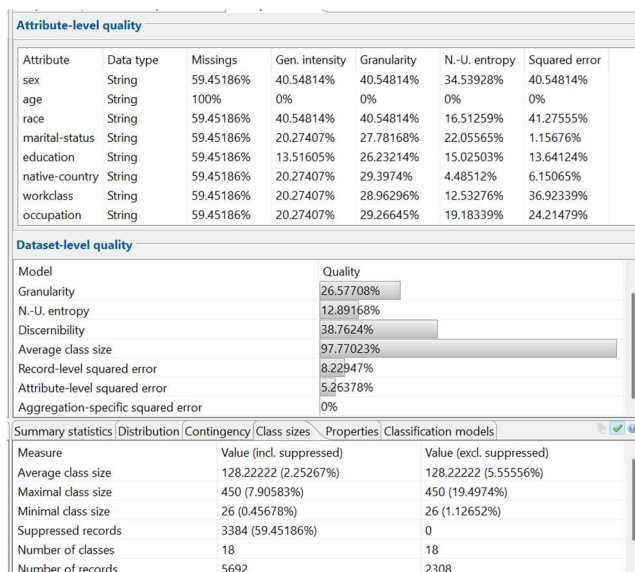


Figura 13 - Métricas de utilidade dos dados após aplicação do Modelo 2

Analisando a **Figura 13**, observam-se os seguintes indicadores de qualidade dos dados após a aplicação do Modelo 2 (k-anonymity + t-closeness):

- **Granularidade:** 26.577089% → indica uma perda significativa de detalhe nos dados, embora não tão severa como os modelos mais restritivos.
- **Non-uniform entropy (Entropia não-uniforme):** 12.89168% → revela elevada homogeneidade, consequência da forte generalização aplicada.
- **Discernibility:** 38.76246% → valor intermédio, indicando penalização moderada pela agregação dos dados em classes.
- **Erro quadrático médio por registo:** 5.26378% → relativamente baixo, o que sugere que, apesar da generalização, os dados permanecem próximos dos valores originais.
- **Tamanho médio das classes:** 97.77023 registos → demonstra que os dados foram agrupados de forma bastante extensa, em conformidade com os requisitos do t-closeness.

Estes resultados evidenciam que o **Modelo 2** oferece uma **proteção reforçada contra inferência estatística**, mas com um impacto considerável na utilidade dos dados. A granularidade e a diversidade informacional foram reduzidas, comprometendo parcialmente a capacidade de análise exploratória e inferencial. Assim, a escolha deste modelo deve ser ponderada em função do **nível de sensibilidade dos dados** e do **grau de anonimato pretendido**, sendo mais adequado para cenários onde a **privacidade absoluta** se sobrepõe à fidelidade analítica.

## 6. Avaliação da Variação de Parâmetros dos Modelos

A etapa final da análise consistiu em avaliar **como a variação de determinados parâmetros** dos modelos de anonimização influencia os resultados em termos de **risco de reidentificação e utilidade dos dados**. Esta análise foi realizada de forma empírica, através da **alteração controlada dos seguintes parâmetros**:



- **Limite de Supressão** (% de registos que podem ser suprimidos para garantir o anonimato)
- **Valor de  $k$**  nos modelos de  $k$ -anonymity

Para cada variação, foram recolhidas métricas de:

- **Utilidade dos dados** (através do score de *Discernibility*)
- **Risco médio de reidentificação** (Average Prosecutor Risk)

Os testes foram realizados no ARX com diferentes configurações, **mantendo os outros parâmetros constantes**, de modo a isolar o efeito individual de cada variável. Os resultados foram convertidos em gráficos para facilitar a interpretação visual.

## 6.1 Limite de Supressão vs Utilidade (Discernibility)

Nesta primeira análise, o objetivo foi perceber como o **aumento do limite de supressão** afeta a **utilidade dos dados**, medida através da métrica *Discernibility Score* — um indicador da granularidade e detalhe informacional presente nos registos após anonimização.

O experimento consistiu em manter o valor de  $k$  fixo ( $k = 25$ ) e aplicar diferentes limites de supressão (5%, 10%, 15%, 20% e 30%), recolhendo os valores de *Discernibility* resultantes para cada caso.

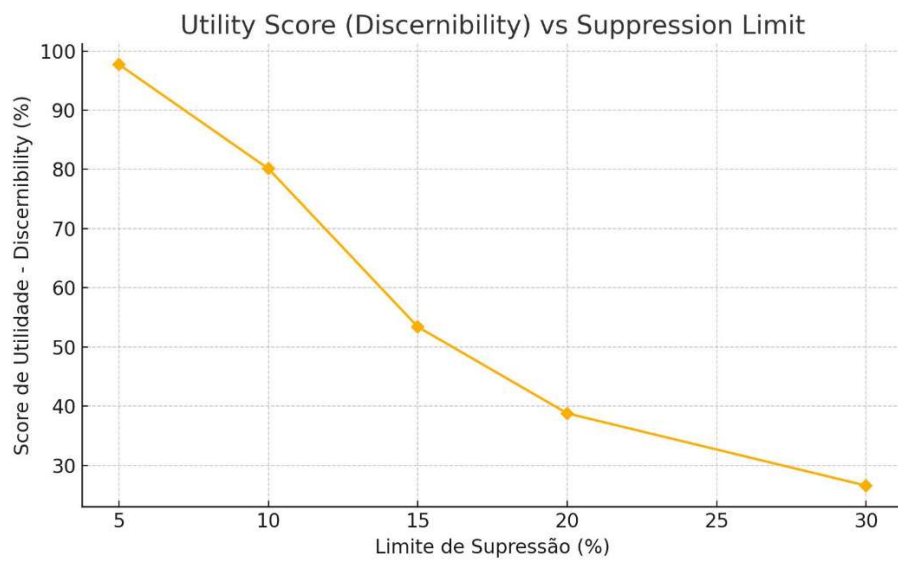


Figura 14 - Variação do *Discernibility Score* em função do limite de supressão

Como se observa na Figura 14, existe uma correlação claramente negativa: à medida que o limite de supressão aumenta, a utilidade (medida por *Discernibility*) diminui drasticamente. O score desce de cerca de 98% com 5% de supressão para 27% com 30% de supressão.

Este comportamento é esperado: quanto mais registos são suprimidos, menos dados permanecem disponíveis para análise, o que reduz a granularidade e a qualidade estatística do

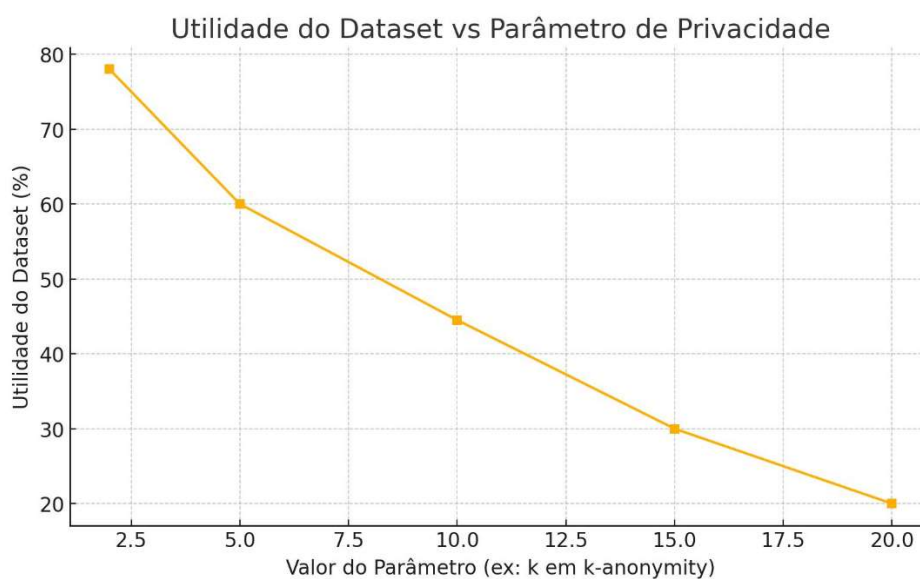


dataset. A supressão, embora útil para proteger a privacidade em casos extremos, deve ser usada com moderação, pois o seu impacto na utilidade é altamente penalizador.

## 6.2 Valor de $k$ vs Utilidade dos Dados

O segundo teste realizado teve como objetivo compreender o **impacto direto do valor de  $k$**  na utilidade do dataset, mantendo os restantes parâmetros constantes.

Neste caso, foram realizadas sucessivas execuções de anonimização com diferentes valores de  $k$  (2, 5, 10, 15 e 20), sendo recolhido, para cada execução, o **Discernibility Score** como métrica de utilidade. Esta métrica penaliza o agrupamento excessivo e a perda de especificidade, sendo **inversamente proporcional ao nível de generalização** aplicado aos dados.



*Figura 15 - Utilidade dos dados (Discernibility Score) em função do valor de  $k$  no modelo  $k$ -anonymity*

A **Figura 15** evidencia uma tendência claramente decrescente: à medida que o valor de  $k$  aumenta, a utilidade dos dados regista uma redução significativa. Os principais pontos observados são:

- Com  $k = 2$ , a utilidade é elevada, próxima dos **78%**, refletindo **grupos de equivalência reduzidos** e elevada especificidade.
- À medida que  $k$  aumenta para 5 e 10, o **Discernibility Score** desce para **60%** e **45%**, respetivamente.
- Com  $k = 20$ , a utilidade atinge um valor mínimo de cerca de **20%**, evidenciando uma generalização intensa e perda considerável de detalhe informativo.

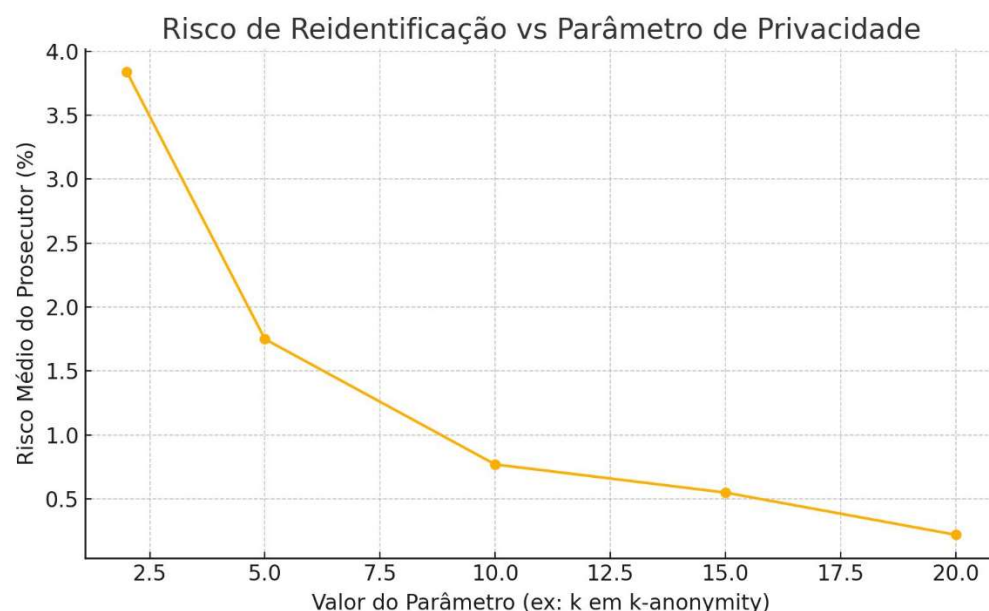
Este comportamento é coerente com o funcionamento do modelo de *k-anonymity*: valores mais elevados de *k* implicam **agrupamentos maiores**, obrigando à generalização de um maior número de atributos para assegurar a indistinguibilidade entre registos. Como consequência, a **granularidade dos dados é gravemente comprometida**, reduzindo a sua utilidade para análises exploratórias ou inferenciais.

No entanto, importa salientar que, para valores de *k* superiores a 15, verifica-se uma **estabilização da perda de utilidade**, o que sugere que o impacto marginal da generalização **tende a diminuir a partir desse ponto** — uma observação relevante na seleção de *k*, sobretudo em cenários que exijam um equilíbrio entre privacidade e valor analítico.

### 6.3 Valor de *k* vs Risco de Reidentificação

Nesta última análise, o objetivo foi avaliar o **efeito do valor de *k* no risco médio de reidentificação**. A métrica utilizada foi o **Average Prosecutor Risk**, que representa a **probabilidade média de um atacante conseguir reidentificar corretamente um registo**, assumindo conhecimento exato de um indivíduo-alvo.

Foram testados os mesmos valores de *k* utilizados anteriormente (2, 5, 10, 15, 20), mantendo constantes os restantes parâmetros de anonimização. Para cada valor de *k*, foi recolhido o risco médio estimado pelo modelo de atacante do tipo *Prosecutor* na ferramenta ARX.



*Figura 16 - Risco médio de reidentificação (Prosecutor Risk) em função do valor de *k* no modelo *k-anonymity**

A **Figura 16** apresenta uma tendência decrescente bastante acentuada, confirmando a eficácia do parâmetro *k* na **mitigação do risco de reidentificação**:

- Com  $k = 2$ , o risco médio atinge **3,85%**, um valor muito elevado e próximo do risco máximo teórico.
- Ao aumentar para  $k = 5$ , o risco desce para **1,75%**, e com  $k = 10$  atinge cerca de **0,75%**.
- A partir de  $k = 15$ , observa-se uma **estabilização**, com reduções marginais até atingir aproximadamente **0,3%** com  $k = 20$ .

Estes resultados validam o princípio subjacente ao modelo de *k-anonymity*: à medida que o valor de  $k$  aumenta, torna-se mais difícil para um atacante distinguir um registo dos restantes, o que **diminui significativamente a probabilidade de reidentificação com sucesso**.

A forma da curva descendente indica ainda que **a maior parte do ganho em proteção ocorre no intervalo entre  $k = 2$  e  $k = 10$** . Para valores superiores, o risco tende a atingir um **ponto de saturação**, no qual os ganhos adicionais em termos de privacidade se tornam marginais, enquanto o **impacto negativo na utilidade dos dados** (como observado na secção 6.2) **se acentua progressivamente** — tornando mais difícil justificar aumentos adicionais de  $k$  sem comprometer o valor analítico dos dados.

## 6.4 Conclusão da Avaliação Paramétrica

A análise empírica da variação dos parâmetros de anonimização demonstra, de forma clara, os compromissos entre **privacidade** e **utilidade dos dados**. O aumento do valor de  $k$  ou do limite de supressão conduz, inevitavelmente, a:

- Uma **redução do risco de reidentificação**, sobretudo nos intervalos mais baixos (por exemplo, de  $k = 2$  até  $k = 10$ );
- Uma **diminuição progressiva da utilidade dos dados**, refletida em métricas como o *Discernibility Score*.

Embora o reforço dos mecanismos de anonimização — através de valores mais elevados de  $k$  ou limites de supressão mais amplos — contribua para **uma maior proteção da privacidade**, os resultados obtidos sugerem que existem **pontos de equilíbrio ideais**. Em particular, valores de  $k$  entre 5 e 10 parecem oferecer uma **segurança razoável** contra reidentificação, **sem comprometer drasticamente a qualidade analítica dos dados**.

Esta avaliação revela-se essencial para uma configuração informada dos modelos de anonimização em contextos reais, permitindo adaptar os **níveis de proteção** às **exigências específicas de privacidade e análise estatística**.

## 7. Comparação Final entre os Modelos

Após a análise detalhada dos dois modelos de anonimização — **Modelo 1: k-anonymity (k = 25) + l-diversity (l = 2)** e **Modelo 2: k-anonymity (k = 25) + t-closeness (t = 0.2)** — foi possível identificar **diferenças substanciais em termos de risco, utilidade e aplicabilidade**.

Ambos os modelos garantiram um nível elevado de proteção da privacidade, com **eliminação total de sample uniques e registos em risco igual a 0%**. No entanto, as estratégias utilizadas impactaram de forma distinta a **utilidade dos dados e o controlo sobre a distribuição do atributo sensível**.

O quadro seguinte resume os principais resultados observados:

Critério	Modelo 1 (k + l)	Modelo 2 (k + t)
Risco médio ( <i>Prosecutor Attacker Model</i> )	0,74818%	0,7799%
Sample & Population Uniques	0%	0%
Registos em risco	0%	0%
Discernibility Score	78,02% ( <i>maior utilidade</i> )	97,70% ( <i>maior penalização</i> )
Granularidade	53,44%	12,89%
Tamanho médio das classes	~98 registos ( <i>menor agregação</i> )	128 registos
Objetivo adicional	Garantia de diversidade ( <i>l-diversity</i> )	Preservação da distribuição estatística ( <i>t-closeness</i> )
Impacto na utilidade	Moderado	Elevado
Aplicabilidade recomendada	Cenários gerais com dados sensíveis	Contextos críticos (saúde, finanças, dados altamente confidenciais)

### Análise Comparativa:

- O **Modelo 1** revelou-se uma **solução equilibrada**, ao conseguir reduzir o risco de reidentificação para níveis residuais, preservando simultaneamente uma **utilidade aceitável** dos dados. A aplicação do critério de *l-diversity* demonstrou eficácia na **prevenção de inferências diretas** sobre o atributo sensível, sem comprometer de forma significativa a estrutura e a granularidade do conjunto de dados.
- O **Modelo 2**, por sua vez, também eliminou os principais riscos de reidentificação, mas **exigiu generalizações substancialmente mais agressivas** para satisfazer a condição de *t-closeness*. Como consequência, a **utilidade informacional foi fortemente penalizada**. No entanto, este modelo oferece uma **camada adicional de proteção contra ataques estatísticos por inferência**, sendo por isso mais indicado para **contextos onde a preservação da confidencialidade do atributo sensível é absolutamente prioritária**.

A escolha entre os modelos deve ser orientada pelo **grau de sensibilidade dos dados**, pela **finalidade da análise** e pelo **nível de proteção** requerido:

- Quando o objetivo principal é **preservar a capacidade analítica dos dados**, mantendo um bom equilíbrio entre utilidade e privacidade, o **Modelo 1** constitui a opção mais adequada.
- Em contrapartida, se a prioridade recair sobre a **proteção máxima do atributo sensível**, evitando qualquer possibilidade de inferência estatística, o **Modelo 2** oferece uma salvaguarda superior — **ainda que com um custo significativo ao nível da utilidade dos dados**.

## 8. Conclusão Final

Este trabalho permitiu não apenas aplicar diferentes modelos de anonimização sobre um conjunto de dados sensível, mas também compreender de forma empírica **as implicações práticas das decisões técnicas** que se tomam neste domínio.

Mais do que atingir níveis formais de *k-anonymity*, *l-diversity* ou *t-closeness*, ficou evidente que **anonimizar é um exercício de equilíbrio**: entre segurança e utilidade, entre proteção legal e viabilidade analítica.

Através de testes sistemáticos e avaliação de métricas objetivas, demonstrou-se que **não existe um modelo “universalmente melhor”**. Cada abordagem oferece vantagens que **devem ser ponderadas em função do contexto**, da sensibilidade dos dados e da finalidade da análise.

Além disso, ao estudar a variação dos parâmetros de anonimização, tornou-se claro que pequenas alterações — como aumentar o valor de *k* ou o limite de supressão — **podem gerar impactos desproporcionados** na qualidade informacional. Essa sensibilidade destaca a importância de **ajustar modelos de forma informada**, em vez de aplicar valores arbitrários ou conservadores por defeito.

O desafio da anonimização não reside apenas na aplicação de técnicas, mas na **tomada de decisões fundamentadas e contextualmente adequadas**. E esse é o verdadeiro valor de uma abordagem consciente.

A realização deste trabalho contribuiu de forma significativa para aprofundar o nosso entendimento prático sobre os desafios da anonimização e a sua importância no contexto da proteção de dados. Compreendemos que decisões técnicas, por vezes vistas como meros parâmetros ou opções de configuração, têm implicações reais na preservação da privacidade e na utilidade dos dados. Esta experiência reforçou a importância de uma abordagem crítica e informada, baseada na experimentação e na análise rigorosa, sempre ajustada ao contexto e aos objetivos da análise.