Trabalho #3 - Anonimização de um dataset com análise de utilidade e risco

Grupo 6 PL4
Carolina Proença - 202306055
Eduarda Neves - 202307178
Maria Morais - 202304201

1. Classificação de Atributos

O dataset disponibilizado contém 9 atributos. Com base nos valores de **distinção** e **separação**, todos os atributos foram classificados como *quasi-identificadores (QIDs)*, à exceção de **"salary-class"**, que foi classificado como **atributo sensível**, por representar a informação confidencial que se pretende proteger. Esta classificação garante que, mesmo com anonimização, a distribuição dos valores de "salary-class" dentro de cada grupo permanece suficientemente próxima ou diversificada, evitando inferências diretas.

A tabela seguinte resume os valores obtidos para os atributos classificados como QIDs:

Atributo	Distinction (%)	Separation (%)
age	0.23	97.81
occupation	0.046	89.46
education	0.053	80.74
marital-status	0.023	65.72
sex	0.0066	43.82
workclass	0.023	43.85
race	0.016	25.10
native-country	0.13	16.00

A métrica **"Distinction"** mede a percentagem de valores únicos de um atributo, logo, os valores baixos indicam muitos registos repetidos. Por outro lado, a **"Separation"** avalia o quanto um atributo separa registos com diferentes valores do atributo sensível, sendo os valores elevados indicativos de maior risco de inferência.

Apesar de apenas quatro atributos (age, occupation, education, marital-status) apresentarem separation superior a 65%, os restantes também foram considerados QIDs devido ao **efeito combinatório**, pois combinações de atributos aparentemente inofensivos podem aumentar significativamente o risco de reidentificação. Exemplos:

- sex + workclass → separation = 69.01%
- race + education → separation = 85.69%
- native-country + marital-status → separation = 71.42%
- sex + race + native-country + workclass → separation = 79.66%

2. Análise dos Riscos de Privacidade (antes da anonimização)

A análise efetuada com o ARX, antes de aplicar qualquer técnica de anonimização, revela **um elevado risco de reidentificação**, como demonstrado pelos resultados dos diferentes modelos de atacantes:

Prosecutor Model (atacante com mais conhecimento):

- **Risco estimado**: 100%
- **Risco mais elevado**: 100%
- **Registos em risco**: 72.86%
- **Registos afetados pelo risco máximo**: 46.49%

Neste cenário, o atacante sabe que o indivíduo está na base de dados e conhece os QIDs. O risco é máximo, com registos identificáveis com certeza absoluta.

Journalist Model (risco intermédio):

- **Risco estimado:** 100%
- Mesmos valores que o modelo anterior, apesar de o atacante ter menos informação (não sabe se o indivíduo está na base e tem apenas informação parcial sobre os QIDs). Mostra que a estrutura dos dados favorece a reidentificação.

Marketer Model (risco mínimo):

- **Risco estimado**: 60.04%
- Mesmo com um atacante aleatório, com base apenas em estatísticas gerais, ainda existe um risco de reidentificação significativo.

Estes resultados evidenciam a necessidade de aplicar técnicas de anonimização ao dataset para garantir conformidade com requisitos de privacidade.

3. Aplicação de Modelos de Privacidade

Com o objetivo de avaliar o desempenho de diferentes técnicas de anonimização, foram aplicados dois modelos de privacidade ao dataset e analisados os seus efeitos tanto no risco de reidentificação como na utilidade dos dados.

Métricas Utilizadas

Para medir **risco de privacidade**, foram utilizadas as seguintes métricas:

- **Average Prosecutor Risk**
- **Estimated Journalist Risk**

A escolha destas duas métricas justifica-se pela sua relevância em contextos distintos de avaliação de risco. A ***Average Prosecutor Risk*** representa o cenário em que o atacante tem conhecimento prévio de que o alvo está presente no dataset e conhece os seus quasi-identificadores, fornecendo uma estimativa direta e rigorosa do risco real de reidentificação. Já a ***Estimated Journalist Risk*** assume um cenário mais conservador, no qual o atacante não sabe se o alvo está na base de dados, sendo útil para avaliar riscos sob perspetivas menos informadas. Entre ambas, dá-se maior ênfase à *Average Prosecutor Risk*, por refletir situações mais críticas e alinhadas com ataques plausíveis em ambientes reais, servindo assim como métrica principal para avaliar a eficácia das técnicas de anonimização aplicadas.

Para avaliar a **utilidade dos dados anonimizados**, recorreram-se às métricas:

- **Discernibility**: Mede a utilidade dos dados com base no tamanho dos grupos de equivalência. No ARX, o valor desta métrica é normalizado/invertido para que seja reportado como uma percentagem de utilidade, logo valores mais altos indicam **maior utilidade**, pois correspondem a menor generalização ou supressão.
- **Normalized-Unit Entropy (NU Entropy)**: Avalia a incerteza média por atributo após a anonimização, com base na entropia da distribuição dos seus valores. Valores mais altos representam **maior preservação da variabilidade original dos dados** e, consequentemente, **maior utilidade**. Valores mais baixos indicam que os atributos foram fortemente generalizados ou suprimidos.

  Estas métricas foram escolhidas por refletirem de forma objetiva o **trade-off entre privacidade e utilidade**, permitindo comparar os efeitos das diferentes configurações de anonimização.

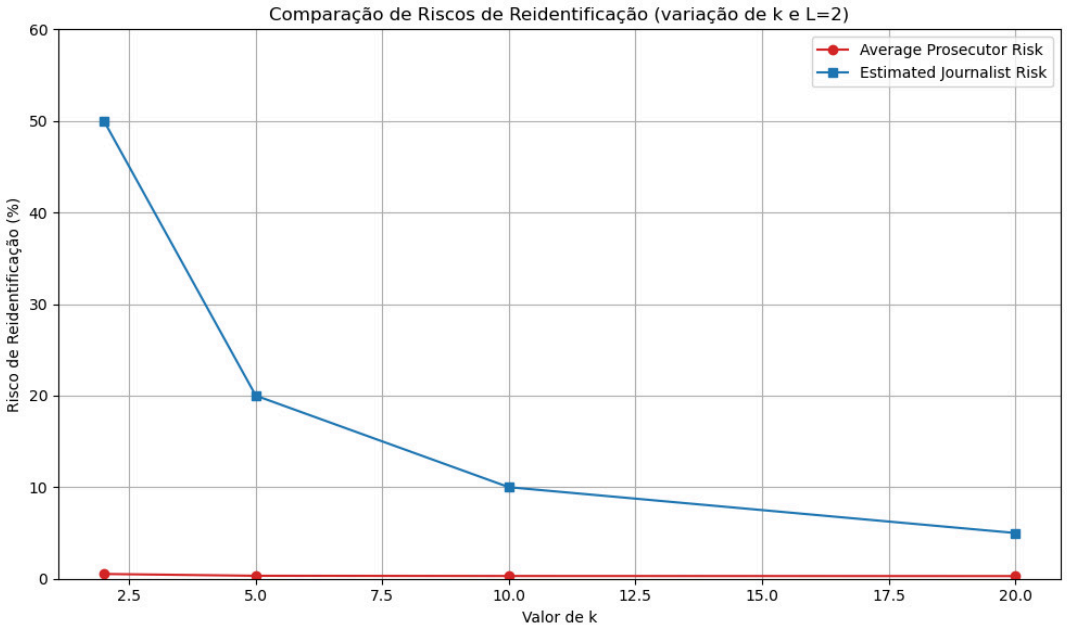
Modelos de Privacidade Aplicados

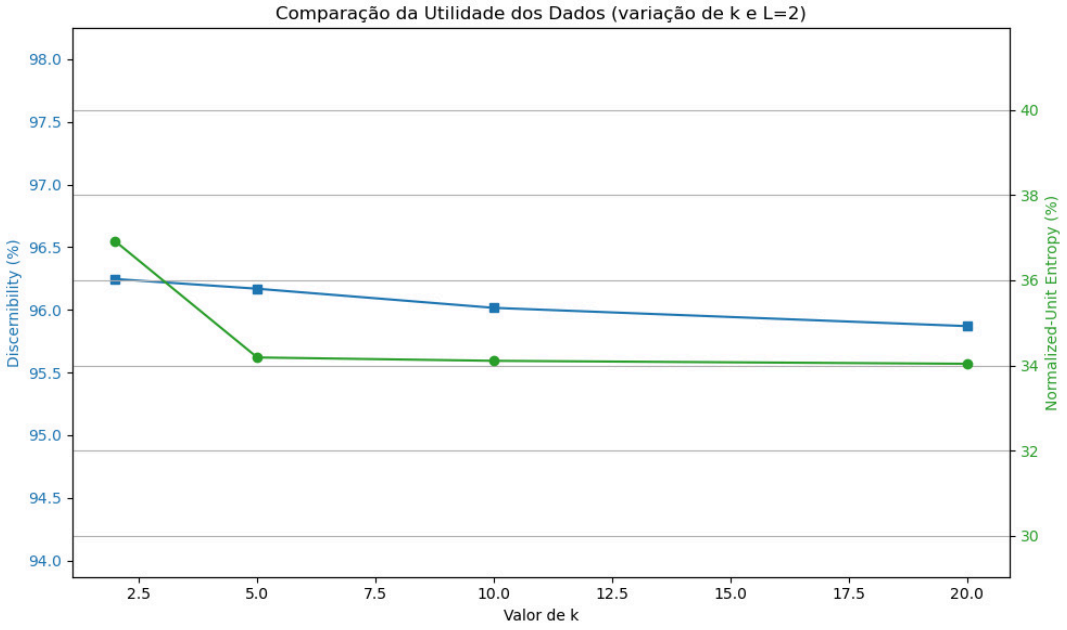
- **k-Anonymity + l-Diversity**
  Este modelo combinado foi utilizado para reforçar a proteção dos dados.

  O **k-anonymity** assegura que cada registo não pode ser distinguido de pelo menos *k−1* outros com base nos QIDs, protegendo contra reidentificação direta.
  Contudo, esta técnica não evita inferências sobre atributos sensíveis. Por isso, foi incluída a **l-diversity**, que impõe que, dentro de cada grupo de k-anónimos, existam pelo menos *l* valores suficientemente distintos do atributo sensível.
  A conjugação destes dois modelos melhora significativamente a proteção contra ataques tanto de reidentificação como de inferência.

Decidimos ver como o modelo combinado reage em 2 situações:

1. Fixar L = 2 e verificar o comportamento a mudar para diferentes valores de k (2, 5, 10, 20)





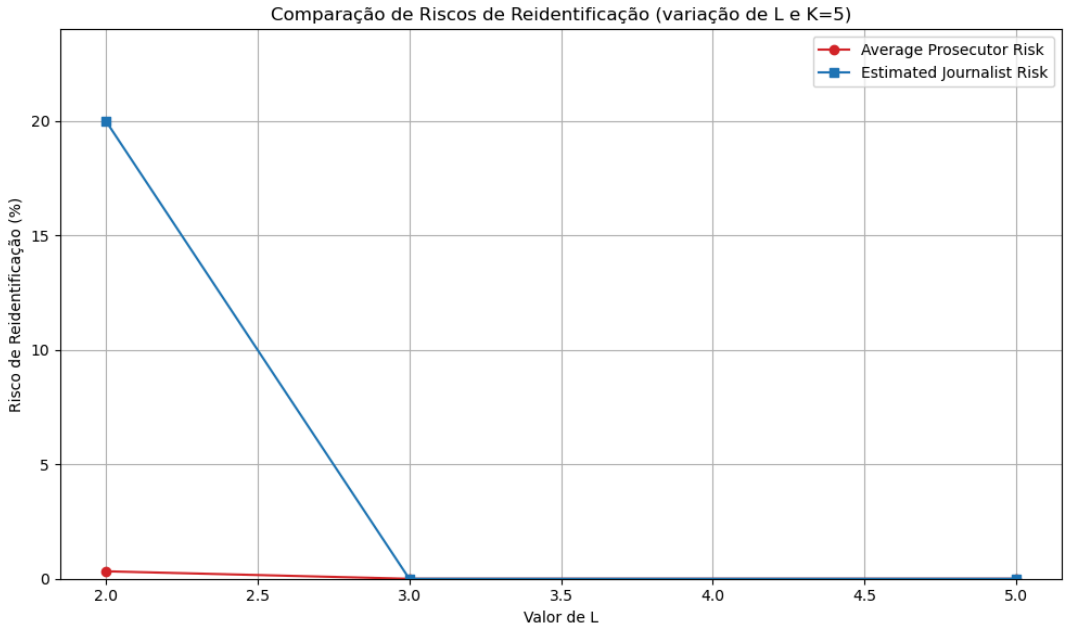
Observa-se que o **risco** estimado pelo modelo de *prosecutor* se mantém quase nulo, enquanto o risco estimado pelo modelo de *jornalista* diminui significativamente com o aumento de *k*, especialmente entre *k = 2* e *k = 5*.

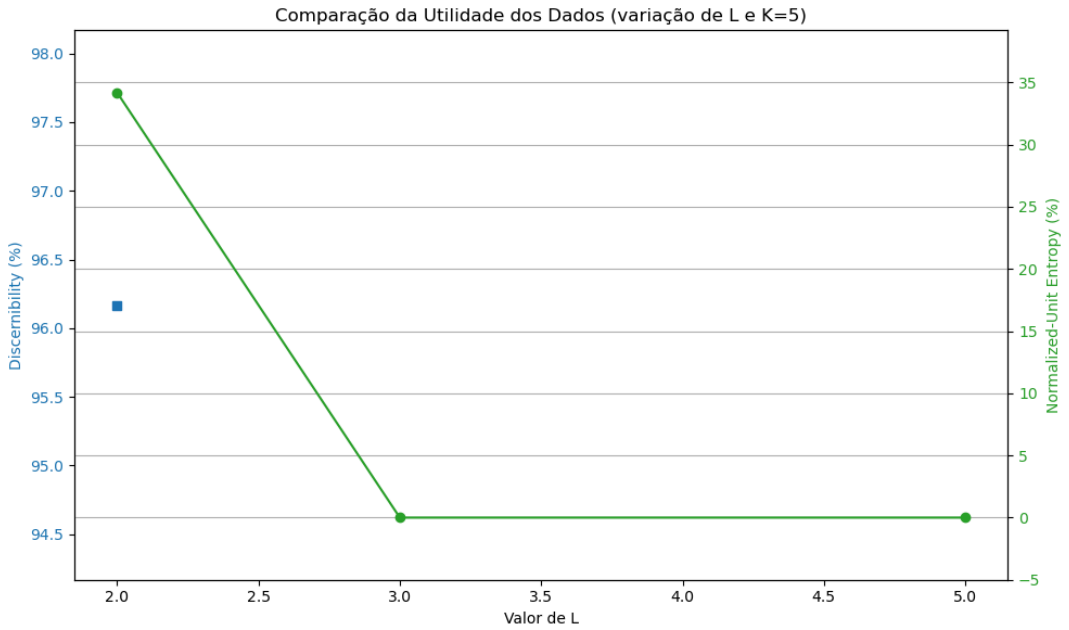
No que toca à **utilidade**, ambas as métricas — *Discernibility* e *Normalized-Unit Entropy* — revelam apenas ligeiras descidas, o que indica que o aumento de *k* tem pouco impacto adicional na utilidade. Os valores elevados de Discernibility (>95%) mostram que a utilidade dos dados se mantém elevada, com pouca generalização ou supressão adicional, e a *NU*

Entropy em torno de **34%** reforça que a generalização adicional afeta pouco a incerteza média dos atributos.

Conclui-se que este modelo permite reduzir substancialmente o risco de reidentificação com o aumento de *k*, mantendo uma utilidade elevada e estável.

2. Fixar K = 5 e verificar o comportamento a mudar para diferentes valores de L (2, 3, 5)





Observa-se que o **risco** de reidentificação, de acordo com as métricas analisadas, se torna nulo a partir de *L* = 3. No entanto, esse aumento de proteção tem um impacto drástico na **utilidade** dos dados: a partir desse ponto, a Discernibility torna-se *NaN* (sendo de 96% em *L* = 2, o que indica elevada utilidade até aí), e a Normalized-Unit Entropy desce para 0%, refletindo utilidade completamente nula.

Este comportamento resulta de uma supressão total dos dados: o ARX aplica generalização até ao nível mais alto das hierarquias (representado por "*"), eliminando qualquer possibilidade de inferência ou reidentificação, mas também destruindo por completo o conteúdo informativo do conjunto de dados.

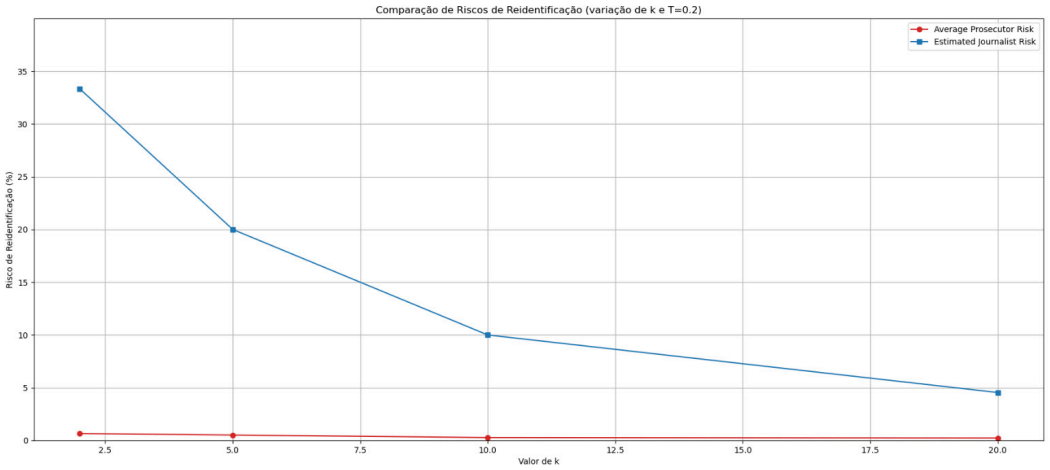
Conclui-se que a exigência de níveis mais altos de diversidade (*L*) pode conduzir a soluções inviáveis para análise de dados, sendo essencial encontrar um equilíbrio entre a proteção da privacidade e a preservação do valor analítico.

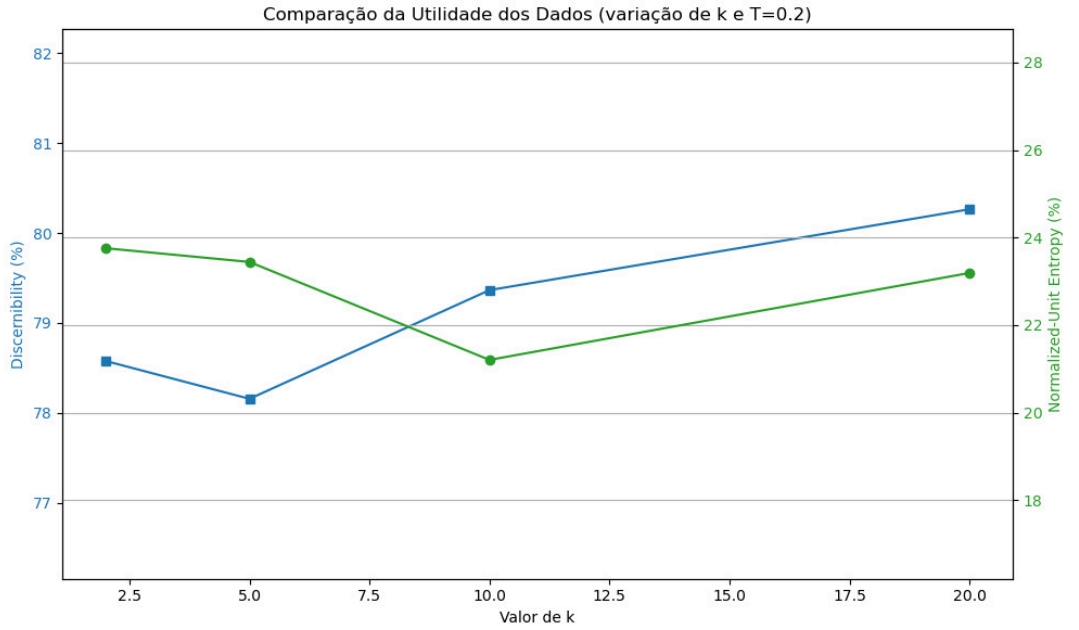
- **K-anonymity + t-closeness**

  Também utilizámos este modelo combinado para reforçar a proteção dos dados.
  Tal como referido anteriormente, o **k-anonymity** impede a reidentificação direta ao garantir que cada registo é indistinguível de pelo menos *k−1* outros com base nos quasi-identificadores.
  No entanto, este modelo não evita inferências quando os valores do atributo sensível são muito homogéneos dentro de um grupo.
  Para ultrapassar essa limitação, aplicámos **t-closeness**, que exige que a distribuição dos valores do atributo sensível em cada grupo seja estatisticamente próxima (a uma distância máxima *t*) da distribuição global.
  Desta forma, a conjugação reduz simultaneamente o risco de reidentificação e o risco de inferência de informação sensível, garantindo uma proteção mais robusta.

De forma análoga, decidimos ver como o modelo combinado reage em 2 situações:

1. Fixar t = 0.2 e verificar o comportamento a mudar para diferentes valores de k (2, 5, 10, 20)



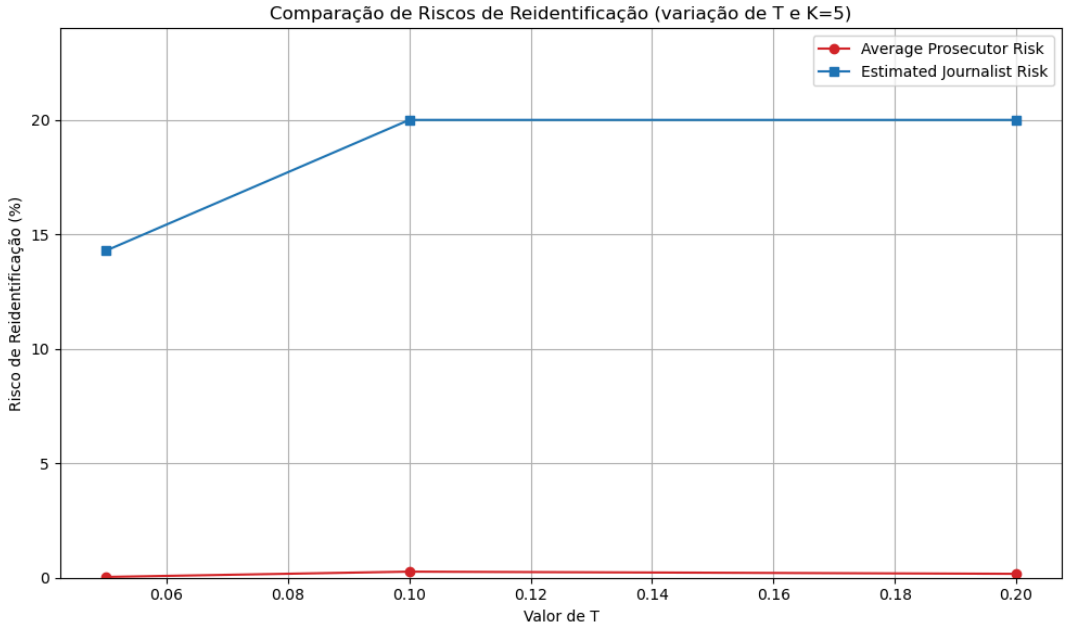


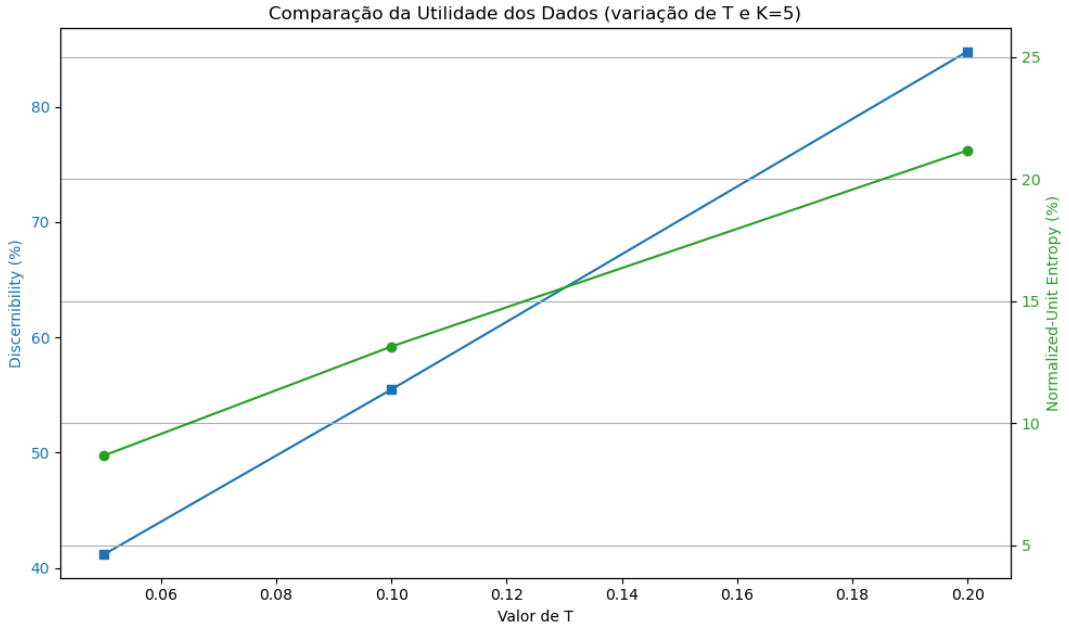
No que respeita ao **risco**, o Average Prosecutor Risk mantém-se praticamente nulo para todos os valores de *k*. Já o Estimated Journalist Risk apresenta uma descida clara com o aumento de *k*, passando de cerca de 30% para 5%. Nota-se ainda que, para *k* = 2, o risco é mais baixo neste modelo (com *t* = 0.2) do que no modelo anterior (com *l* = 2), mantendo-se semelhante para os restantes valores de *k*.

Em termos de **utilidade**, os resultados mostram que o aumento de *k* tem impacto muito reduzido. A Discernibility varia apenas entre 78% e 80%, indicando que a utilidade dos dados se mantém relativamente estável, com níveis constantes de generalização ou supressão. A Normalized-Unit Entropy varia entre 21% e 24%, o que confirma que a incerteza média dos atributos é pouco afetada pela variação de *k*.

Assim, este modelo revela-se eficaz na redução do risco (particularmente no modelo do jornalista), com uma utilidade moderada e estável, mesmo para valores mais elevados de *k*.

2. Fixar K = 5 e verificar o comportamento a mudar para diferentes valores de t (0.05, 0.1, 0.2)





De forma análoga ao conjunto anterior, ao fixar $k = 5$, o Average Prosecutor **Risk** mantém-se praticamente nulo. A principal diferença surge no Estimated Journalist Risk, que apresenta um aumento de aproximadamente 5% ao passar de $t = 0.05$ para $t = 0.1$, mantendo-se estável para $t = 0.2$.

No que diz respeito à **utilidade**, observam-se melhorias claras em ambas as métricas: a Discernibility aumenta de 40% para 85%, refletindo uma preservação significativamente maior da estrutura dos dados, e a Normalized-Unit Entropy cresce de 8% para 21%, indicando um aumento da incerteza média por atributo — e, consequentemente, maior utilidade informativa.

Em suma, o aumento de *t* resulta numa melhoria substancial da utilidade dos dados, com impacto mínimo no risco de reidentificação, reforçando a eficácia do compromisso entre privacidade e qualidade da informação.

Configuração ótima: equilíbrio entre privacidade e utilidade

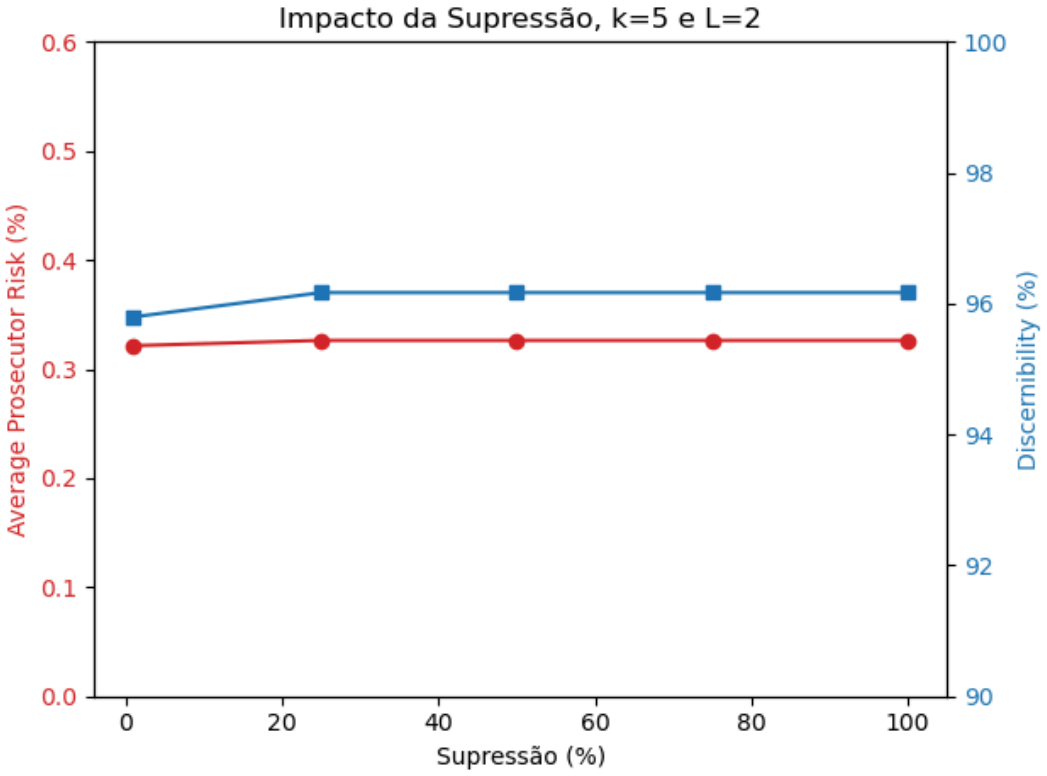
Após uma análise às combinações de modelos que testamos, apesar de todas apresentarem diferenças positivas em relação a antes da aplicação de qualquer modelo, concluímos que a melhor em termos de equilíbrio entre risco e utilidade foi a configuração **K=10, T=0.2**.
O *average prosecutor risk* é de apenas **0.27431%**, com *estimated journalist* e *prosecutor risk* em **10%**, garantindo baixo risco de reidentificação. Ao mesmo tempo, as métricas de utilidade — **discernibility (79.36%)** e **N.-U. entropy (21.20%)** — indicam boa preservação da estrutura e variabilidade dos dados.
Enquanto as configurações com menor risco apresentaram perdas significativas de utilidade, outras com maior utilidade mantinham riscos mais altos. Por isso, essa configuração mostrou-se a mais equilibrada e por esse motivo é a que está i**mplementada no projeto ARX entregue.**

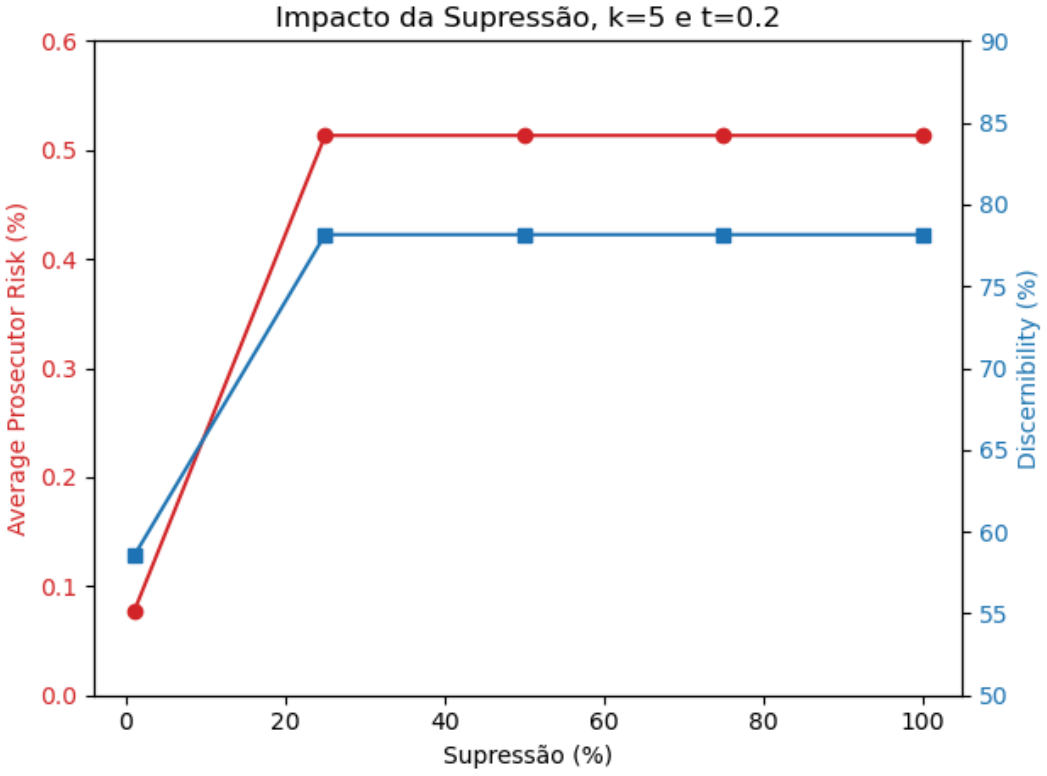
4. Variação dos parâmetros dos modelos de privacidade

Para os dois modelos de privacidade anteriormente definidos - $k = 5 + l = 2$ e $k = 5 + t = 0.2$ - avaliou-se o impacto da variação de certos parâmetros na **utilidade** (medida pela métrica *Discernibility*) e no **risco de reidentificação** (medido pelo *Average Prosecutor Risk*).

1. **Supression limit (para as análises acima, este valor era = 100%):**Este parâmetro define a percentagem máxima de registos que podem ser suprimidos (isto é, ocultados ou removidos) durante o processo de anonimização. A supressão é aplicada quando a generalização, por si só, não é suficiente para satisfazer os requisitos de privacidade. Um *suppression limit* mais baixo impõe uma maior preservação dos dados, mas pode dificultar o cumprimento dos critérios de anonimização. Por outro lado, valores mais elevados facilitam alcançar os níveis de privacidade desejados, à custa de uma maior perda de informação. Este parâmetro,

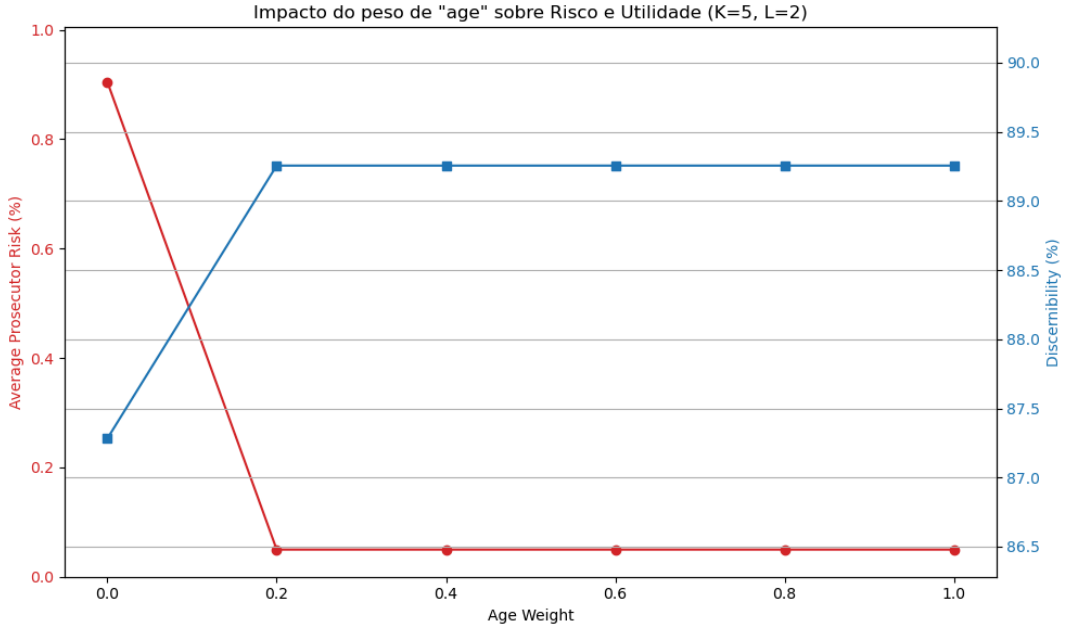
portanto, desempenha um papel crucial no equilíbrio entre a proteção da privacidade e qualidade dos dados anonimizados.

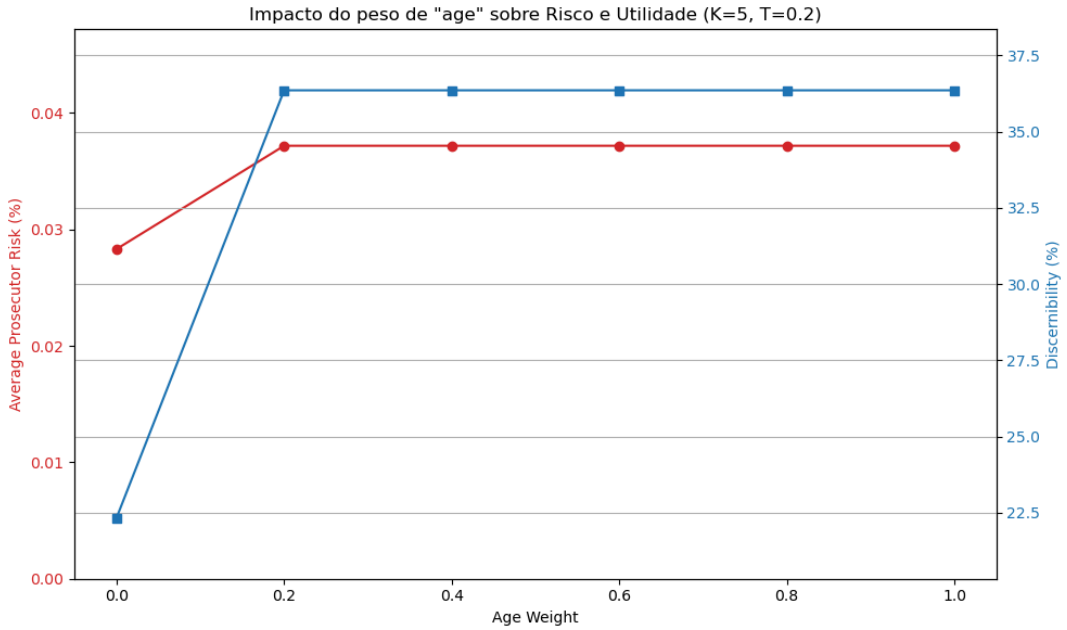

Impacto da Supressão, k=5 e L=2
Average Prosecutor Risk (%)
Discernibility (%)
Supressão (%)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
90
92
94
96
98
100
0
20
40
60
80
100


Impacto da Supressão, k=5 e t=0.2
Average Prosecutor Risk (%)
Discernibility (%)
Supressão (%)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
50
55
60
65
70
75
80
85
90
0
20
40
60
80
100

Observa-se que, no primeiro caso com ***k=5* e *L=2***, tanto o risco quanto a utilidade permanecem praticamente constantes, indicando que a supressão tem efeito limitado quando combinada com uma proteção mais robusta como a *L-diversity*.
Já no segundo cenário, com ***k=5* e *t=0.2***, a supressão provoca um aumento expressivo no risco, mas principalmente na utilidade, sugerindo que, sob um modelo de privacidade menos rigoroso (baseado em *t-closeness*), a supressão pode comprometer a privacidade enquanto melhora a utilidade dos dados. Esses resultados reforçam a importância da escolha adequada do modelo de privacidade em conjunto com técnicas de supressão.

2. **Variar peso do atributo "age"** (com todos os restantes pesos = 0): Escolhemos este pois inicialmente concluímos que era o atributo com maior separação. Variações no peso de um atributo influenciam diretamente a forma como o ARX prioriza a generalização ou supressão desse atributo durante a anonimização. Atribuir um peso maior a um atributo indica que ele é mais importante para a utilidade dos dados, levando o modelo a preservá-lo melhor, mesmo que isso implique generalizar mais outros atributos.





Observa-se que, ao variar o peso do atributo "idade", os valores de risco e de utilidade se mantêm praticamente constantes a partir dos 20% de peso. As principais alterações ocorrem apenas no intervalo entre 0% e 20%, indicando que, acima deste limiar, o modelo estabiliza o compromisso entre privacidade e utilidade.
No cenário com ***k=5*** **e** ***L=2***, as métricas mantêm-se relativamente estáveis independentemente da percentagem de peso, sugerindo que este modelo já proporciona uma proteção eficaz sem depender fortemente disso.
Por outro lado, no cenário com ***k=5*** **e** ***t=0.2***, tanto o risco como a utilidade são mais baixos quando comparados com os valores obtidos com 100% de supressão e todos os atributos com peso 0.5, situando-se agora a descernibility em torno dos 35%. Destaca-se o salto observado entre 0% e 20%, que é especialmente relevante para a discernibilidade, com um ganho superior a 20%, o que demonstra o impacto direto que a preservação deste atributo pode ter na utilidade dos dados anonimizados.