Anonimização de um *Dataset*

Trabalho de Segurança e Privacidade

Realizado pelos alunos:

Maximiliano Vítor Phillips e Sá (up202305979),

Rita Maria Pinho Moreira (up202303885),

e Samuel José Sousa Ventura da Silva (up202305647)

0. Índice

1. Introdução

A proteção da privacidade de dados pessoais tornou-se um pilar fundamental na era digital, especialmente em contextos onde a partilha e análise de grandes volumes de informação são essenciais para avanços científicos e tecnológicos. Neste contexto, a anonimização de dados emerge como uma técnica fundamental para viabilizar a partilha e análise de informações sensíveis sem comprometer a identidade dos titulares dos dados.

Este relatório descreve o processo de anonimização de um *dataset* real, utilizando a ferramenta ARX (*Anonymization and Risk eXploration*), reconhecida pela sua robustez na aplicação de técnicas de anonimização e avaliação de riscos. Inicialmente, os atributos do *dataset* foram classificados nas categorias de *Identifying*, *Quasi-Identifying* (QID), *Sensitive* e *Insensitive*, tendo em conta as melhores práticas de classificação e as métricas de distinção e separação recomendadas. O *dataset* inclui 9 variáveis sobre 30162 registos: *sex* (sexo), age (idade), *marital-status* (estado civil), *education* (educação), *native-country* (país de origem), *workclass* (setor de trabalho), *occupation* (emprego), *race* (raça) e *salary-class* (faixa salarial).

Posteriormente, foi realizada uma análise dos riscos de reidentificação associados ao *dataset* original e após a aplicação de diferentes modelos de privacidade, nomeadamente *k-anonymity*, *l-diversity* e *t-closeness*. Para cada modelo, avalia-se o impacto na redução do risco e na utilidade dos dados, permitindo comparar as vantagens e limitações de cada abordagem.

Desta forma, conseguimos ter uma análise mais crítica sobre o equilíbrio entre privacidade e utilidade dos dados, contribuindo para a adoção de melhores práticas em contextos reais de tratamento de informação sensível.

2. Classificação de atributos

Os seguintes atributos são categorizados como *Identifying*, *Quasi-Identifier (QID)*, *Sensitive* ou *Insensitive* para orientar a escolha das técnicas de anonimização.

- *Identifying*: Permitem a ligação direta a um indivíduo específico e devem ser removidos ou substituídos. (ex. nome, número de identificação)
- *Quasi-Identifier (QID)*: Não identificam alguém isoladamente, mas quando combinados entre si ou com fontes externas, podem reidentificar indivíduos.
- *Sensitive*: Contêm informação intrinsecamente delicada cujo acesso não autorizado pode resultar em danos ou discriminação.
- *Insensitive*: Baixo risco para a privacidade e não contribuem significativamente para a reidentificação nem contêm informação sensível.

Segue abaixo uma tabela com cada classificação de atributo e a sua justificação.

Nome do atributo	Classificação	Justificação
Sex	QID	Distinção: 0.0066% Separação: 43.8284% Baixo poder de distinção direta, mas em conjunto com outros atributos aumenta a unicidade dos registos. Necessita generalização para k-anonimato.
Age	QID	Distinção: 0.2387% Separação: 97.8117% Alta cardinalidade (muitas faixas etárias) e maior poder de distinção, mas só facilita reidentificação com outros QIDs. Necessita generalização para k-anonimato.
Race	*Sensitive*	Distinção: 0.0166% Separação: 25.1017% Intrinsicamente sensível e não permite identificacão direta. Exige *l-diversity* ou *t-closeness* para evitar inferências discriminatórias.
Marital-status	QID	Distinção: 0.0232% Separação: 65.7201% Contribui para distinção de perfis e não é sensível em si, mas pode reidentificar em conjunto com outros QIDs. Necessita generalização para k-anonimato.
Education	QID	Distinção: 0.0531% Separação: 80.7438% Moderada cardinalidade (níveis de escolaridade) e reforça a distinção quando combinado com outros QIDs. Necessita generalização para k-anonimato.
Native-country	QID	Distinção: 0.1359% Separação: 16.7895% Alta cardinalidade (vários países) e maior poder de distinção, mas vulnerável a reidentificação quando combinado com outros QIDs ou bases externas (valor de separação aumenta muito).
Workclass	QID	Distinção: 0.0232% Separação: 43.8471%

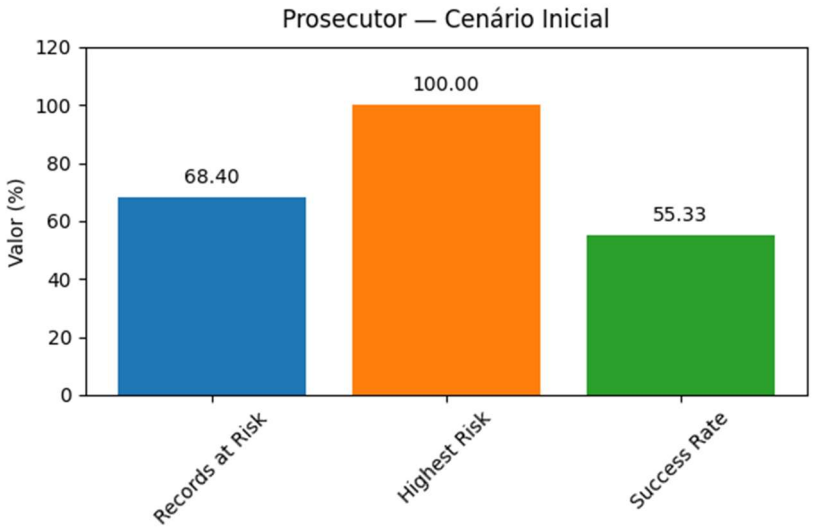
		Várias classes laborais, mas não é sensível e contribui para reidentificação de registos em conjunto. Necessita generalização para k-anonimato.
Occupation	QID	Distinção: 0.0464% Separação: 89.4622% Alta cardinalidade (várias profissões) e reforça a distinção quando combinado com outros QIDs. Necessita generalização para k-anonimato.
Salary-class	*Sensitive*	Distinção: 0.0066% Separação: 37.3933% Revela informação delicada (faixa de rendimento) e é sensível por poder causar discriminação. Exige *l-diversity* ou *t-closeness* para evitar inferências discriminatórias.

Nota: os valores de distinção e separação são os valores de cada atributo quando não está combinado com outro atributo.

3. Riscos de privacidade do *dataset* original

Este capítulo apresenta uma análise dos riscos de privacidade no *dataset* original, considerando diferentes perfis de atacantes (Modelos de Promotor, Jornalista e Marketing), e os seus respetivos limiares de risco.

3.1. Modelo do Promotor



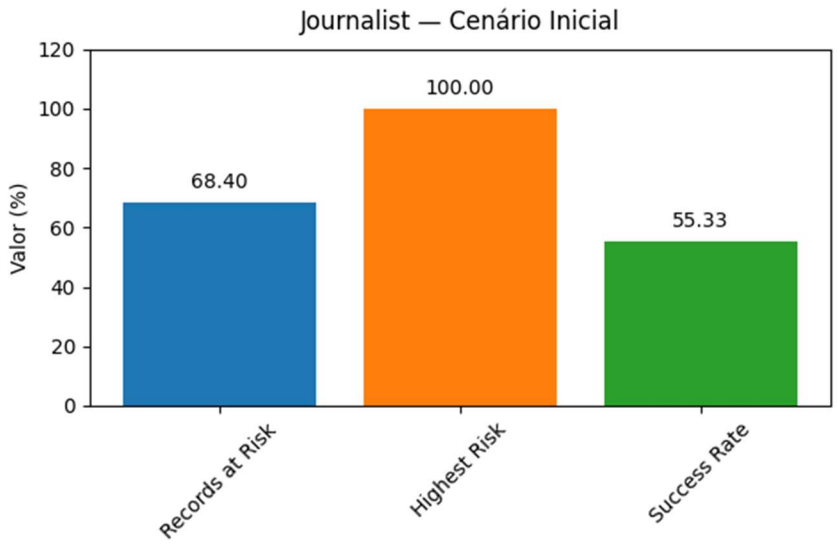
No modelo de promotor-atacante, o invasor tem como alvo um indivíduo específico e presume-se que ele já saiba que dados sobre ele estão contidos no conjunto de dados.

Um promotor público, com acesso a recursos legais e informações complementares, tem alta probabilidade de ligar registos do *dataset* a indivíduos em mais de metade dos casos.

De acordo com a figura 1, aproximadamente 68.4% dos registos do *dataset* original podem ser reidentificados por um promotor público, mas 100% dos mesmos atingem o risco máximo de reidentificação sob esse modelo. Assim, a taxa de sucesso na reidentificação de registos é de 55.33%.

3.2. Modelo Jornalístico

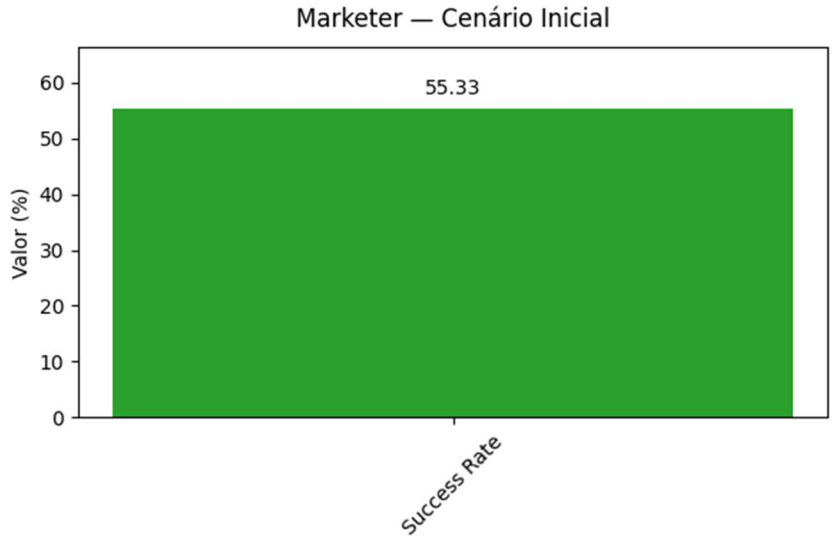
Neste modelo, o agressor tem como alvo um indivíduo específico, mas não se espera que ele tenha conhecimento prévio sobre a filiação.



Um jornalista da área de investigação, com técnicas de correlação de dados públicos, alcança níveis de risco semelhantes aos do promotor, destacando a vulnerabilidade das informações pessoais.

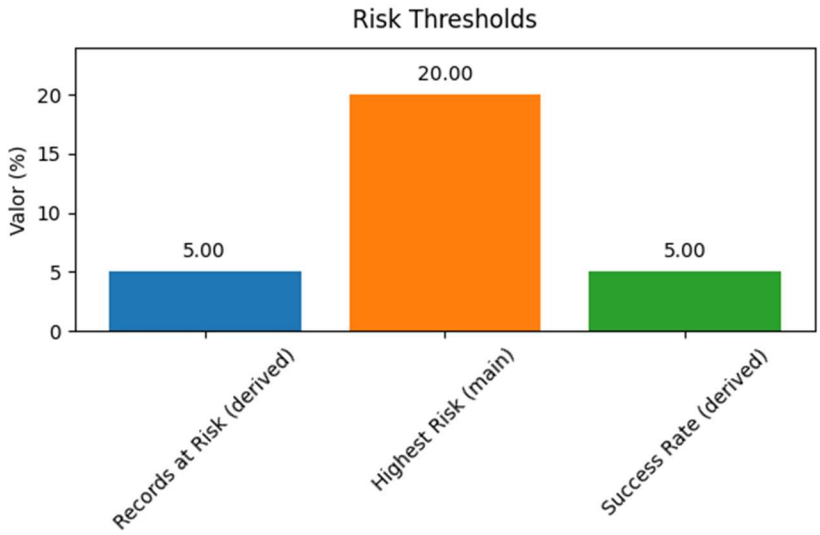
Como está apresentado na figura 2, cerca de 68.4% dos registos podem ser reidentificados por um jornalista, mas 100% dos registos atingem o risco máximo de reidentificação sob esse modelo. Assim, tal como no Modelo do Promotor, a taxa de sucesso na reidentificação de registos é de 55.33%.

3.3. Modelo de Marketing



No modelo de marketing, o invasor não tem como alvo um indivíduo específico, mas sim a reidentificação de um grande número de indivíduos. Portanto, um ataque só pode ser considerado bem-sucedido se uma fração grande dos registos puder ser reidentificada. Assim, a sua taxa de sucesso médio na reidentificação de registos é de 55.33%, visto que o invasor possui menos informação de contexto.

3.4. Limiares de Risco



O limiar definido como "alto risco" do risco principal é ultrapassado por 20% dos registos. Já no limiar de risco derivado, 5% dos registos caem no mesmo, tendo assim uma taxa de sucesso de 5%.

3.5. Panorama Geral

Medida	Valor [%]
Menor risco (Promotor)	2.17391%
Registos afetados pelo menor risco	0.15251%
Risco Médio (Promotor)	55.3279%
Maior risco (Promotor)	100%
Registos afetados pelo maior risco	41.03176%
Risco estimado (Promotor)	100%
Risco estimado (Jornalista)	100%
Risco estimado (*Marketing*)	55.3279%

Unicidade na amostra	41.03176%
Unicidade na população	1.78134%
Modelo populacional	PITMAN
QIDs	Age, education, marital-status, native-country, occupation, sex, workclass

Os resultados indicam elevado risco de reidentificação sob os modelos de promotor e jornalista, ambos alcançando 100% de risco estimado. O modelo de marketing apresenta risco moderado. A unicidade na população é baixa (1.78%), mas na amostra atinge 41.03%, reforçando a necessidade de técnicas de anonimização. O modelo populacional usado, *Pitman*, é, no contexto de anonimização de dados e de ferramentas como o ARX, um dos métodos estatísticos usados para estimar a unicidade de um registo na população total, a partir de informação observada numa amostra, ou *dataset*.

4. Modelos de Privacidade

4.1. Conceitos Fundamentais

Um modelo de privacidade é um conjunto de princípios, regras ou técnicas, usado para proteger dados pessoais contra reidentificação ou exposição indevida durante a partilha ou análise desses dados. No contexto deste projeto, um modelo de privacidade define matematicamente o significado de "proteger a privacidade", estabelecendo assim critérios que uma base de dados deve satisfazer para ser considerada "privada o suficiente". Neste projeto vamos usar três modelos de privacidade:

- ***k-Anonymity***: técnica de anonimização de dados usada para proteger a privacidade de indivíduos numa base de dados. Envolve generalização de dados, *data masking*, ou a substituição de PII por pseudónimos para assegurar que nenhum indivíduo pode ser identificado. É vulnerável a ataques homogéneos e de *background knowledge*;

- ***l-Diversity***: É uma extensão do modelo *k-Anonymity*, desenvolvida para reduzir a granulação da representação de dados numa base de dados. O modelo *l-Diversity* assegura que nenhuma informação de um indivíduo pode ser identificada pelo menos por L outros indivíduos da base de dados, baseado num atributo sensível, protegendo assim ambos atributos sensíveis como gerais. É vulnerável a ataques *skewness* e de similaridade;

- ***t-Closeness***: É um refinamento da extensão *l-Diversity*, pois trata os valores de um atributo de forma distinta, tendo em conta a distribuição dos mesmos. Diz-se que uma

classe equivalente possui *t-Closeness* se a distância entre a distribuição de um atributo sensível nessa classe e a distribuição do atributo na tabela completa não excede *t* (*threshold*).

Para cada modelo aplicado, serão analisados os riscos e a sua utilidade.

4.1.1. Análise de Risco

Na análise de risco, tal como na análise inicial, serão avaliados três modelos de atacante: Promotor, Jornalista e Marketing, e dentro dos mesmos são fornecidos valores para *Records at Risk, Highest Risk* e *Success Rate*, exceto no modelo de Marketing, que só possui *Success Rate*.

A métrica *Records at Risk*, também intitulado de "*proportion of records at risk*", indica a fração (ou número) de registos na base de dados cuja probabilidade de reidentificação excede um certo limiar, que segundo a análise inicial é de 5%). É calculada estimando, para cada registo i, a sua probabilidade de ser reidentificado (p_i). Definindo o limiar τ (20% ou 0.20), conta-se quantas probabilidades p_i são iguais ou superiores ao limiar, dividindo o valor resultante pelo número total de registos N:

$$Records\ at\ Risk = \frac{|\{ i : p_i \geq \tau \}|}{N}.$$

Após calculado o valor, este é filtrado usando o limiar de alerta de *Records at Risk* (5%), derivado do limiar principal *Highest Risk* (20%) e não parte do cálculo intrínseco da métrica. Assim, quanto maior for esse valor, mais registos estão vulneráveis a um ataque, indicando uma pior proteção de privacidade.

A métrica *Highest Risk*, também chamada de "*maximum re-identification risk*", captura o pior caso de risco dentro do *dataset*, isto é, num conjunto de probabilidades de reidentificação excedentes do limiar de *Records at Risk*, este será o seu valor máximo. Mostra, assim, o registo mais vulnerável ("ponto fraco") mesmo que a maioria esteja bem protegida.

Por último, o *Success Rate* de um modelo de atacante mede a efetividade prática do adversário simulado em reidentificar registos ou atributos. Usa um limiar derivado do *Highest Risk* (20%), de valor 5%, ou seja, a partir daí é considerado um modelo inseguro. Possui duas variações:

- *Success Rate* de Reidentificação Global: Proporção de tentativas de reidentificação feitas pelo atacante que acertam o verdadeiro ID ou valor sensível,

$$Success\ Rate = \frac{Nº\ de\ acertos}{Nº\ de\ tentativas}$$

- *Success Rate em Top-k*: Se o atacante produz um *ranking* de ***k*** candidatos possíveis para cada registo, é a proporção de vezes que o verdadeiro alvo está dentro dos ***k*** primeiros resultados. Traduz, portanto, o risco teórico num desempenho simulado de ataque, sendo assim o indicador mais prático de vulnerabilidade.

Juntas, estas métricas permitem avaliar tanto o risco médio quanto o risco extremo e a eficiência real do atacante sob diferentes modelos de privacidade.

4.1.2. Análise de Utilidade

A análise de utilidade em contextos de anonimização de dados avalia o quanto a informação original é preservada após a aplicação de técnicas de privacidade. O objetivo é garantir que os dados continuem úteis para análises estatísticas, *Machine Learning,* entre outras aplicações, mesmo após a proteção da privacidade.

Existem dois tipos de métricas de utilidade: Qualidade ao nível de Atributo (*Attribute-Level Quality*) e Qualidade ao Nível de Conjunto de Dados (*Dataset-Level Quality*).

A qualidade ao nível de atributo avalia o impacto da anonimização em atributos específicos. Para isso, são usadas métricas como:

- *Missings*: número de valores ausentes ou suprimidos após a anonimização;

- Intensidade de Generalização (*Gen. Intensity*): Nível de generalização aplicado aos dados, isto é, mede o quão "largo" foi o recodificar dos atributos. Quanto maior for a intensidade, maior será a perda de detalhe,

$$GenIntensity = \frac{1}{|QI|}\sum_{QI}\frac{nível\ de\ recodificação}{máximo\ nível\ possível}\ ;$$

- Granularidade: Mede o nível de detalhe mantido nos dados após a generalização,

$$Granularity = 1 - \frac{\sum_i E_i}{N\ \times\ |QI|}\ ;$$

- Entropia Normalizada (*N. -U. entropy*): Avalia a diversidade de valores num atributo. Uma menor entropia indica perda de informação,

$$Entropia = \frac{Entropia\ anonimizada}{Entropia\ original}\ ;$$

- Erro Quadrático ao Nível de Atributo (*Attribute-Level Squared Error*): Calcula o erro médio quadrático entre os valores originais e anonimizados de um atributo numérico.

A qualidade ao Nível de Conjunto de Dados avalia o impacto da anonimização no conjunto como um todo, usando métricas como:

- Intensidade de Generalização;

- Granularidade;

- Entropia Normalizada;

- Discernibilidade: Mede a capacidade de distinguir registos após a anonimização, ou seja, mede a "penalidade" infligida às classes de equivalência. Valores mais altos indicam uma menor utilidade,

$$Discernibility = \sum_i |E_i|^2 \ ;$$

- Tamanho Médio de Classes (De Equivalência): Média de número de registos em cada grupo indistinguível, usando-se a proporção dessa média em relação ao tamanho ideal (k). Tamanhos maiores indicam maior perda de especificidade,

$$Avg\ Equivalence\ Class\ Size = \frac{\frac{1}{N}\sum_i |E_i|}{k} \ ;$$

- Erro Quadrático;

- Erro Quadrático ao Nível do Registo (*Record-Level Squared Error*): Calcula o erro quadrático entre registos originais e anonimizados individualmente;

- Erro Quadrático Específico de Agregação (*Aggregation Specific Squared Error*): Avalia o erro introduzido por técnicas de agregação específicas aplicadas durante a anonimização.

Na análise de utilidade de dados anonimizados, é importante considerar tanto as métricas ao nível de atributo quanto ao nível do conjunto de dados.

4.2. k-Anonymity com l-Diversity

A primeira combinação de modelos escolhida foi *k-Anonymity* com *l-Diversity*, para lidar com atributos sensíveis (*race, salary_class*).

4.2.1. Variação de *k* e *l*

Foram escolhidos diferentes parâmetros para *k-Anonymity* e *l-Diversity*, de forma a analisar a mudança de valores percentuais de risco e utilidade:

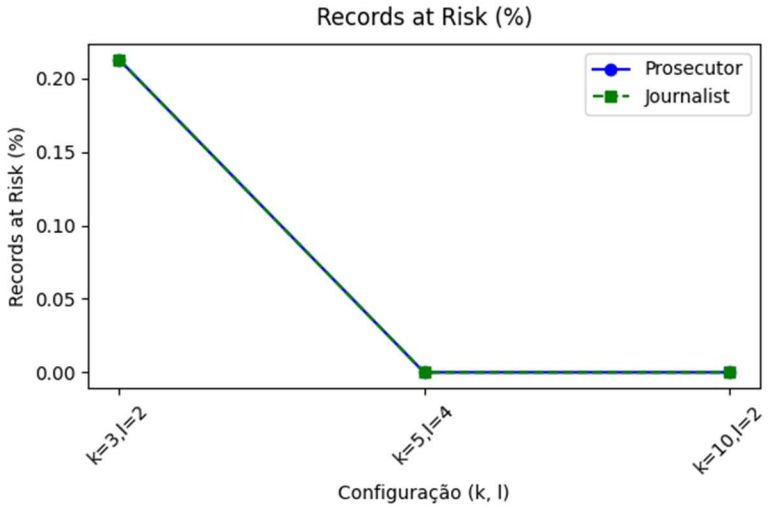
- k = 3 e l = 2;

- k = 5 e l = 4;

- k = 10 e l = 2.

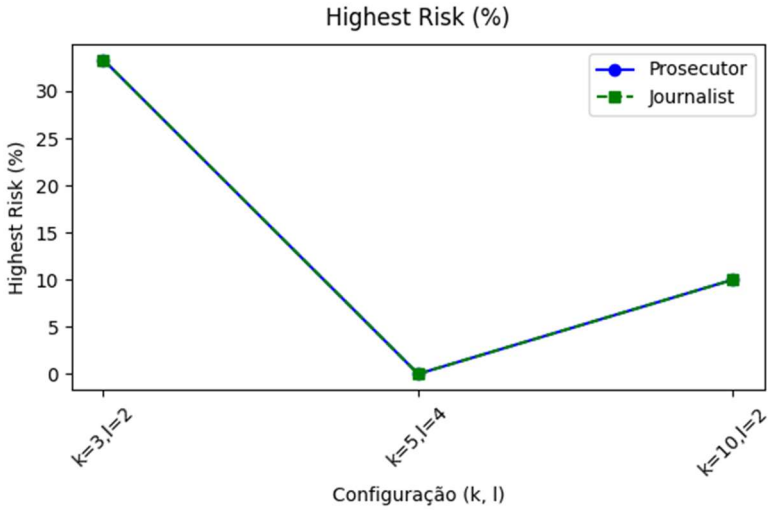
Em todas as anonimizações feitas, foram utilizadas medida de utilidade *Loss*, limite de supressão de 100%, e pesos de atributos padrão, de 0.5.

Análise de Riscos



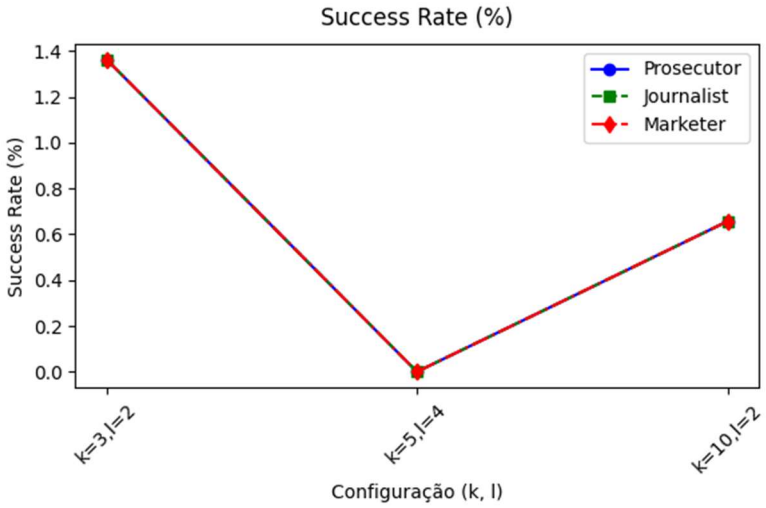
É possível analisar na figura acima que, quando k = 3 e l = 2, 20% de registos excede o limiar de 5% em relação à probabilidade de reidentificação, tornando-se um risco inaceitável, tanto num modelo de Promotor (*Prosecutor Attacker Model*), como no modelo de Jornalista (*Journalist Attacker Model*). Quanto maior é k, menor é este valor, indicando assim menor risco de identificação.

É possível observar isso quando k = 5 (e l = 4), e k = 10 (e l = 2), pois nenhum registo excede o limiar de 20% em nenhum dos modelos. Comparativamente à análise inicial, de valor 68.4%, observa-se um claro decréscimo nesta probabilidade.



Na figura seguinte, é possível observar a percentagem de risco máximo para ambos os modelos de atacante. É percetível um pico quando k = 3 e l = 2, de mais de 30%, ultrapassando o limiar, mas, ao contrário da análise anterior, o aumento de k não é a solução para um valor eficaz, mas talvez seja o aumento de l, do modelo *l-Diversity*.

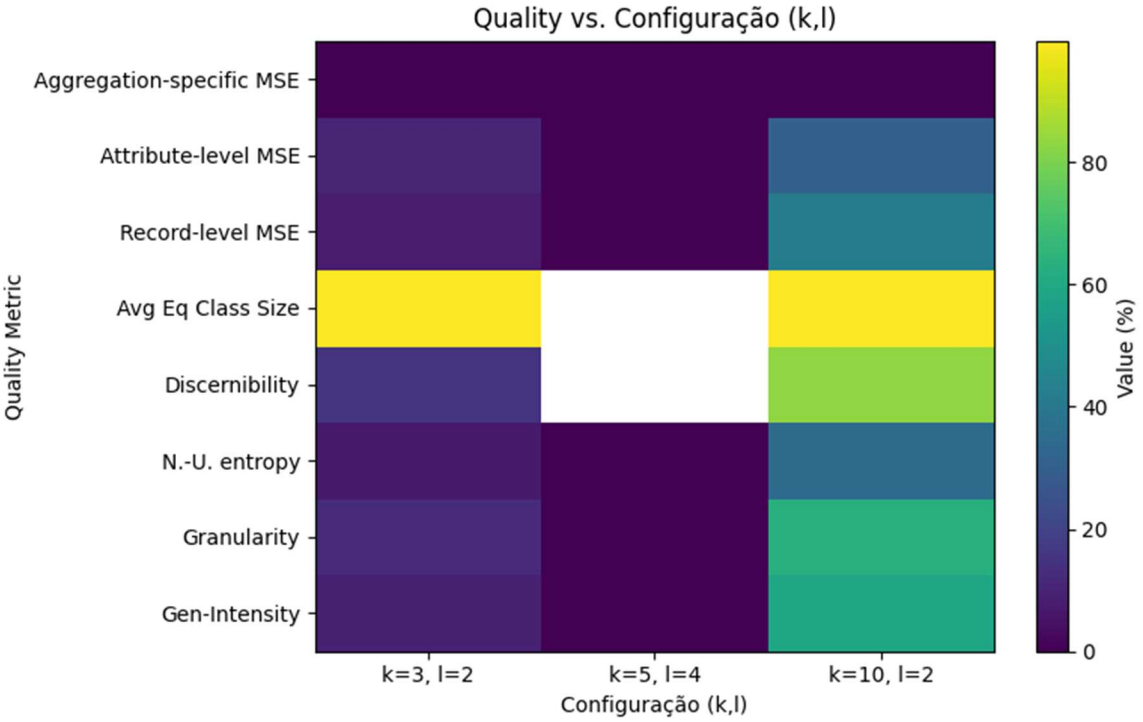
Assim, o ponto ótimo deste risco máximo ocorre quando k = 5 e l = 4, de valor 0%. Em relação à análise inicial, com um risco máximo de 100%, verifica-se um anulamento do mesmo, pois a probabilidade de reidentificação máxima encontrada com esses parâmetros é nula. Já com k = 10 e l = 2, a probabilidade máxima é de 10%, que não supera o limiar, mas ainda é reportado.



Por último, o *Success Rate* é significativamente baixo em todos os modelos, mas há um que se destaca: k = 5 e l = 4, com 0%, o que significa que nenhum modelo tem possibilidade de atacar com sucesso a base de dados. Quando k = 3 e l = 2, a taxa de sucesso é de quase 1.4%, isto é, 0.014 é a proporção de tentativas em que o atacante acerta no valor sensível ou ID. Já com k = 10 e l = 2, a taxa de sucesso é de

aproximadamente 0.7%, demonstrando que, apesar de baixo, ainda existem padrões que permitem a reidentificação em cerca de 1 em cada 140 tentativas. Em todos os cenários, nenhum destes valores é considerado perigoso, visto que o limiar de *Success Rate* é de 5%, e nenhum excede este valor, embora o par k = 5 e l = 4 seja o mais seguro.

Análise de Utilidade (Qualidade do *dataset*)



No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha de parâmetros (k,l). No eixo Y estão as métricas usadas na avaliação da qualidade do conjunto de dados, e no eixo X estão as configurações escolhidas para a anonimização dos dados.

Analisando o mesmo de baixo para cima, é possível tirar diversas informações acerca da utilidade do conjunto de dados.

Na primeira configuração, (k = 3, l = 2), a intensidade de generalização é baixa (9%), isto é, pouca generalização foi aplicada e os dados do conjunto foram pouco transformados via taxonomias. Curiosamente, a granularidade também é baixa (12%), o que pode dever-se a supressão intensa, ou agrupamentos prévios, logo quase nenhum detalhe ou precisão foi mantido. De seguida, a entropia normalizada e a discernibilidade foram igualmente baixas (7% e 15%), o que pode ser interpretado como a existência de valores muito repetidos ou apagados, que se distinguem entre si, isto é, o conteúdo tornou-se pouco diverso e distintivo. A média de tamanhos de

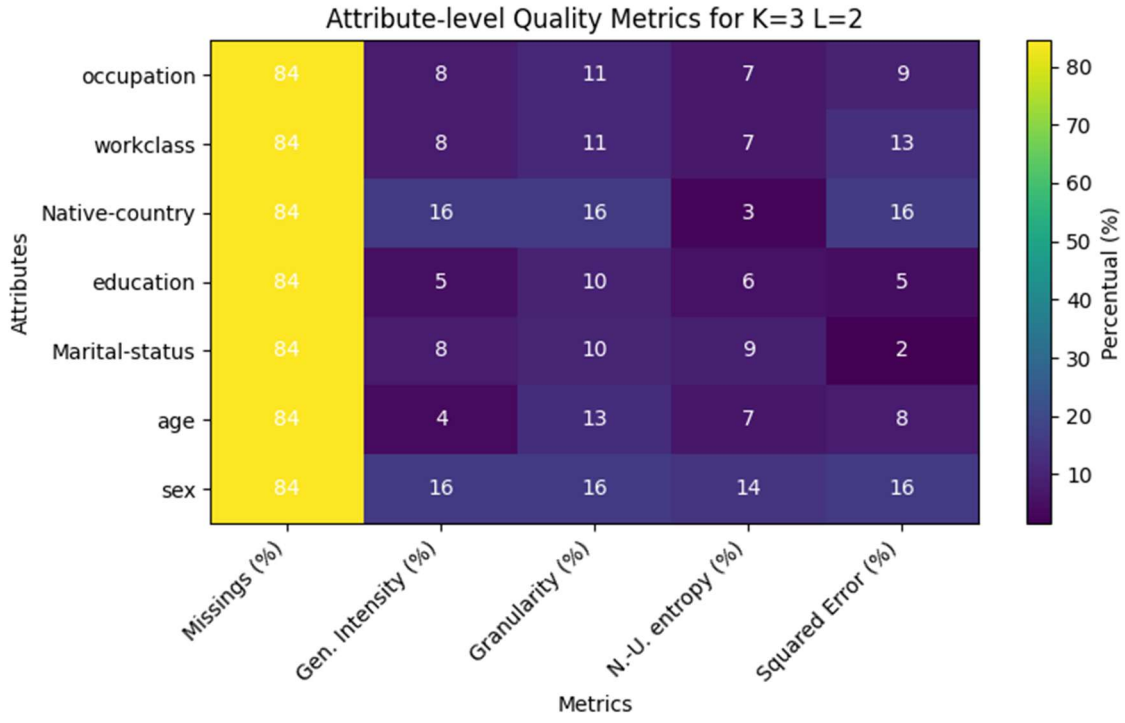
classes equivalentes (99%) prova que muitos registos foram agrupados nos mesmos grupos, e os erros quadráticos serem tão baixos (quase 0%) comprova que pouca transformação ocorreu. Este é um dos piores cenários em termos de utilidade, pois o conjunto de dados possui uma baixa transformação controlada, mas com forte supressão ou agrupamento automático. Por outras palavras, os dados têm pouca distinção entre registos, a informação foi quase toda perdida, mas grupos grandes existem. Este é um cenário onde não houve anonimização estruturada, mas sim uma destruição generalizada da informação.

Na segunda configuração (k = 5, l = 4), observa-se um cenário semelhante. Nesta situação, todas as métricas de qualidade têm o valor de 0%, exceto a discernibilidade e a média de tamanhos de classes equivalentes, que não são aplicáveis (N/A). Este cenário é considerado ainda mais extremo, e significa ou que nenhuma anonimização foi aplicada, ou que a anonimização foi tão forte que os dados foram totalmente apagados. Como na análise de riscos constatou-se que os valores eram igualmente 0%, o que não se comparava à análise inicial, conclui-se que estamos perante do segundo caso. Ao não haver dados, a entropia é 0% pois não há diversidade, os erros quadráticos são 0%, pois não há o que comparar, e a discernibilidade e média de tamanhos de classes equivalentes não se aplicam pois não existem classes equivalentes (grupos indistinguíveis). Assim, o conjunto de dados obtido é inutilizável, e percebe-se que os riscos obtidos são 0% pois não há dados para o atacante reidentificar um indivíduo.

Por último, na terceira configuração, (k=10, l = 2), existe uma alta generalização (59%), o que significa que os dados foram fortemente agrupados, e uma alta granularidade (35), que é contraditório, pois, apesar da generalização intensa, muitos valores preservaram níveis de detalhe. Neste caso, pode ter ocorrido generalização seletiva (em poucos atributos), ou há muitos valores diferentes dentro de cada grupo. A entropia é de 35%, um valor que demonstra que os dados ainda têm alguma variedade, e possui uma discernibilidade alta (83%), isto é, a maioria dos registos ainda é distinguível, o que indica um certo risco de reidentificação, analisado no capítulo anterior. A média de tamanhos de classes equivalentes (99%) sugere heterogeneidade entre atributos, ou seja, os grupos de indistinguíveis ficaram grandes, ocorrendo menos individualização e mais anonimato, mas ainda em risco devido à discernibilidade. Por último, erros quadráticos ao nível de atributos e registos é de 42% e 31%, o que sugere que alterações foram feitas, mas de forma moderada, e o erro quadrático de agregação é de 0%, o que demonstra que agregações específicas não foram distorcidas, implicando uma boa preservação estatística geral. Assim, obtém-se uma privacidade razoável em alguns atributos, havendo um risco considerável de reidentificação apesar de bons indicadores de utilidade, devido à alta discernibilidade. É ideal para cenários onde a utilidade estatística é crítica, mas não recomendado se o foco principal for proteção contra ataques a registos individuais.

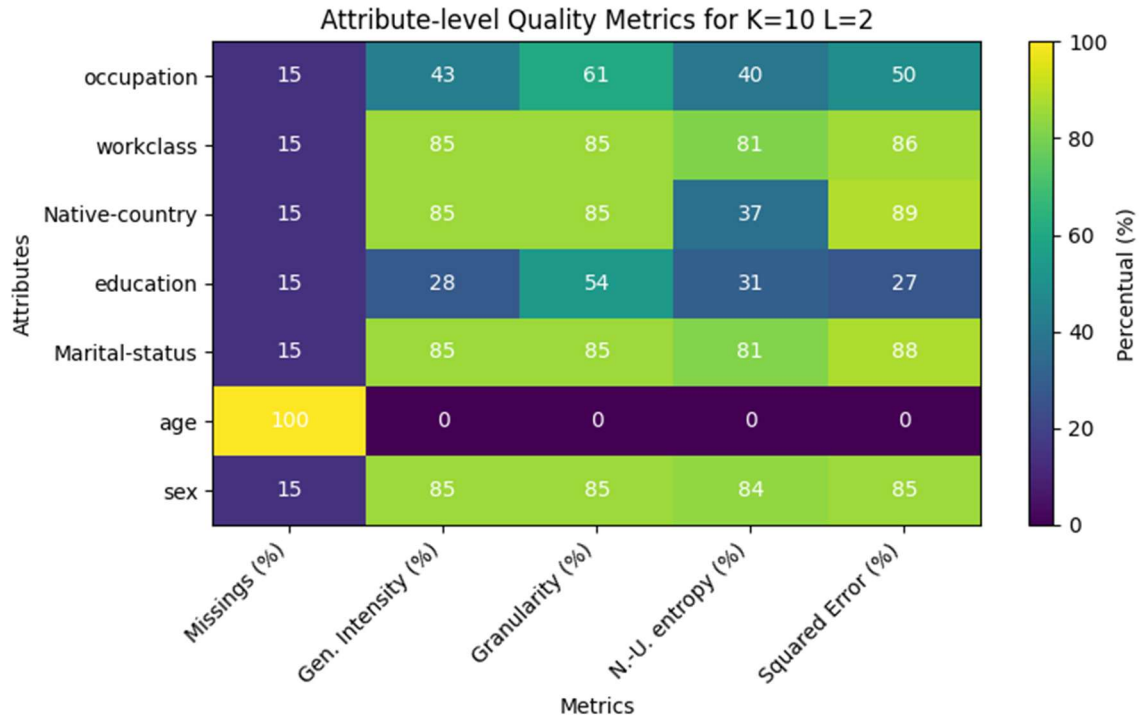
Conclui-se assim que, se o objetivo principal for privacidade máxima, a configuração (k = 5, l = 4) é a mais segura, pois todos os riscos são de 0%, isto é, nem o atacante mais forte consegue reidentificar, mas os dados não têm qualquer utilidade. É ideal para contextos sensíveis onde nenhuma violação de privacidade pode ocorrer, mesmo que o conjunto de dados não seja usado posteriormente (ex: casos judiciais, dados pessoais sensíveis, políticas de *privacy-by-design*). Se o objetivo principal for utilidade prática com privacidade razoável, a configuração (k = 10, l = 2) é o melhor compromisso, pois o risco está abaixo de todos os limiares e a utilidade é preservada de forma equilibrada, mesmo que a discernibilidade seja alta. É ideal para contextos como estatísticas públicas, estudos de mercado e *Machine Learning,* onde manter estrutura e valor analítico é essencial, mas com controle de risco. O pior cenário é a configuração (k=3, l =2), pois nem protege efetivamente, nem preserva utilidade, devendo ser descartado como opção viável.

Análise de utilidade (Qualidade dos atributos)



Na configuração (k=3, l=2), a supressão de todos os atributos é de 84%, indicando que a maior parte dos registos foi suprimida para atingir as classes de equivalência de tamanho maior ou igual a três. A intensidade de generalização mais alta, de aproximadamente 15,6%, ocorre em *sex* e *native-country*. Isso sugere que, para esses atributos, houve maior generalização, provavelmente agrupando valores raros sob categorias “Outros” ou faixas amplas. Os atributos, *age*, *marital-status* e *education* apresentam intensidade de generalização muito baixa, entre 3,9% e 7,8%). Para

garantir que cada combinação de valores apareça ao menos três vezes, idealmente aumentaríamos a generalização em age, usando, por exemplo, faixas decenais mais largas, e em *education*, agregando níveis de escolaridade semelhantes. A entropia para todos os atributos é baixa, indicando baixa variedade de valores distintos em cada grupo, isto é, as generalizações ainda não garantem diversidade interna. O erro quadrático, mesmo que variado, está sempre relativamente próximo. Isso sinaliza que, para alcançar diversidade, foi feita alguma perturbação, mas insuficiente, pois ainda se vê muita homogeneidade dentro das classes anónimas.



Na configuração (k=10, l=2), só 15% de valores em praticamente todos os QIDs foram suprimidos, exceto *age*, que foi totalmente suprimido. Isso indica que, para alcançar grupos de tamanho maior ou igual a dez, a maioria dos registos foi mantida, o que é um ponto positivo na utilidade dos dados. Houve uma generalização elevada (85%), envolvendo *workclass, native-country, marital-status* e *sex*, e uma granularidade moderada. Constata-se assim que os valores originais foram praticamente todos convertidos em categorias amplas ou "Outros". Mesmo assim, devido ao modelo de privacidade *l-Diversity*, esta generalização pode estar relacionada ao facto de ser necessário garantir pelo menos dois valores distintos em cada grupo. Para esses QIDs, a entropia foi igualmente alta, tal como o erro quadrático, sugerindo perda de informação. Analisando ao detalhe, nos atributos *sex, marital-status, native-country* e *workclass*, onde houve 85% de generalização e entre 83% e 88% de erro quadrático, pouco resta da informação original, dificultando qualquer análise mais granulada, como diferenças de renda por sexo ou país. O QID age foi suprimido a 100%, ou seja, não resta nenhum dado desse atributo. Por último, os atributos *education* e

occupation possuem níveis intermédios de generalização (28-43%) e erro quadrático (27-50%), sugerindo que ainda há alguma variação, como por exemplo entre "Licenciatura vs. Pós-graduação" ou "Saúde vs. Educação vs. I.T.".

Assim, com k = 10, a maioria dos registos são mantidos, mas a idade é totalmente apagada, e vários QIDs são fortemente generalizados, o que resulta em perdas significativas de informação. Já com k = 3, ocorre generalização mínima, porém suprime a quase totalidade da base, o que pode inviabilizar quase qualquer análise. Conclui-se que, k = 10 dá uma melhor cobertura, mas utilidade reduzida nos atributos mais generalizados, e k = 3 prioriza exatidão local, mas perde a maior parte da base, não tendo muito por onde analisar.

4.2.2. Variação de definições de transformação

Para o estudo abrangente de anonimização de dados, foram também variados outros parâmetros, não relacionados aos modelos de privacidade aplicados, mas sim às transformações de dados. Assim, foi escolhida uma configuração de k = 5 e l = 2, e foram variados parâmetros como o limite de supressão, a medida de utilidade e o peso de atributos.

4.2.2.1. Limite de supressão

O limite de supressão é o percentual máximo permitido de registos que podem ser suprimidos para atender aos critérios de privacidade. Foi estudado inicialmente o seu impacto, variando-o da seguinte forma:

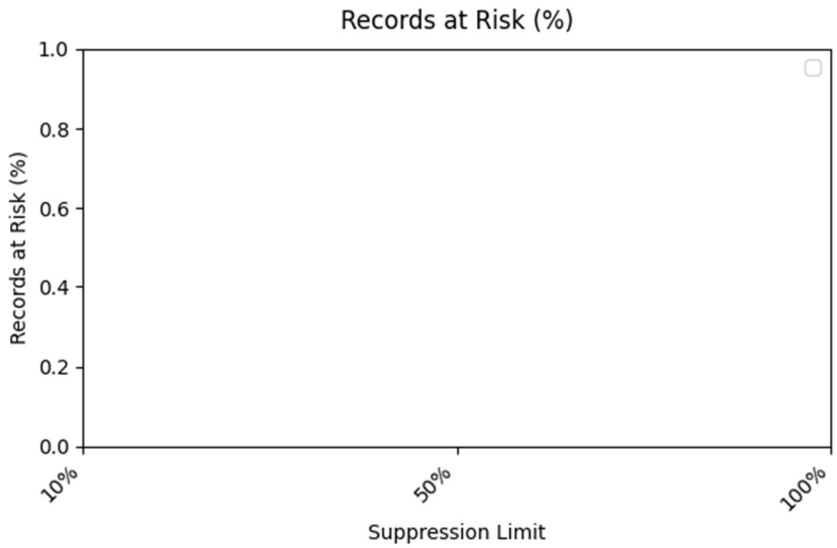
- Limite = 10%,

- Limite = 50%,

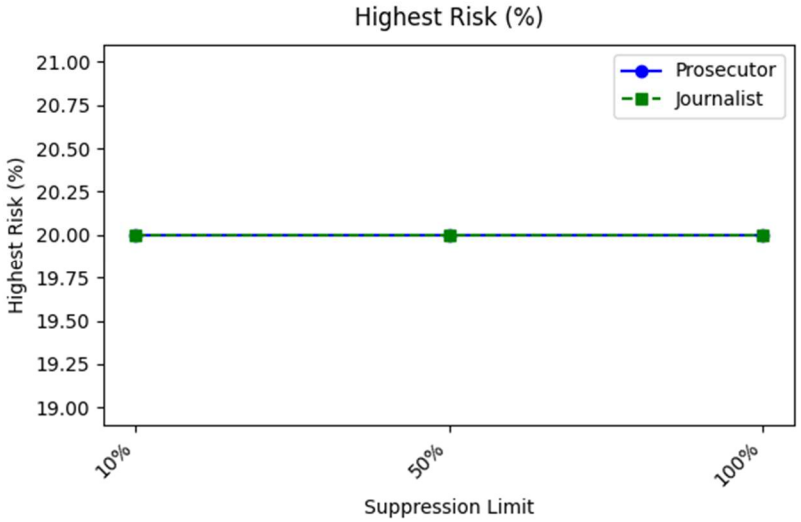
- Limite = 100%,

Onde, para todos os casos, foi usada a medida de utilidade *Discernibility*, e todos os atributos possuíam o peso de 0.5.

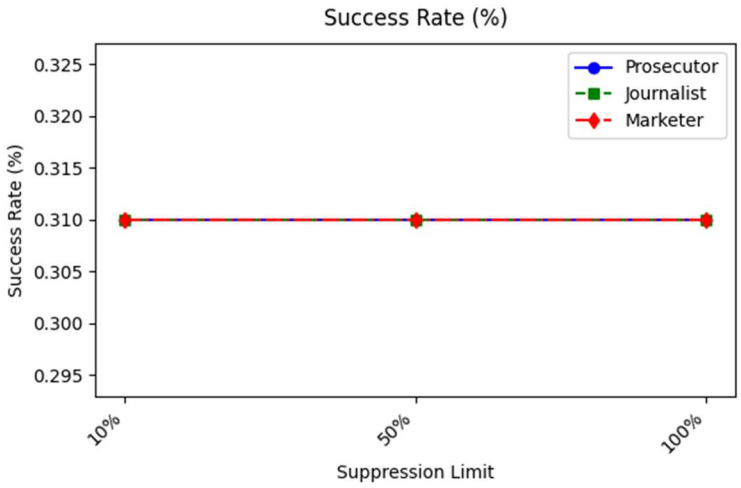
Análise de Riscos



Analisando o seguinte gráfico, não há nenhum registo com probabilidade de reidentificação de 100% (0%), isto é, não existe nenhum registo isolado, logo nenhum indivíduo é unicamente identificável, o que confirma a correta aplicação de *k-Anonymity*.



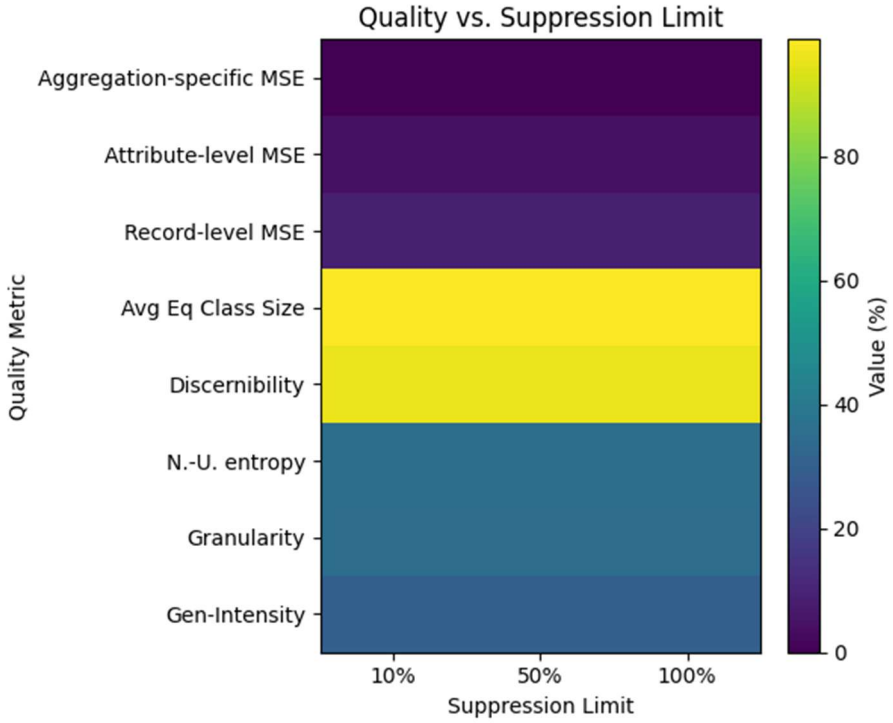
De seguida, é possível observar que o *Highest Risk* para os três limites de supressão é de 20%, o que pode ser visto como a menor classe de equivalência, que tem exatamente 5 registos, logo o pior caso de reidentificação é de 1/5 (20%).



Concluindo, a taxa de sucesso para todos os casos analisados é de 0.31%. Se todas as classes de equivalência tivessem exatamente tamanho 5, então a taxa seria de 20%, mas, no conjunto estudado, várias classes saíram com um tamanho superior a 5, o que faz descer a média para 0.31%. Isto demonstra que a generalização gerou muitas classes maiores do que o mínimo, reforçando a proteção média.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha do limite de supressão.



Primeiramente, a intensidade de generalização é de 30.5%, o que significa que, em média, cada atributo QID foi generalizado a aproximadamente 30% da profundidade máxima. A granularidade é de 35.3%, o que indica que pouco mais de 30% da "resolução" original de QID's sobreviveu. Já a entropia normalizada é mais baixa, com 35.7%, o que revela a variabilidade original restante dos dados e reflete a capacidade de suportar análises estatísticas que dependem da dispersão. A discernibilidade deste estudo foi de 96%, o que indica alta generalização. O tamanho médio de classes de equivalência é de 99%, o que indica que o tamanho médio ficou praticamente igual a 5 (k), e mostra que quase todos os grupos têm exatamente k registos. Os erros quadráticos ao nível de registo (8.8%) e atributo (4.5%) mostram que pouca precisão foi perdida. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Depreende-se assim que todas as métricas se mantêm idênticas para supressão máxima de 10%, 50% e 100%, porque nenhuma supressão foi aplicada efetivamente. A utilidade do conjunto de dados é determinada só pela generalização, não pela supressão, nestes parâmetros. Assim, variar o limite de supressão não altera nenhum valor.

4.2.2.2. Medidas de utilidade

Uma medida de utilidade é uma função que avalia a qualidade dos dados após a aplicação de técnicas de anonimização como generalização ou supressão. Estas ajudam a quantificar o quanto os dados transformados mantêm a sua utilidade para análises futuras, equilibrando a proteção da privacidade com preservação de informação. Foram, assim, analisados os impactos de três medidas diferentes no risco e na utilidade:

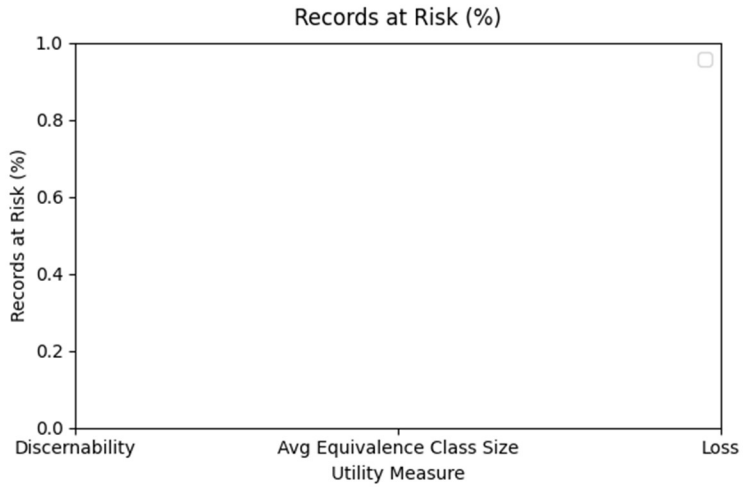
- *Loss* (ou granularidade),

- Discernibilidade,

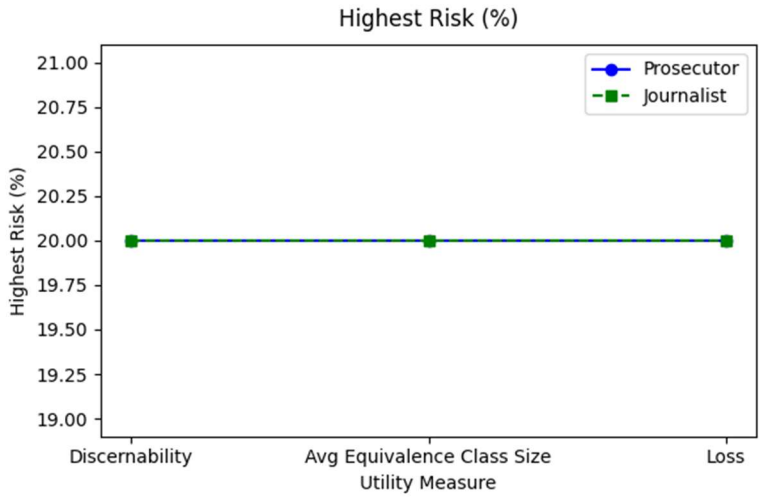
- Tamanho médio de classes equivalentes,

Em que para todos os casos, foi usado um limite de supressão de 100%, e todos os atributos possuíam o peso de 0.5.

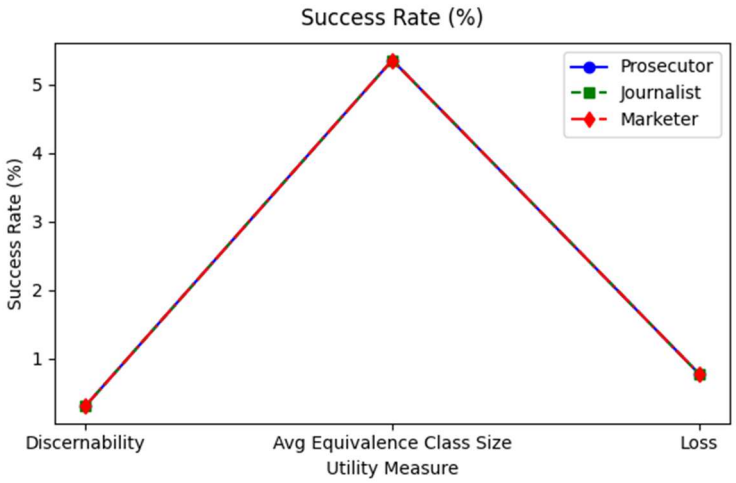
Análise de Riscos



É possível analisar no gráfico acima que para nenhuma das medidas de utilidade usadas, existem classes unitárias. Logo, para todas, *Records at Risk* é de 0%.



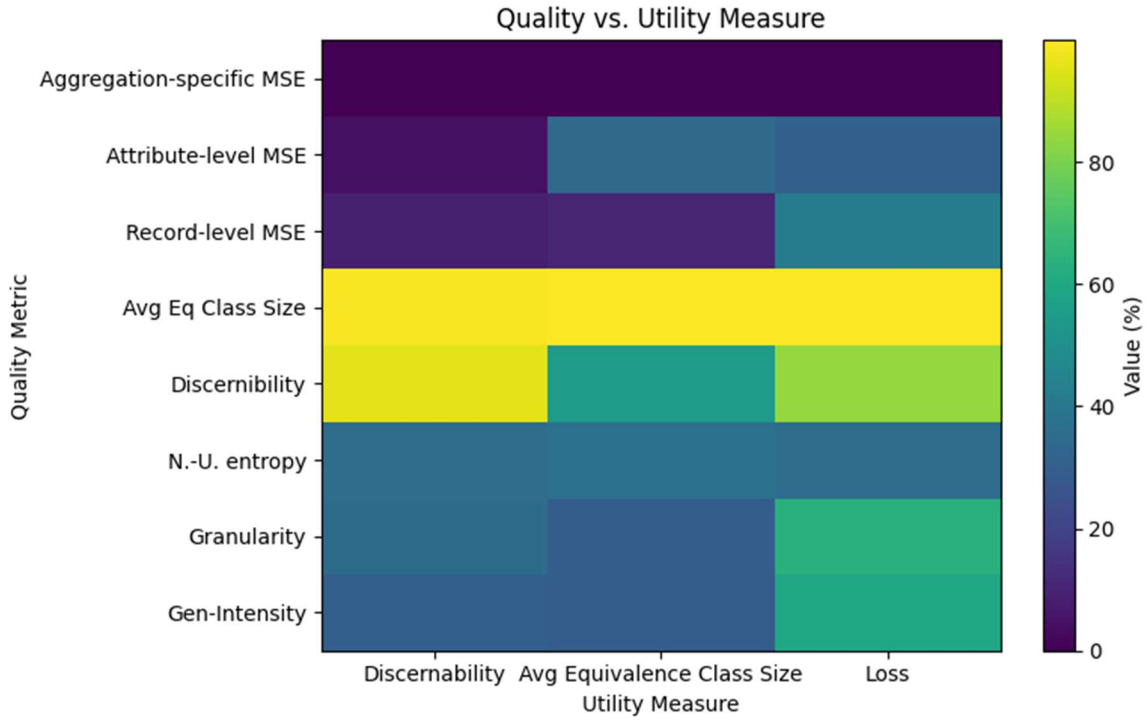
Já o risco máximo, ainda idêntico para as medidas diferentes, é de 20%, o que pode ser visto como a menor classe de equivalência, que tem exatamente 5 registos, logo o pior caso de reidentificação é de 1/5 (20%).



Em contraste com as análises anteriores, a taxa de sucesso da medida Discernibilidade é de 0.30%, pois as classes de equivalência ficaram grandes e homogéneas. Usando a medida de tamanho médio de classes equivalentes, a taxa de sucesso de reidentificação de um indivíduo é de 5.35%. Por último, usando *Loss*, a taxa desce para 0.78%.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha da métrica de utilidade.



Na primeira medida, Discernibilidade, a generalização teve uma intensidade de 30% e granularidade de 35%, que indicam uma moderada generalização e boa preservação de detalhe nos QIDs. A entropia é de 35.6% e a discernibilidade é de 96%, que revela que, embora exista anonimização, as classes ficaram relativamente distintas. O tamanho médio de classes de equivalência é de 99%, o que indica que o tamanho médio ficou praticamente igual a 5 (k), e mostra que quase todos os grupos têm exatamente k registos. Os erros quadráticos são baixos, de 8.8% e 4.5%, mostrando que a aproximação aos valores originais foi muito fiel. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Usando a segunda medida, *Average Equivalence Class Size*, essa mesma métrica é otimizada, com quase 100%. A generalização e granularidade são similares, com 29%, indicando baixa generalização e razoável preservação de detalhe. A entropia é de 38% e a discernibilidade de 56%, sinalizando classes menos diversas. Apesar dos erros quadráticos serem baixos, o de atributo é elevado (35%), indicando que, para atingir classes do tamanho ideal, houve perda considerável de precisão em atributos individuais.

Por último, com a medida *Loss*, a intensidade de generalização é alta (60%), tal como a granularidade (63%), o que demonstra que o foco em minimizar a *Loss* leva a generalizações mais intensas, porém preserva mais variedade nas classes. A entropia é de 36% e a discernibilidade é alta (84%), indicando classes ainda muito distintas. Os erros são mais elevados (31-42%), exceto o de agregação, que é sempre 0%, visto que a prioridade foi manter a utilidade global em vez de minimizar distorções pontuais.

Numa comparação geral:

- A discernibilidade é a melhor escolha se for necessária máxima fidelidade aos valores originais;

- Se o essencial for ter sempre k registos por grupo, a melhor opção é o tamanho médio de classes;

- Para um compromisso geral entre precisão local e global, *Loss* seria a melhor escolha.

4.2.2.3. Peso de Atributos

O peso de um atributo é a importância atribuída a cada atributo na análise de utilidade. Atributos com pesos maiores têm mais influência na avaliação da qualidade dos dados.

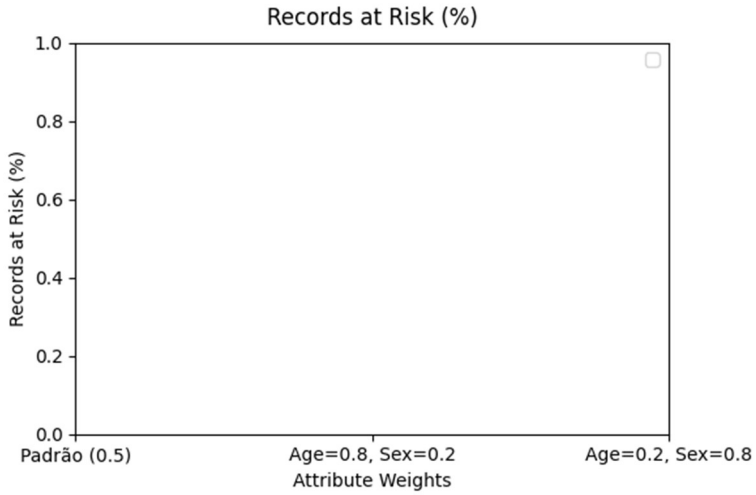
Para o seu estudo, usaram-se os seguintes dados:
- Peso padrão (0.5 para todos os atributos),

- Atributo *Age* com peso de 0.8 e atributo *Sex* com peso de 0.2, e restantes com peso padrão,

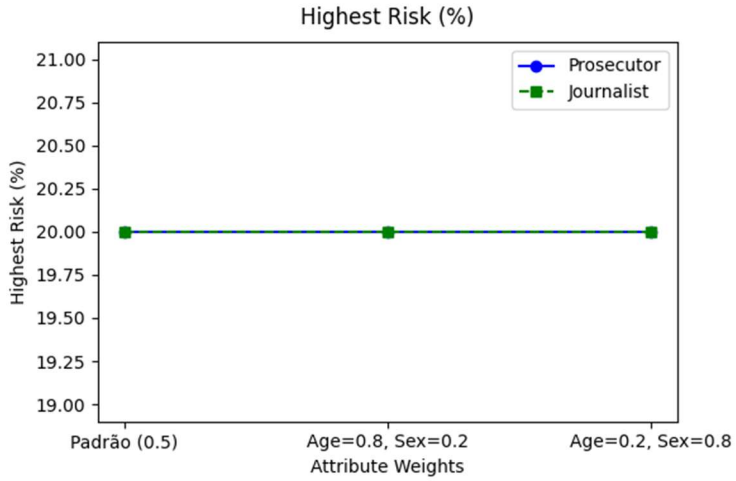
- Atributo *Age* com peso de 0.2 e atributo *Sex* com peso de 0.8, e restantes com peso padrão,

Em que para todos os casos, foi usado um limite de supressão de 100% e a medida de utilidade *Loss*.

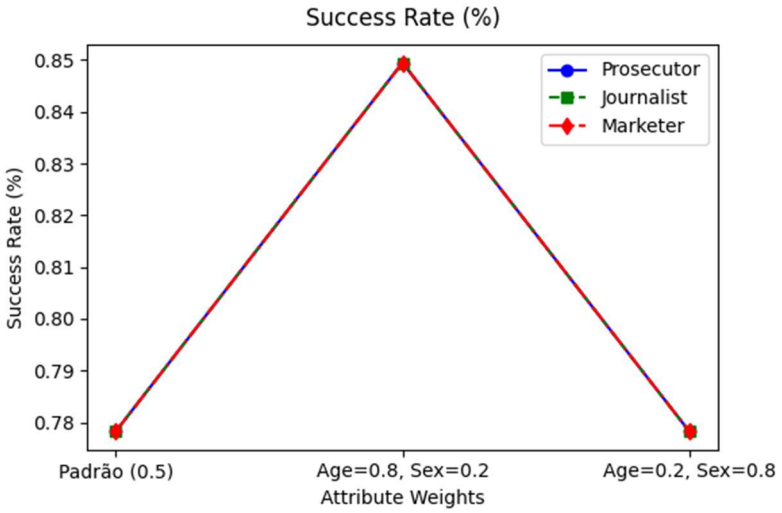
Análise de Risco



Observando o gráfico acima, é possível concluir que *Records at Risk* mantém-se invariável (0%), indicando que não existem classes unitárias**.**

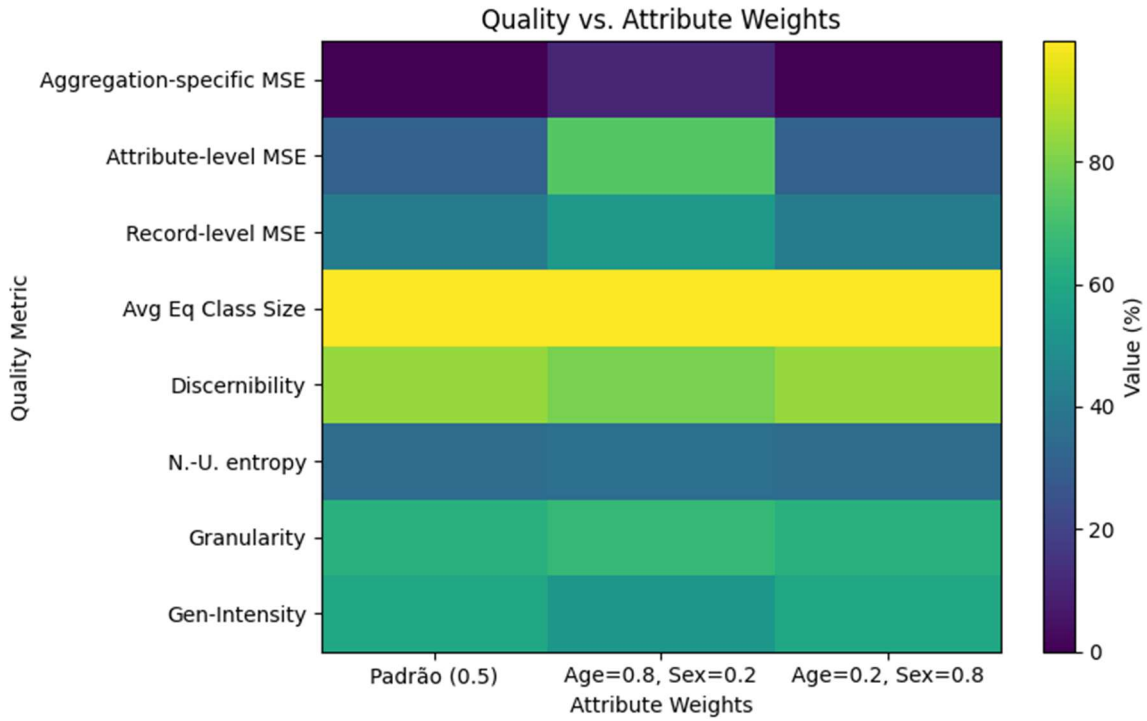


Tal como na situação anterior, o *Highest Risk* é invariável (20%), confirmando que todos os grupos têm pelo menos 5 registos.



O *Success Rate* sobe ligeiramente, de 0.78% para 0.85%, quando se dá muito peso no atributo age. Isso significa que, em média, um atacante tem mais chance de acerto quando o algoritmo se preocupa mais em preservar a informação etária.

Análise de Utilidade



Com o cenário padrão, observa-se uma utilidade moderada. A generalização foi relativamente intensa (60%), necessária para garantir *k-Anonymity*, mas implica perda

de detalhe nos QIDs. Ainda assim, a granularidade e discernibilidade são altas (63% e 84%), mostrando que sobra diversidade suficiente. O erro a nível de registo (42%) e atributo (31%) não é desprezível, visto que análises que dependam de valores exatos ou limiares apertados, podem ser afetadas. Por outro lado, a preservação integral das agregações garante fiabilidade total em sumários estatísticos. O tamanho médio de classes quase constante em k (99.6%) sugere que a generalização se distribuiu uniformemente, sem criar grupos muito maiores que o mínimo, ajudando assim a manter o risco médio baixo (0.78%).

Aumentando o peso do atributo *age* para 0.8 e diminuindo o peso do atributo *sex* para 0.2, é possível ver algumas variações na utilidade. A intensidade de generalização cai de 60% para 53%, pois o algoritmo faz menos generalização em idade para proteger esse atributo. A granularidade aumenta (67%), refletindo mais variedade nos QIDs em geral. Os erros disparam, sobretudo a nível de atributo, que passa de 31% para 73%, e o erro de agregação aparece, com 11%, porque outros QIDs (sexo) são sacrificados para preservar a idade. O *Sucess Rate* sobe, visto que os grupos de equivalência se tornam ligeiramente menores em média, logo o risco médio é maior.

Por fim, diminuindo o peso do atributo *age* para 0.2 e aumentando o peso do atributo *sex* para 0.8, é possível observar que os valores são idênticos aos do cenário padrão. Isto sugere que, na prática, dar mais peso ao atributo *sex* não altera a escolha de generalizações/supressões, possivelmente porque o sexo já era o atributo menos variado ou menos influente na configuração padrão.

Conclui-se que, dar peso ao sexo não traz benefício prático adicional face ao padrão e, para um compromisso equilibrado, o melhor cenário é o padrão, visto que dar pesos extra a atributos pode mesmo comprometer o conjunto de dados e aumentar o risco de reidentificação.

4.3. k-Anonymity com t-Closeness

A segunda combinação de modelos escolhida foi *k-Anonymity* com *t-Closeness*, para lidar com atributos sensíveis (*race, salary_class*).

4.3.1. Variação de *k* e *t*

Foram escolhidos diferentes parâmetros para k-Anonymity e t-Closeness, de forma a analisar a mudança de valores percentuais de risco e utilidade:

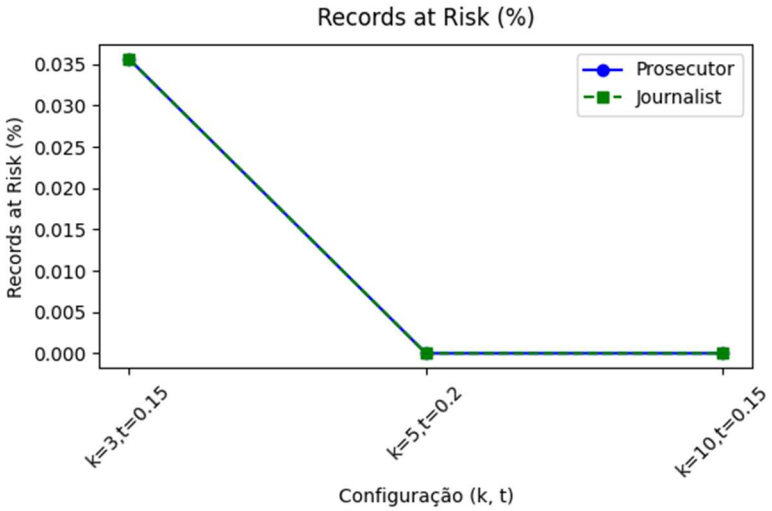
- k = 3 e t = 0.15;

- k = 5 e t = 0.2;

- k = 10 e t = 0.15.

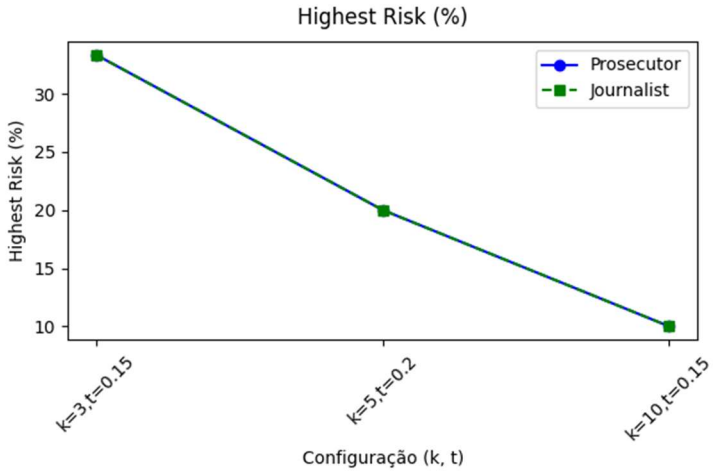
Em todas as anonimizações feitas, foram utilizadas medida de utilidade *Loss*, limite de supressão de 100%, e pesos de atributos padrão, de 0.5.

Análise de Riscos



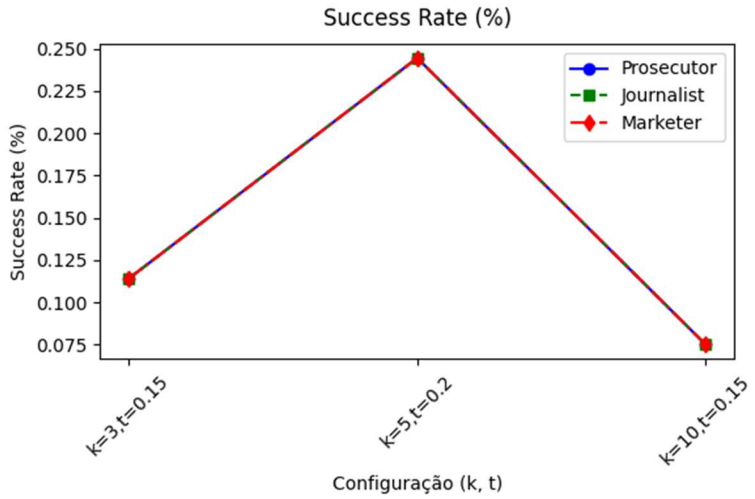
É possível analisar na figura acima que, quando k = 3 e t = 0.15, 0.035% de registos estão em risco de ser reidentificados tanto num modelo de Promotor (*Prosecutor Attacker Model*), como no modelo de Jornalista (*Journalist Attacker Model*). Quanto maior é k, menor é este valor, indicando assim menor risco de identificação.

É possível observar isso quando k = 5 (e t = 0.2), e k = 10 (e t = 0.15), pois ambos os casos têm 0% dos recordes em risco. Comparativamente à análise inicial, de valor 68.4%, observa-se um claro decréscimo nesta probabilidade.



Na figura acima, é possível observar a percentagem de risco máximo para ambos os modelos de atacante. É percetível um pico quando k = 3 e t = 0.15, de mais de 30%, ultrapassando o limiar. Quanto maior é k, menor é este valor, indicando assim menor risco de identificação.

Assim, o ponto ótimo deste risco máximo ocorre quando k = 10 e t = 0.15, de valor 10%. Em relação à análise inicial, com um risco máximo de 100%, verifica-se uma melhoria do mesmo.



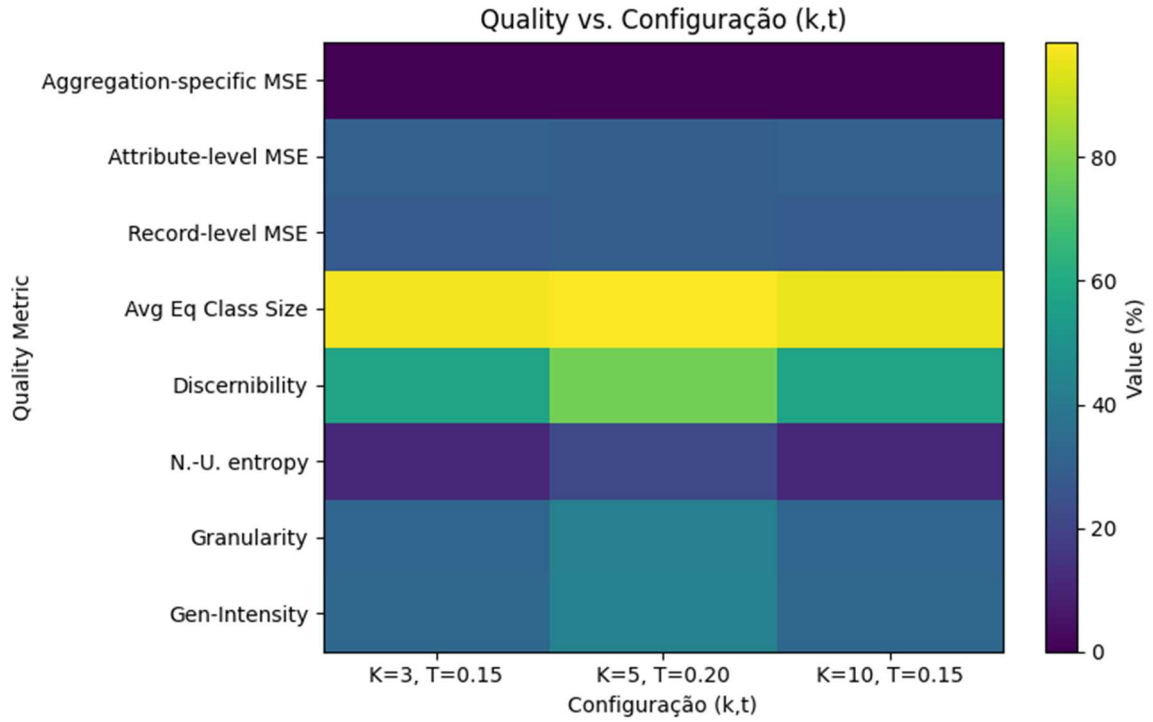
Por último, o *Success Rate* é significativamente baixo em todos os modelos, mas há um que se destaca: k = 10 e t = 0.15, com aproximadamente 0.075%, o que significa que nenhum modelo tem grande possibilidade de atacar com sucesso a base de dados. Quando k = 3 e t = 0.15, a taxa de sucesso é de quase 0.125%. Já com k = 5 e t = 0.2, a taxa de sucesso é de aproximadamente 0.25%, demonstrando que, apesar de baixo, ainda existem padrões que permitem a reidentificação. Em todos os cenários, nenhum destes valores é considerado perigoso, visto que o limiar de *Success Rate* é de 5%, e nenhum excede este valor, embora o par k = 10 e t = 0.15 seja o mais seguro.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha de parâmetros (k,t). No eixo Y estão as métricas usadas

na avaliação da qualidade do conjunto de dados, e no eixo X estão as configurações escolhidas para a anonimização dos dados

Analisando o mesmo de baixo para cima, é possível tirar diversas informações acerca da utilidade do conjunto de dados.



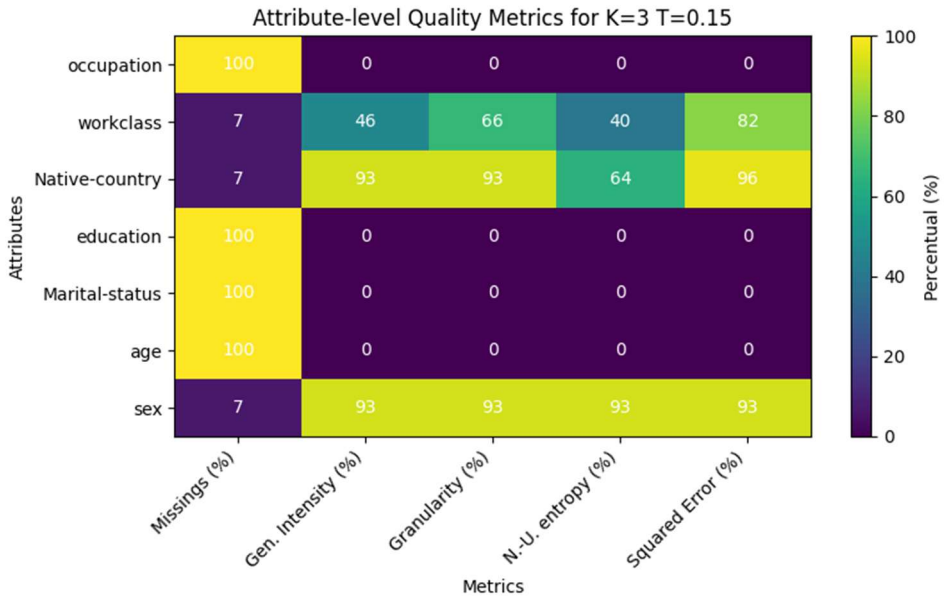
Na primeira configuração, (k = 3, t = 0.15), a intensidade de generalização é 33%, ou seja, pouca generalização foi aplicada e os dados do conjunto foram pouco transformados (aprox. 1/3) via taxonomias. Curiosamente, a granularidade é 32%, quase igual a intensidade de generalização, o que significa uma baixa qualidade dos dados e que a maioria dos dados é agregada. A entropia normalizada é baixa (11%) que significa pouca variedade nos dados e a discernibilidade é mais alta (57%), o que pode ser interpretado como a existência de valores pouco repetidos ou apagados, que se distinguem entre si, isto é, o conteúdo tornou-se mais diverso e distintivo. A média de tamanhos de classes equivalentes (97%) mostra que muitos registos foram agrupados nos mesmos grupos. Os erros quadráticos (*Record-level MSE*: 28%, *Attribute-level MSE*: 31% e *Aggregation-specific MSE*: 0%) significam que existe um nível moderado de alteração nos registos e atributos (ligeiramente alterados - atributos e registos completos). Isto significa que os dados têm uma qualidade reduzida, mas ainda mantêm um alto nível de utilidade. Este é um cenário onde não houve anonimização estruturada, mas sim uma destruição generalizada da informação.

Na segunda configuração (k = 5, t = 0.2), a intensidade de generalização e a granularidade são ambos de 43% o que significa uma generalização moderada (quase metade dos dados foi transformada) e que os dados foram relativamente agregados,

com pouco detalhe. A entropia normalizada baixa (21%) e a discernibilidade alta (77%) significam que existe uma baixa diversidade nos dados e uma alta distinção entre registos. A média de tamanhos de classes equivalentes é alta e significa que quase todos os registos foram agrupados em grandes classes que reforça o k-anonimato. Os erros quadráticos (*Record-level MSE*: 29%, *Attribute-level MSE*: 30% e *Aggregation-specific MSE*: 0%) significam que existe uma distorção moderada nos registos completos e atributos individuais sem perdas nas agregações estatísticas.

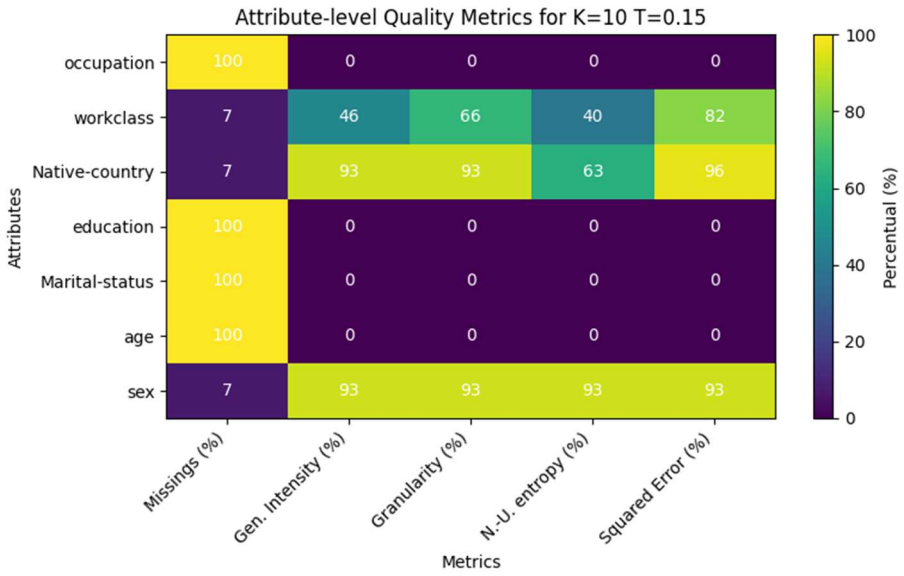
Por último, na terceira configuração, (k=10, t = 0.15), a intensidade de generalização e a granularidade são ambos aproximadamente 33% o que significa há pouca generalização e menor precisão dos valores originais. A entropia normalizada baixa (11%) e a discernibilidade moderada (57%) mostram que existe uma baixa diversidade de valores e cerca de metade dos dados são identificáveis. A média de tamanhos de classes equivalentes é alta (95%) significa que os registos foram fortemente agrupados o que garante anonimato. Os erros quadráticos (*Record-level MSE*: 28%, *Attribute-level MSE*: 31% e *Aggregation-specific MSE*: 0%) significam que existe uma distorção moderada nos registos completos e atributos individuais sem perdas nas agregações estatísticas.

Concluímos que todas as configurações demostram um compromisso razoável entre utilidade e privacidade para análises agregadas. A configuração (k = 5, t = 0.2) tem uma proteção à privacidade melhor, mas os dados são mais distortidos. A configuração (k = 5, t = 0.15) preserva mais detalhes, mas oferece menor proteção à privacidade. Em todos os casos, a ausência de erro nas agregações reforça a utilidade para estudos estatísticos. A configuração melhor é (k = 10, t = 0.15) porque tem o melhor equilíbrio entre privacidade e utilidade. A configuração melhor em termos de privacidade é (k = 5, t = 0.2) e a configuração melhor em termos de utilidade é (k = 10, t = 0.15).



Na configuração (k=3, t=0.15), o algoritmo recorreu a supressão total ou a generalizações extremas para cumprir simultaneamente os dois modelos de privacidade usados. A generalização nos atributos *sex* e *native-country* foi de 93%, com erro quadrático também em torno de 93-96%. Por outras palavras, quase toda a informação original desses atributos foi convertida em categorias tão amplas que pouca resta de discriminante. Já a *workclass* apresenta uma generalização mais moderada, de 46%, mas ainda tem um erro quadrático de 82% e entropia interna baixa (39%), indicando pouca variabilidade preservada em cada classe de equivalência.

A supressão total de 4 em 7 atributos e a generalização quase completa dos demais tornam o *dataset* praticamente inútil para qualquer análise descritiva ou preditiva minimamente granulada. Embora o critério de *t-closeness*, que mantém a distribuição de cada atributo sensível a uma distância menor ou igual a 0.15 do original, e *k-anonymity* estejam atendidos, perdeu-se toda a estrutura informacional.



Na configuração (k=10, t=0.15) ocorrem supressões variadas. Em *sex* e *native-country* a maior parte dos registos foi preservada, com uma supressão de 7%, mas ao custo de generalizações extremas (93%), sendo quase toda a informação original transformada em "Outros" ou categorias muito amplas. Nos atributos *age*, *marital-status*, *education* e *occupation*, esses campos foram totalmente removidos para satisfazer *t-closeness* e um k maior ou igual a 10. O QID *workclass* recebeu uma generalização moderada (46-66%), e supressão baixa (7%), mas ainda apresenta um erro quadrático elevado (82%) e entropia baixa (40%), indicando homogeneidade interna.

Com quatro de sete atributos totalmente suprimidos, e os demais com generalização massiva, o *dataset* perde quase toda a variação necessária para análises

demográficas, estatísticas ou preditivas. Embora a configuração seja atendida, o valor analítico dos dados torna-se praticamente nulo.

4.3.2. Variação de Definições de Transformação

Para o estudo abrangente de anonimização de dados, foram também variados outros parâmetros, não relacionados aos modelos de privacidade aplicados, mas sim às transformações de dados. Assim, foi escolhida uma configuração de k = 5 e t = 0.15, e foram variados parâmetros como o limite de supressão, a medida de utilidade e o peso de atributos.

4.3.2.1. Limite de Supressão

O limite de supressão é o percentual máximo permitido de registos que podem ser suprimidos para atender aos critérios de privacidade. Foi estudado inicialmente o seu impacto, variando-o da seguinte forma:

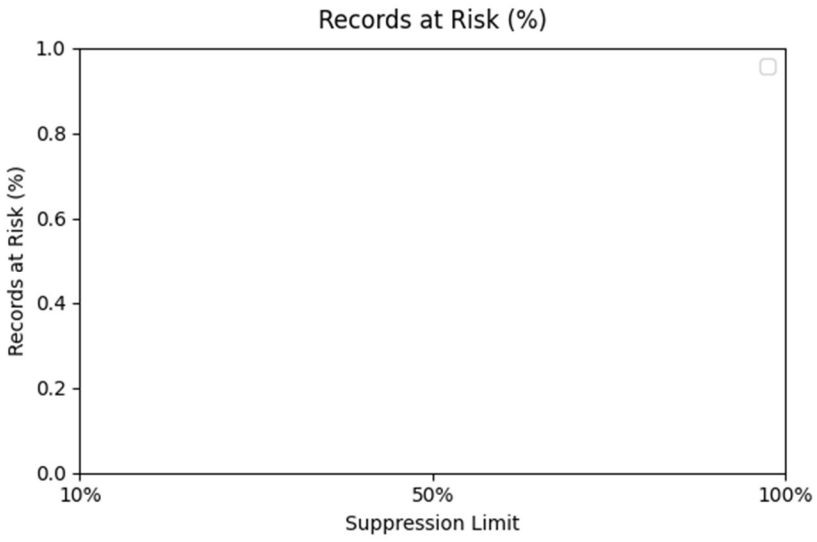
- Limite = 10%,

- Limite = 50%,

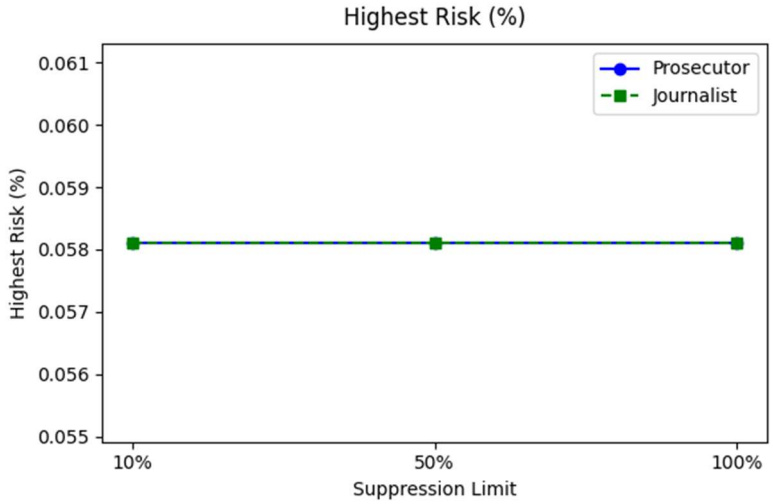
- Limite = 100%,

Em que para todos os casos, foi usada a medida de utilidade *Discernibility*, e todos os atributos possuíam o peso de 0.5.

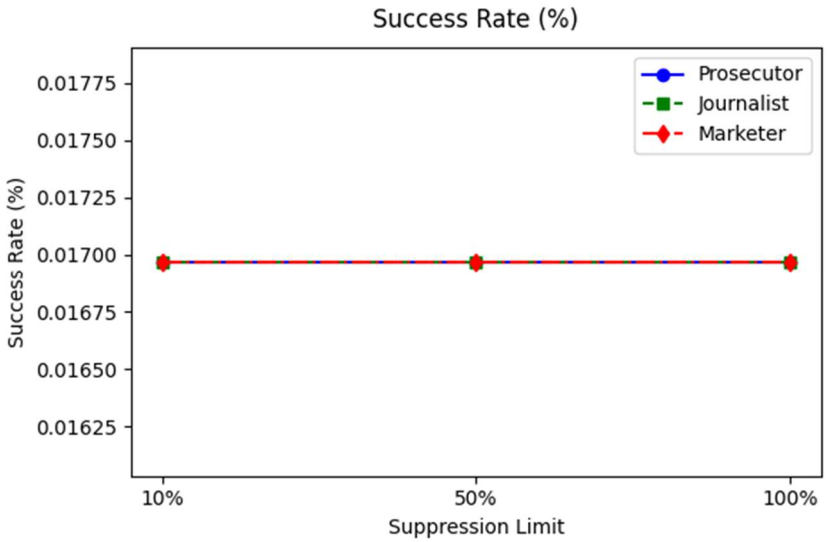
Análise de Riscos



Analisando o seguinte gráfico, não há nenhum registo com probabilidade de reidentificação de 100% (0%), isto é, não existe nenhum registo isolado, logo nenhum indivíduo é unicamente identificável.



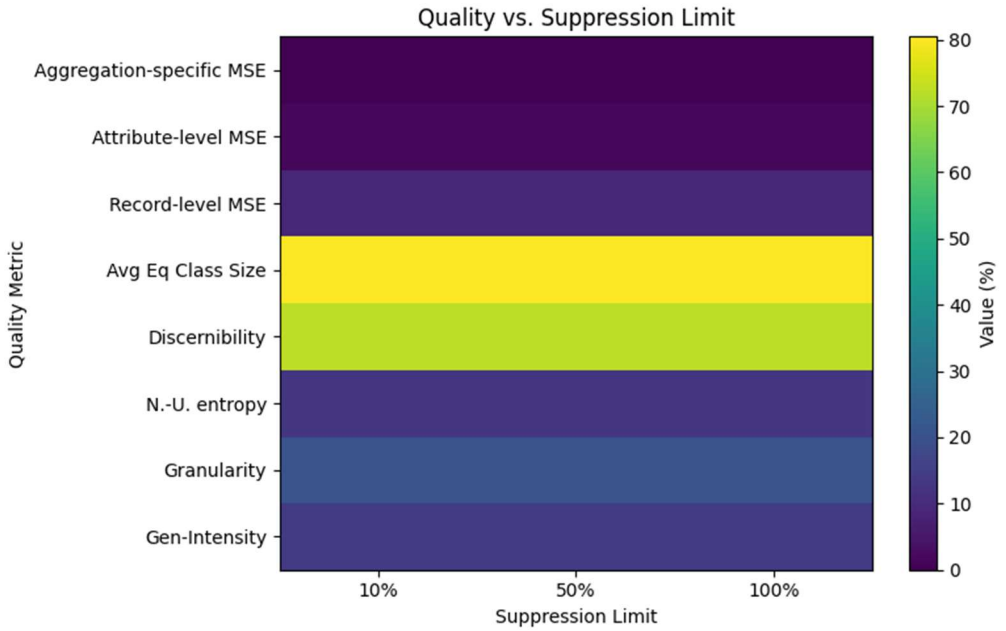
De seguida, é possível observar que o *Highest Risk* para os três limites de supressão é aproximadamente de 0.058%, logo o pior caso de reidentificação é de 1/2000 (aprox. 0.05%).



Concluindo, a taxa de sucesso para todos os casos analisados é de 0.017%. Isto demonstra que a generalização gerou muitas classes maiores do que o mínimo, reforçando a proteção à privacidade média.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha do limite de supressão



Primeiramente, a intensidade de generalização é de 13.95%, o que significa que, em média, cada atributo QID foi generalizado a aproximadamente 14% da profundidade máxima e que a maioria dos dados são quase iguais dos originais. A granularidade é de 20.69%, o que indica que há pouca agregação. Já a entropia normalizada é baixa, com 12.81%, o que revela uma diversidade pequena nos atributos. A discernibilidade é de 72.41%, o que indica um alto nível de identificação. O tamanho médio de classes de equivalência é de 80.47%, o que indica que os grupos são mais pequenos e há um maior risco de reidentificação devido a menor anonimização. Os erros quadráticos ao nível de registo (8.94%) e atributo (1.65%) mostram que os registos foram minimalmente alterados e que pouca precisão foi perdida. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Deduz-se assim que todas as métricas se mantêm idênticas para supressão máxima de 10%, 50% e 100%, porque nenhuma supressão foi aplicada efetivamente. A utilidade do conjunto de dados é determinada só pela generalização, não pela supressão, nestes parâmetros. Assim, variar o limite de supressão não altera nenhum valor.

4.3.2.2. Medidas de Utilidade

Uma medida de utilidade é uma função que avalia a qualidade dos dados após a aplicação de técnicas de anonimização como generalização ou supressão. Estas ajudam a quantificar o quanto os dados transformados mantêm a sua utilidade para análises futuras, equilibrando a proteção da privacidade com preservação de informação. Foram, assim, analisados os impactos de três medidas diferentes no risco e na utilidade:

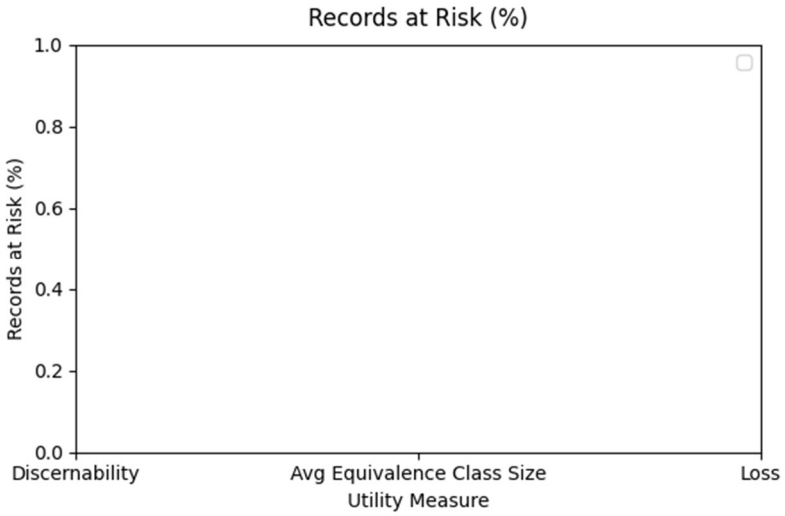
- *Loss* (ou granularidade),

- Discernibilidade,

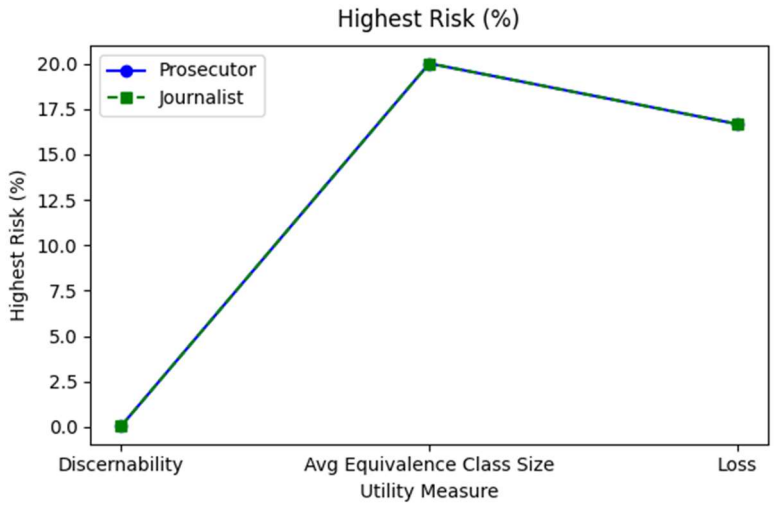
- Tamanho médio de classes equivalentes,

Em que para todos os casos, foi usado um limite de supressão de 100%, e todos os atributos possuíam o peso de 0.5.

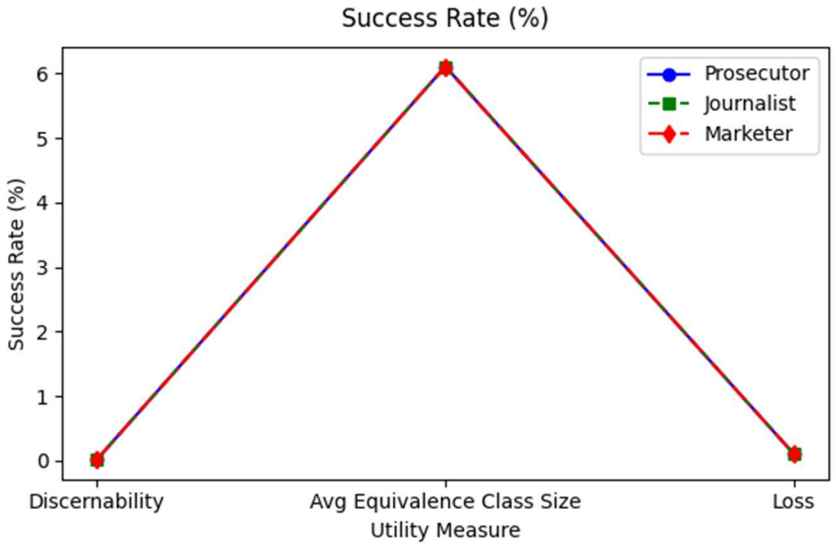
Análise de Riscos



É possível analisar no gráfico acima que para nenhuma das medidas de utilidade usadas, existem classes unitárias. Logo, para todas, *Records at Risk* é de 0%.



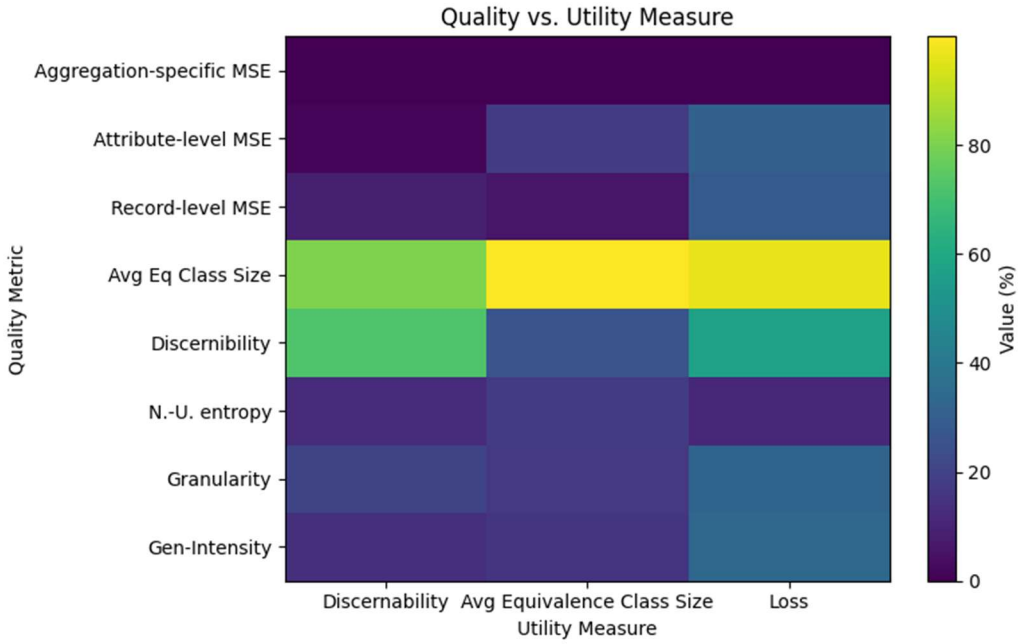
O risco máximo varia para as medidas diferentes onde Discernibilidade não apresenta nenhum risco enquanto 'Tamanho médio de classes equivalentes' apresenta um risco de 20%. *Loss* também apresenta um risco relativamente alto (aprox. 17.5%). Logo, é melhor usar Discernibilidade como a medida de utilidade.



A taxa de sucesso da medida Discernibilidade é de 0.01% e a medida de *Loss* é de 0.1% o que indica melhor proteção à privacidade. Usando a medida de tamanho médio de classes equivalentes, a taxa de sucesso de reidentificação de um indivíduo é de 5.35%, pois as classes de equivalência ficaram grandes e homogéneas.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e a escolha da métrica de utilidade.



Na primeira medida, Discernibilidade, a generalização teve uma intensidade de 13.9% e granularidade de 20.6%, que indicam uma pequena generalização e boa preservação de detalhe nos QIDs. A entropia é de 12.8% e a discernibilidade é de 72.4%, que revela que, embora exista anonimização, as classes ficaram relativamente distintas. O tamanho médio de classes de equivalência é de 80.4%, o que indica que as classes de equivalência agrupam bem os registos, mas não garantem anonimato forte. Os erros quadráticos são baixos, de 8.9% e 1.6%, mostrando que a aproximação aos valores originais foi muito fiel. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Na segunda medida, *Average Equivalence Class Size* a generalização e granularidade são similares, com 15.6% e 16.7% respetivamente, indicando baixa generalização e razoável preservação de detalhe. A entropia é de 17.6% e a discernibilidade de 26.2%, sinalizando classes menos diversas. O tamanho médio das classes é de 99.9% o que indica que quase todos os registos foram agrupados em grandes classes, o que reforça o anonimato. O erro quadrático por registo é de 6.1% e o erro quadrático por atributo é de 17.6%, mostrando que os registos foram pouco alterados, mas os valores individuais sofreram alguma distorção. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Por último, com a medida *Loss*, a intensidade de generalização é moderada (60%), tal como a granularidade (32.2%), o que demonstra que o foco em minimizar a '*loss*' leva a generalizações mais intensas e perda de detalhe, porém preserva mais variedade nas classes. A entropia de 11.4% é baixa e a discernibilidade é moderada (57.7%),

indicando classes ainda mais distintas, mas com pouca diversidade de valores nos atributos, o que pode prejudicar algumas algumas análises. O tamanho médio das classes é de 96.8% o que indica que quase todos os registos foram agrupados em grandes classes, o que reforça o anonimato. Os erros são mais elevados (28-31%), exceto o de agregação, que é sempre 0%, visto que a prioridade foi manter a utilidade global em vez de minimizar distorções pontuais.

Concluímos:

- A primeira configuração (Discernibilidade) apresenta baixa distorção e boa preservação de dados, ideal para alta utilidade com alguma proteção. (máxima utilidade).

- A segunda configuração (*Average Equivalence Class Size*) garante o anonimato mais forte com classes muito grandes. (máxima privacidade).

- A terceira configuração (*Loss*) aplica generalização mais intensa, o que compromete utilidade individual dos dados, mas mantém a consistência das análises agregadas. (preservação estatística com proteção moderada).

4.3.2.3. Peso de Atributos

O peso de um atributo é a importância atribuída a cada atributo na análise de utilidade. Atributos com pesos maiores têm mais influência na avaliação da qualidade dos dados.

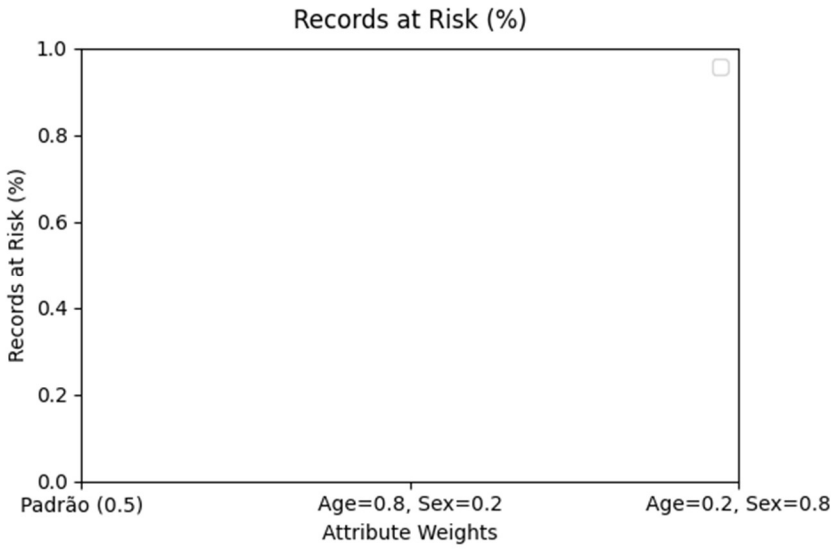
Para o seu estudo, usaram-se os seguintes dados:
- Peso padrão (0.5 para todos os atributos),

- Atributo Age com peso de 0.8 e atributo *Sex* com peso de 0.2, e restantes com peso padrão,

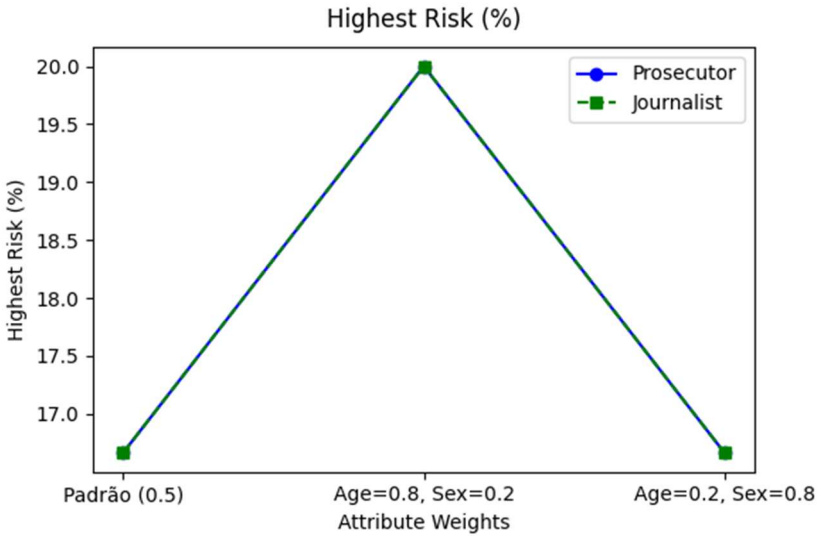
- Atributo *Age* com peso de 0.2 e atributo *Sex* com peso de 0.8, e restantes com peso padrão,

Em que para todos os casos, foi usado um limite de supressão de 100% e a medida de utilidade *Loss* com (k = 5, t = 0.15).

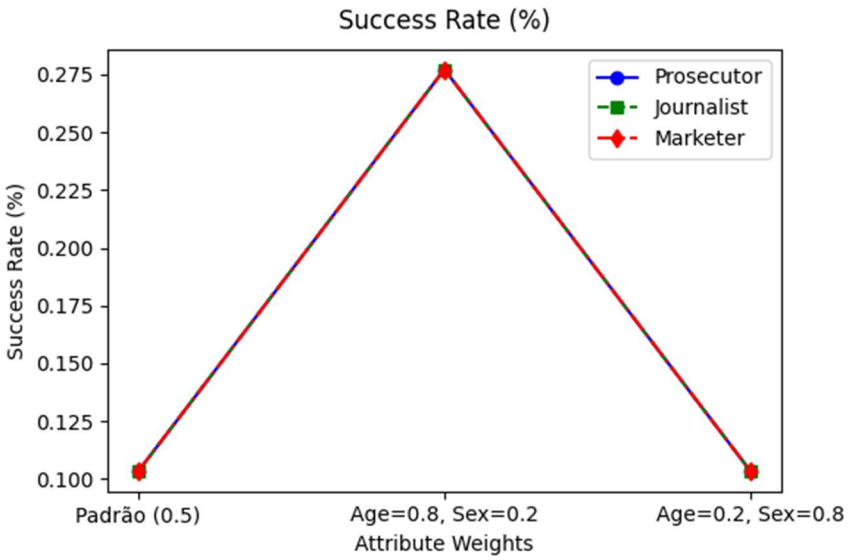
Análise de Riscos



Observando o gráfico acima, é possível concluir que *Records at Risk* mantém-se invariável (0%), indicando que não existem classes unitárias**.**



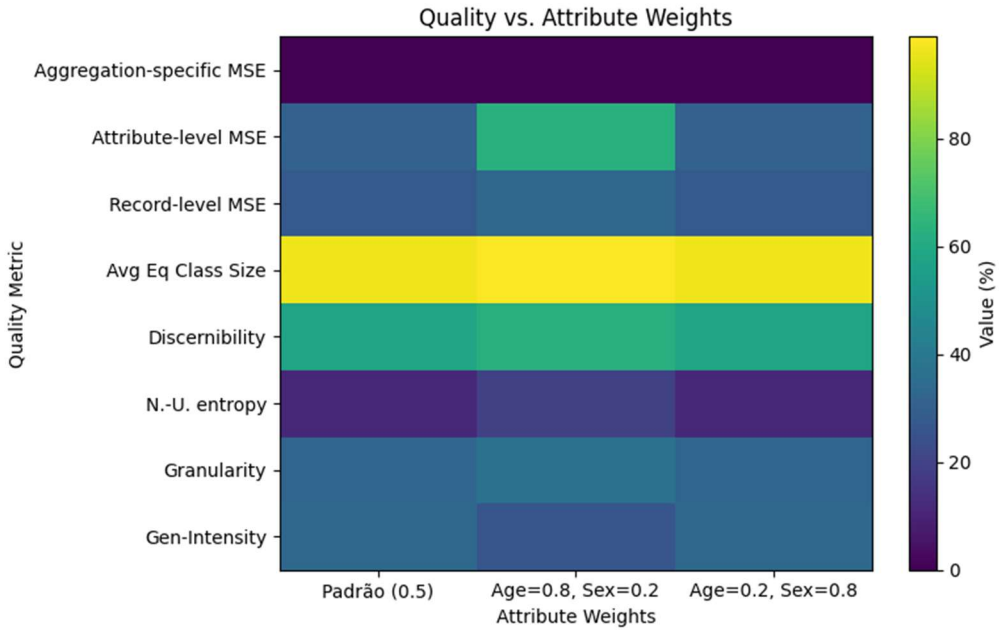
De seguida, é possível observar que o *Highest Risk* para o limite de supressão (*Age*=0.8 e *Sex*=0.2) é de 20% e, o limite padrão e o limite (*Age*=0.2 e *Sex*=0.8) são de menos de 17%. Estas observações sugerem que idade é um atributo mais identificável neste conjunto de dados enquanto as outras duas configurações apresentam menor risco devido ao menor peso no atributo idade. Idade contribui mais para o risco de reidentificação do que o sexo e o risco maior ocorre quando esse atributo tem peso predominante.



Concluindo, a taxa de sucesso para o limite de supressão (*Age*=0.8 e *Sex*=0.2) é de 0.275% e, o limite padrão e o limite (*Age*=0.2 e *Sex*=0.8) são de 0.1%. Podemos concluir que idade é o atributo que mais contribui para o sucesso na reidentificação e que deve ser prioritariamente protegido na anonimização dos dados. A consistência entre os tipos de atacantes indica que os riscos não variam com a motivação do atacante, mas sim com a informatividade dos atributos.

Análise de Utilidade

No gráfico de calor abaixo, é possível verificar a relação entre Qualidade ao Nível do Conjunto de Dados e o peso dos atributos.



Na primeira configuração (padrão 0.5), a generalização teve uma intensidade de 33.2% e granularidade de 32.2%, que indicam uma moderada generalização e uma transformação moderada de detalhe nos QIDs. A entropia é de 11.4% e a discernibilidade é de 57.7%, que revela que, embora exista anonimização, as classes ficaram relativamente distintas com pouca diversidade nos valores de atributos. O tamanho médio de classes de equivalência é de 96.8%, o que indica que as classes de equivalência agrupam bem os registos, e garantem anonimato forte. Os erros quadráticos são moderados, de 28.1% e 31.1%, mostrando que os valores alteraram moderadamente. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Na segunda configuração (*age*=0.8 e *sex*=0.2), a generalização teve uma intensidade de 22.3% e granularidade de 36.8%, que indicam uma moderada generalização e uma perda de detalhe significativa nos QIDs. A entropia é de 19.4% e a discernibilidade é de 61.9%, que revela que, embora exista anonimização, as classes ficaram relativamente distintas com pouca diversidade nos valores de atributos. O tamanho médio de classes de equivalência é de 98.8%, o que indica que as classes de equivalência agrupam bem os registos, e garantem anonimato forte. Os erros quadráticos são moderados/altos, de 33.1% e 62.5%, mostrando que os valores alteraram moderadamente com distorção elevada nos atributos. Por último, o erro quadrático de agregação é de 0%, o que indica que todas as estatísticas de agregação escolhidas foram idênticas, e garante que análises agregadas não perdem exatidão.

Na terceira configuração (*age*=0.2 e *sex*=0.8), a qualidade foi exatamente igual a configuração dois (*age*=0.8 e *sex*=0.2).

Concluímos que a configuração padrão oferece o melhor compromisso entre privacidade e qualidade dos dados detalhados. A escolha ideal dependerá do contexto: se a prioridade for anonimato absoluto, as configurações com maior distorção são mais adequadas e se for necessário equilíbrio entre anonimização e análise granular, a configuração padrão é preferível.

5. Discussão Ética e Regulatória

5.1. Fundamentos Éticos

Segundo o Princípio da Minimização de Dados (*Privacy by Design*/ *Data Minimization*), a anonimização deve seguir o princípio de coletar apenas o mínimo necessário para a finalidade. Muitas vezes, mesmo que os dados possam ser anonimizados, a coleta anterior pode não ter seguido princípios de consentimento informado.

No âmbito da transparência e responsabilidade, interessados deveriam saber, ou poder saber, que os seus dados podem vir a ser anonimizados e compartilhados para uma finalidade específica.

É, também, fundamental reconhecer que “0%” nos modelos de ARX não significa “privacidade absoluta em qualquer cenário futuro”. Em determinadas configurações, existe sempre o conceito de “risco residual”, combinando atributos auxiliares ou mesmo descobrindo novas técnicas de correlação em ataques futuros.

5.2. Conformidade Regulatória

O Regulamento Europeu de Proteção de Dados (GDPR) está em vigor em toda a União Europeia desde 2018 e define “pseudonimização” e “anonimização” como medidas de segurança:

- Artigo 6 (Legalidade do tratamento): Se os dados forem anonimizados de forma irreversível, já não são considerados “dados pessoais” sob o GDPR e, portanto, muitas obrigações, como consentimento explícito, deixam de se aplicar plenamente;

- Artigo 25 (Proteção de Dados desde a Conceção, *Privacy by Design*): A anonimização deve ser considerada desde o início do projeto de coleta/tratamento;

- Artigo 32 (Segurança do Tratamento): Exige que medidas técnicas e organizacionais, como as descritas no relatório, sejam implementadas para proteger os dados.

6. Conclusão

Podemos concluir que, a partir da extensa análise feita, o dataset original apresentava elevado risco de reidentificação e alta vulnerabilidade na amostra analisada. Assim, as técnicas de anonimização aplicadas reduziram fortemente esse risco, onde configurações como (k=5, l=4) ou (k=10, t=0.15) resultaram em 0% de records at risk e taxas de sucesso de ataque muito baixas, mostrando alta eficácia na proteção contra a reidentificação. Apesar disso, existe um *trade-off* entre Privacidade e Utilidade. A configuração (k=5, l =4) garantiu privacidade máxima, mas resultou num *dataset* inutilizável, devido à perda total de informação. Já as configurações (k=10, l=2) ou (k=10, t=0.15) oferecem melhor equilíbrio entre utilidade e privacidade, algo útil para análises com algum risco residual controlado. Outra conclusão retirada deste estudo é que a utilidade dos dados varia conforme os parâmetros e métricas. A escolha de medidas de utilidade (*Loss*, Discernibilidade e Tamanho médio de classes de equivalência) impacta diretamente o resultado da anonimização. O modelo baseado em Discernibilidade mostrou-se o mais equilibrado, com boa preservação de estrutura e baixa distorção. A atribuição de peso aos atributos também afeta os riscos: atributos como idade têm maior poder de reidentificação e devem ser mais protegidos.

Assim, a configuração que traz privacidade extrema é (k=5, l=4), mesmo que sem nenhuma utilidade associada. Configurações como (k=10, l=2), (k=10, t=0.15), (k=5, l = 2 + Loss), ou (k=5, t=0.15 + *Loss*) trazem o equilíbrio ideal entre privacidade e utilidade. As piores configurações, que não protegem bem nem mantêm utilidade, são (k=3, l=2) e (k=3, t=0.15), isto é, configurações com o k mínimo.

Assim, o modelo escolhido para anonimizar os dados foi (*k-anonymity* + *l-diversity*) onde k = 10 e l = 2.

7. Referências

[1] Anon, (n.d.). *Privacy models | ARX - Data Anonymization Tool*. [online] Available at: https://arx.deidentifier.org/overview/privacy-criteria/.

[2] AnyDesk. (2024). *Política de privacidade do AnyDesk*. [online] Available at: https://anydesk.com/pt/privacidade.

[3] Linkedin.com. (2024). *You're managing data access. What are the most user-friendly data governance tools to use?* [online] Available at: https://www.linkedin.com/advice/3/youre-managing-data-access-what-most-user-friendly-uffxe.

[4] OpenAI (2025). *ChatGPT*. [online] chatgpt.com. Available at: https://chatgpt.com.

[5] PRIVACIDADE E PROTEÇÃO DE DADOS PESSOAIS PRIVACY AND PERSONAL DATA PROTECTION. (n.d.). Available at: https://cetic.br/media/docs/publicacoes/2/20240901120340/privacidade-e-protecao-de-dados-2023.pdf.

[6] Richman, A. (2023). *The L Diversity Data Anonymization Model: Extending K Anonymity*. [online] www.k2view.com. Available at: https://www.k2view.com/blog/l-diversity/.

[7] S.A, P.I. (n.d.). *Dicionário Priberam, Dicionário Online de Português Contemporâneo*. [online] Dicionário Priberam. Available at: https://dicionario.priberam.org.

[8] Trotino, G. (2024). *What is K Anonymity?* [online] www.k2view.com. Available at: https://www.k2view.com/blog/what-is-k-anonymity.

[9] Wikipedia Contributors (2022). *t-closeness*. Wikipedia.