FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

ANONYMIZATION OF DATASETS WITH PRIVACY, UTILITY AND RISK ANALYSIS

Segurança e Privacidade

Síntese

O objetivo principal deste trabalho é fazer uma análise detalhada de um processo de anonimização de um dataset.

Carlos Matos 2020245868
Mariana Magueijo 2020246886

Índice

Caracterização do Dataset

Descrição e objetivo

O dataset é composto por dados clínicos de diversos voluntários que providenciaram amostras de sémen, de modo a determinar a sua qualidade. Decidimos adaptar este dataset para um banco de doação de esperma, no qual iremos fazer, mais à frente, alterações de modo a ir de encontro ao nosso objetivo.
Para além da infertilidade, homossexualidade e mulheres sem companheiro que decidem ser mães, também homens com distúrbios genéticos ou doenças contagiosas decidem recorrer a este tratamento. Precisamente por estas razões, é crucial garantir que não existam anomalias genéticas ou hereditárias, minimizando assim a possibilidade de malformações e portação de doenças no feto. Para tal, é importante criar modelos de previsão de doenças, tendo em conta o histórico e genética do dador, e anonimizar estes dados de modo a não comprometer a privacidade dos indivíduos.

Tratamento do dataset

Para ir de encontro ao nosso objetivo, achámos que seria pertinente acrescentar e retirar certas colunas. Uma vez que se trata de uma doação, é uma mais valia que os recetores da mesma saibam certas características físicas do dador, tais como a raça, cor de olhos, altura e peso, atributos estes que adicionamos ao *dataset*. Consideramos que características psicológicas, tais como o percurso escolar, área de trabalho e orientação sexual, e ideais, como as preocupações no mundo atual, ajudaria a formar uma imagem mais concreta, realista e tendenciosa. Não podíamos deixar de parte a avaliação médica, especificamente os resultados da análise psicológica, análise ao sangue, qualidade do esperma e genéticos, que determinam se o dador é saudável e apto para prosseguir com o procedimento.
Consideramos que o número de horas sentado por dia e a quantidade de vezes que teve febre no último ano não seriam dados relevantes para o trabalho proposto.

Atributos

Este *dataset* contém 3 tipos de informação: objetiva, de examinação e subjetiva.

- A informação objetiva refere-se a dados concretos e observáveis, que não dependem da opinião ou interpretação pessoal e que são inerentes ao indivíduo.
- A informação subjetiva, por outro lado, abrange os dados influenciados pela interpretação pessoal, sentimentos, opiniões individuais e acontecimentos partilhados pelo indivíduo.
- As informações de examinação têm de ser avaliadas e verificadas por profissionais.

Todas as informações de examinação, sendo elas as doenças em criança, intervenção cirúrgica, análise ao sangue, qualidade do esperma, análise genética e análise psicológica são respostas simples

Atributo	Possíveis opções					
Número de utente	-					
Idade	>18					
Raça	white non-Hispanic	asian	hispanic	caucasian	black	
Cor dos olhos	black	brown	green	blues		
Altura	[140,230] cm					
Peso	[40,150] kg					
Escolaridade	12 grade	banchelor	master	phD	other	
Emprego	-					
Orientação sexual	heterosexual	homosexual	bisexual	other		
Consumo de álcool	once a week	hardly ever or never'	several times a week	several times a day	every day	
Hábito de fumar	occasional	daily	never			
Qualidade do esperma	average	low	high			
Resultados de amostras de sangue	-					
Acidentes ou traumas graves	-					
Avaliação psicológica	-					
Resultados genéticos	-					
Doenças em criança	-					
Intervenção cirúrgica	-					
Preocupações	global warming	homophobia	machism	xenophobia	animal abuse	racism

Primeiramente começamos por retirar os atributos que não faziam sentido estar na tabela e renomear os achávamos que não estavam explícitos, através das funções *drop* e *rename* do *python*, respetivamente. De seguida, acrescentamos colunas que acreditamos que fazem sentido no nosso *dataset*, e para isso criamos variáveis para serem escolhidas aleatoriamente e atribuídas às linhas de cada coluna. Após termos a tabela inicial criada, contendo todas as colunas desejadas, acrescentamos 1400 linhas de modo ao *dataset* ficar mais rico em informação, ficando assim com 1500 linhas e 19 colunas. Esta informação em certos casos não foi acrescentada totalmente de forma aleatória, como por exemplo, fizemos uma pesquisa e obtivemos que mais ou menos 80% das pessoas são heterossexuais, e por isso, para manter a credibilidade do *dataset* aplicamos essa percentagem nele.

PPDP

É importante conseguirmos divulgar dados publicamente ou para terceiros para a sua análise sem divulgar o proprietário de dados confidenciais, mesmo com a presença de qualquer conhecimento prévio do adversário, obtido por outras fontes. Com o intuito de preservar a privacidade, existe o PPDP (*privacy-preserving data publishing*) que pode ser visto como uma resposta técnica para complementar as políticas de privacidade, diretrizes e acordos.

- **Identificadores**: atributos que identificam explicitamente e exclusivamente um indivíduo. Neste *dataset* temos o número de utente de saúde.
- **Quasi-identifyings**: atributo que não identificam explicitamente um usuário, mas pode ser combinado com outros para a sua re-identificação. São eles a idade, peso, altura, cor de olhos e raça.
- **Sensíveis**: atributos confidenciais específicos do indivíduo que não podem ser divulgados publicamente. Pode também estar vinculado para identificar indivíduos. É considerado sensível as doenças em criança, intervenção cirúrgica, os acidentes e traumas, a frequência com que fuma ou bebe álcool, os resultados das análises ao sangue, genéticas, psicológica e de qualidade do esperma, a orientação sexual e as preocupações/opiniões.
- **Não sensíveis**: contém todos os atributos que não se enquadram nas três categorias anteriores: o emprego e o percurso académico entram nesta categoria.

Riscos de privacidade

Com o aumento da quantidade de informações pessoais a serem guardadas e processadas, é natural que também aumente a preocupação em relação à proteção desses dados.
Apesar dos dados pessoais serem atualmente utilizados para inúmeros fins, os autores dos mesmos têm pouco ou nenhum controlo sobre quem os usa, podendo levantar questões de segurança e privacidade.
Reconhecendo que a informação contida no *dataset* escolhido é sensível, uma vez que contém dados médicos e pessoais, é necessário propor mecanismos que maximizem a proteção de dados, garantindo assim a confidencialidade e integridade dos clientes.
A re-identificação refere-se ao processo de correspondência de dados não anonimados com fontes externas de informação, com o objetivo de re-identificar indivíduos cujas identidades deveriam ser protegidas pelo processo de anonimização.

Existem inúmeros riscos de violação de privacidade, entre eles:

- Identificação de pacientes: quando cruzados dados como morada, idade, raça, cor de olhos, peso e altura ou número de cartão de utente é facilmente divulgada a identificação da pessoa.
- Divulgação de informações sensíveis: Se um conjunto de dados médicos for comprometido por meio de uma violação de dados, informações pessoais de pacientes, como histórico médico, diagnósticos, procedimentos e medicamentos, podem ser expostas.
- Discriminação:
    - no seguro: se uma companhia de seguros de saúde tiver acesso a informação médica relevante, tal como o histórico de doenças, pode usá-la para aumentar ou diminuir o preço do serviço ou negar cobertura.

    - nos empréstimos: Os bancos podem utilizar as mesmas informações para tomar decisões de empréstimo.
    - na habitação: Se um proprietário de um imóvel for preconceituoso em relação a raça ou orientação sexual, e tiver acesso a estes dados, podem negar o arrendamento ou compra de habitação a um indivíduo.
    - social
    - no emprego
- Marketing: as empresas podem usar informações médicas para escolher anúncios ou campanhas de marketing com base nos problemas e doenças dos clientes, criando assim um público-alvo.

Para medir o risco de re-identificação de alguém é preciso ter uma visão sobre o modelo de ataque e a qualidade dos dados. A qualidade dos dados refere-se ao grau com que estes atendem ao uso pretendido dos requisitos, como precisão, integridade, consistência e oportunidade. Assumem-se condições ideais sobre a qualidade dos dados do *dataset* e sobre a informação que o adversário pode usar para atacar o mesmo. Dito isto, é possível que leve a uma visão irrealista, pois quanto mais qualidade têm os dados, mais provável é de o adversário conseguir re-identificar uma pessoa, resultando em estimativas de risco conservadora.
Existem também 3 tipos de modelos de ataque a considerar, o que é usado para medir o sucesso do processo de anonimização. Ter uma forma para medir este risco é importante para decidir se este está muito alto e quanto de anonimização é preciso fazer para o reduzir. São eles:

- **Prosecutor:** o alvo é um indivíduo específico e o adversário sabe se este se encontra no dataset.
- **Journalist:** o atacante não tem nenhum alvo específico, fazendo uma escolha random. A re-identificação de qualquer pessoa irá atingir o objetivo.
- **Marketer:** O objetivo do adversário é ter tantos alvos para a sua re-identificação quanto possível, considerando-se assim um ataque bem sucedido.

Prosecutor (%)			Journalist (%)			Marketer (%)
Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate
100	100	100	100	100	99.96035	99.96035

Através desta tabela, é possível inferir que o nosso dataset sem qualquer modificação não se encontra protegido contra nenhuma ameaça, sendo muito fácil a identificação de indivíduos.

Análise de potênciais Quasi-identifiers

Após a importação e a separação de cada atributo pelos grupos definidos pelo PPDP para o ARX, diferentes QIDs são sugeridos para o nosso *dataset*, dependendo dos valores de distinção e separação apresentados por cada atributo. De modo a melhorar o processo de anonimização de dados, deve-se escolher QIDs com valores elevados nos campos referidos anteriormente.
Porém, durante a escolha dos QIDs, é importante ter em consideração, não só os valores de distinção e separação, mas também o tipo de atributo sugerido. Isto é, atributos identificadores não devem ser selecionados, visto que podem causar a re-identificação de um indivíduo. Para além destes, atributos sensíveis também não devem ser considerados, pois contém dados confidenciais.

Para medir o risco de re-identificação é necessário determinar a distinção e separação dos potenciais *quasi-identifiers*, que se traduz no grau no qual as variáveis fazem os registos distintos e no grau em que a combinação de variáveis separam os registos, respetivamente.

$$Distinção\ (\%) = \frac{número\ de\ valores\ distintos\ de\ um\ atributo}{número\ total\ de\ registos\ de\ um\ atributo} \times 100$$

$$Separação\ (\%) = \frac{número\ de\ tuplos\ onde\ não\ há\ repetição\ de\ valores}{número\ de\ tuplos\ que\ é\ possível\ formar} \times 100$$

Para escolher o melhor QID é feita a combinação entre os atributos classificados com *quasi-identifiers*, assim como a apresentação dos respetivos valores de distribuição e separação. De seguida encontra-se a tabela com os dados referidos anteriormente.

Atributos	Distinção (%)	Separação (%)
Idade	0.67	89.99
Raça	0.33	80.03
Cor dos olhos	0.27	74.97
Altura	6.7	98.91
Peso	7.4	99.09
Idade, Raça	3.33	97.99
Idade, Cor dos olhos	2.67	97.49
Idade, Altura	47.6	99.88
Idade, Peso	54.4	99.91
Raça, Cor dos olhos	1.33	94.99
Raça, Altura	29.13	99.78
Raça, Peso	34.08	99.81
Cor dos olhos, Altura	23.9	99.73
Cor dos olhos, Peso	28.8	99.77
Altura, peso	93.67	99.99
Idade, Raça, Cor dos olhos	13.33	99.5
Idade, Raça, Altura	85.87	99.98
Idade, Raça, Peso	86.73	99.98
Idade, Cor dos olhos, Altura	80.53	99.97
Idade, Cor dos olhos, Peso	84.8	99.98
Idade, Altura, Peso	98.87	99.99
Raça, Cor dos olhos, Altura	69.13	99.95
Raça, Cor dos olhos, Peso	71.73	99.95
Raça, Altura, Peso	98.87	99.99
Cor dos olhos, Altura, Peso	98.33	99.99
Idade, Raça, Cor dos olhos, Altura	95.07	99.99
Idade, Raça, Cor dos olhos, Peso	96.6	99.99
Idade, Raça, Altura, Peso	99.8	99.99
Idade, Cor dos olhos, Altura, Peso	99.67	99.99
Raça, Cor dos olhos, Altura, Peso	99.67	99.99
Idade, Raça, Cor dos olhos, Altura, Peso	99.93	99.99

Através dos valores de distinção dos *quasi-identifiers*, que são compostos apenas por um atributo, é visível que estes são muito reduzidos. Porém isso seria de esperar, uma vez que temos um *dataset* com 1500 indivíduos e os QIDs apresentam poucos valores únicos: idade tem 10, raça tem 5, cor dos olhos tem 4, altura tem 91 e peso tem 111.

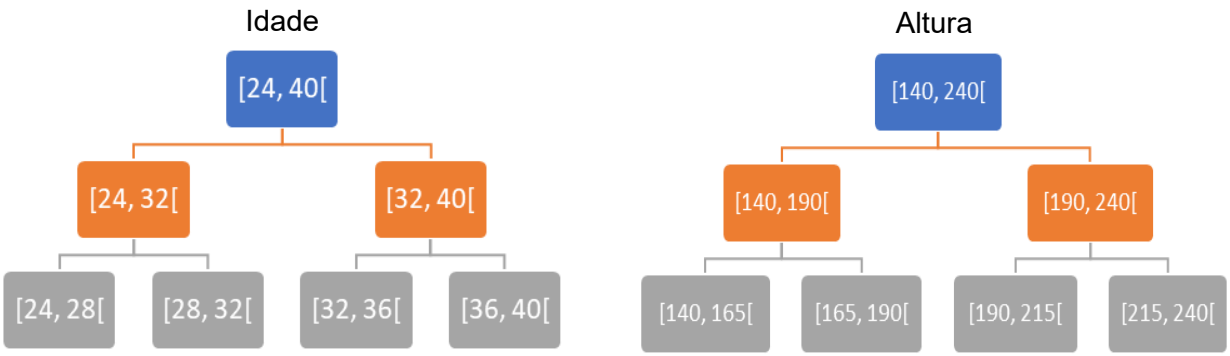
Coding models

ARX implementa métodos que oferecem meios dinâmicos para equilibrar os riscos de privacidade com a utilidade dos dados. Os dados são transformados com modelos de codificação, em particular generalização e supressão de valores de atributos, para garantir que cumprem os requisitos de privacidade especificados.

Sendo que a nossa visão para este *dataset* é auxiliar o desenvolvimento de modelos que permitam avaliar as condições ideias para a doação de esperma, existem atributos que não são importantes nem têm utilidade para este fim, não obtendo nenhuma informação adicional relevante utilizando os mesmos, tal como o número de utente de saúde (identificador) e a educação e profissão. Deste modo, aplicámos o *coding model* de supressão (*column-wise*) a estas 3 colunas, que removem/suprimem todos os valores de um atributo numa tabela.
Atributos considerados sensíveis não podem sofrer qualquer tipo de codificação, uma vez que é crucial para atingir o objetivo pretendido. Dito isto, os resultados das análises ao sangue, análise psicológica, qualidade do esperma, análise genética, orientação sexual e preocupações devem permanecer intactos.
Resta-nos as colunas da idade, altura, peso, raça e cor de olhos que irão ser codificadas com generalização. Para as duas primeiras, iremos aplicar uma generalização por hierarquia baseada numa gama de valores, enquanto para as outras, é pela frequência.

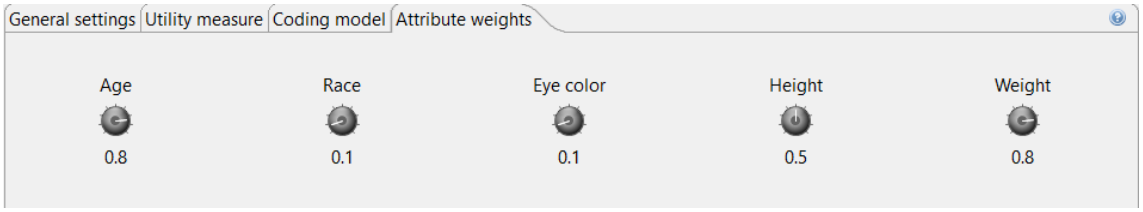
Atributos	Valor mínimo	Valor máximo
Idade	27	36
Altura	140	230

Tendo em conta estes valores, optamos por criar 4 intervalos, ou seja, 3 níveis de hierarquia para a idade e a altura. Já para o peso, visto que a gama de valores é mais ampla, criamos 8 intervalos (4 níveis de hierarquia). A escolha dos intervalos foi feita da seguinte forma:



Após isto, decidimos alterar alguns parâmetros nos coding models. Na aba General Settings, aumentamos o supression limit para 100%, deste modo será possível utilizar todos os níveis de hierarquia que foram definidos. Na aba Coding Model, colocamos também a 100% o valor de generalização, permitindo o haver um maior número de combinações dos níveis de hierarquia.

Atribuímos pesos diferentes aos diferentes atributos pois consideramos que a raça e a cor de olhos não é, de todo, relevante para o propósito do *dataset*. Por outro lado, a idade e o peso, em comparação com a altura, pode influenciar a qualidade da doação assim como possíveis doenças e, por consequente, influenciar o objetivo pretendido.



Feita esta hierarquização, não observamos grandes mudanças no gráfico de distribuição de riscos. Uma possível explicação para isto, é a geração de dados aleatórios na criação das novas colunas, fazendo com que os dados fiquem distribuídos quase uniformemente.

K-Anonymity

Quando uma pessoa deseja compartilhar um conjunto de dados, muitas das vezes removem dados confidenciais e sensíveis. No entanto, a simples remoção de alguns ou de todos os atributos sensíveis não garante a privacidade, pois vários conjuntos de dados podem ser cruzados por invasores para desanonimizar as informações, tendo sido criado o K-Anonymity. Este modelo limita a capacidade de vinculação de informações dos quasi-identifiers, de modo que não seja possível identificar alguém através deles.

O K refere-se ao número de vezes que uma combinação de valores aparece num *dataset*. Portanto, se K=2, os pontos de dados foram mascarados para que existam pelo menos dois conjuntos de cada combinação de dados. Isto significa, por exemplo, que se um determinado conjunto de dados contiver código postal e idades de um grupo de pessoas, estes atributos precisariam ser anonimizados para que cada par de código postal e idade apareça pelo menos duas vezes.

K-Anonymity funciona com base no princípio de que, se você combinar dados com atributos semelhantes, poderá ocultar informações de identificação sobre qualquer indivíduo que contribua para esses dados. É basicamente a capacidade de desaparecer no meio de uma multidão – uma vez que um atributo de dados confidenciais mascarado

usando K Anonymity pode, na verdade, corresponder a qualquer indivíduo no conjunto de dados agrupados.

L-Diversity

Este modelo procura estender as classes de equivalência que criamos no modelo anterior por generalização e mascaramento dos QIDs também para os atributos sensíveis. O princípio da L-Diversity exige que, em cada grupo QIDs, existam pelo menos L valores diferentes para os atributos sensíveis.

No caso de múltiplos atributos sensíveis, a L-Diversity deve ser satisfeita para cada um deles. Assim, à medida que o número de atributos sensíveis aumenta, será necessário mais mascaramento dos Quasi-identifiers para alcançar a diversidade necessária.

T-Closeness

A distribuição dos valores sensíveis de cada grupo deve ser próxima da correspondente distribuição original. Isto é, se num dataset temos como atributo sensível a condição médica, e a distribuição dos valores da mesma fossem 20% asma, 40% hipertensão e 40% diabetes, teríamos de garantir que, depois de aplicados os dois modelos anteriores, estas percentagens permaneceriam o mais idênticas possível.

Differential Privacy

Este modelo introduz um ruído aleatório nos dados antes da divulgação, tornando as respostas aproximadas, em vez de precisas, o que dificulta a identificação de indivíduos.

ε: controla a quantidade máxima de ruído que pode ser introduzida. Quanto maior o valor de ε, mais ruído é permitido, mas menos proteção de privacidade é oferecida. A escolha de ε é um compromisso crítico entre privacidade e utilidade.

δ: controla o risco de divulgação acidental. É usado para definir um limite para a probabilidade de que uma consulta específica possa divulgar informações não autorizada. Um valor δ muito baixo torna a proteção de privacidade mais rigorosa, mas pode tornar as consultas mais limitadas.

Aplicação de modelos de privacidade

Uma vez que o ARX não permite a aplicação do modelo K-Anonymity sem a definição de modelos de privacidade para atributos sensíveis, começamos por aplicar o L-Diversity.

Escolha do L para o L-Diversity

Atributo	Número de valores distintos
Doenças em criança	2
Acidentes ou traumas graves	2
Intervenções cirúrgicas	2
Consumo frequente de álcool	5

Hábitos tabagísticos	3
Resultados de amostras de sangue	2
Qualidade do esperma	3
Resultados genéticos	2
Avaliação psicológica	2
Orientação sexual	20
Preocupações	687

Podemos observar que existe uma grande diferença entre o número de valores distintos da orientação sexual e das preocupações para o resto dos atributos. Dito isto, decidimos aplicar ao L os respetivos valores distintos e fomos mudando os valores do L da orientação sexual e das preocupações para atingir o melhor L-Diversity possível. Como, depois destes atributos, o maior valor corresponde ao consumo frequente de álcool (5) começamos por experimentar este valor.

	Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[4, 5, 4, 1, 10]	0	0.38023	0.26667	0	0.38023	0.26667	0.26667	0
[4, 5, 4, 0, 10]	0	7.69231	5.45455	0	7.69231	5.45455	5.45455	1335
[3, 4, 5, 0, 10]	0	8.33333	5.98291	0	8.33333	5.98291	5.98291	1383
[2, 5, 4, 0, 10]	0	7.69231	7.69231	0	7.69231	7.69231	7.69231	1487
[1, 5, 4, 0, 10]	0	8.33333	8.33333	0	8.33333	8.33333	8.33333	1488
Para L em preocupações = 5 a 10 e L em orientação sexual = 5								

	Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[4, 5, 4, 1, 10]	0	0.38023	0.26667	0	0.38023	0.26667	0.26667	0
[4, 5, 4, 0, 10]	0	4.34783	4.34783	0	4.34783	4.34783	4.34783	1477
Para L em preocupações = 20 e L em orientação sexual = 5								

	Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[4, 5, 4, 1, 10]	0	0.38023	0.26667	0	0.38023	0.26667	0.26667	0
[4, 5, 4, 0, 10]	0	8.33333	5.72139	0	8.33333	5.72139	5.72139	294
[3, 4, 5, 0, 10]	0	10	6.27376	0	10	6.27376	6.27376	448

[2, 5, 4, 0, 10]	0	16.66667	9.33661	0	16.66667	9.33661	9.33661	1093
[1, 5, 4, 0, 10]	0	20	10.06944	0	20	10.06944	10.06944	1212
[0, 5, 4, 0, 10	0	14.28571	14.28571	0	14.28571	14.28571	14.28571	1493
Para L em preocupações = 2 e L em orientação sexual = 2								

Atendendo aos valores acima, apesar de com a hierarquia [4, 5, 4, 1, 10] os resultados serem os mesmos para a primeira tabela, onde L em preocupações e em orientação sexual é 5, ou para a terceira tabela, onde L em preocupações e orientação sexual é 2, decidimos optar pela segunda opção uma vez que apresenta menos registos apagados apesar de ter uma maior taxa de risco.

Atributo	Valor de L
Doenças em criança	2
Acidentes ou traumas graves	2
Intervenções cirúrgicas	2
Consumo frequente de álcool	5
Hábitos tabagísticos	3
Resultados de amostras de sangue	2
Qualidade do esperma	3
Resultados genéticos	2
Avaliação psicológica	2
Orientação sexual	2
Preocupações	2

Escolha do K para K-Anonymity

Para escolher o valor inicial de K, já tendo o valor de L decidimos ir ver qual era o tamanho mínimo de uma classe após aplicado o modelo L-Diversity. Como este é 263, usando um valor de K menor ou igual a percentagem de risco de qualquer modelo de ataque permaneceria igual, isto porque o K-Anonymity tenta criar classes de tamanho igual ou superior a K.

Summary statistics | Distribution | Contingency | Class sizes | Properties | Classification models

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	375 (25%)	375 (25%)
Maximal class size	425 (28.33333%)	425 (28.33333%)
Minimal class size	263 (17.53333%)	263 (17.53333%)
Suppressed records	0 (0%)	0
Number of classes	4	4
Number of records	1500	1500

Com base neste valor, realizamos vários testes para encontrar o melhor valor para k, tendo em conta o número de registos perdidos e o risco dos modelos de ataque.

		Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Valor de k	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[4, 5, 4, 1, 10]	263	0	0.38023	0.26667	0	0.38023	0.26667	0.26667	0
[4, 5, 4, 1, 10]	363	0	0.24938	0.16168	0	0.24938	0.16168	0.16168	263
[4, 5, 4, 3, 10]	363	0	0.08084	0.08084	0	0.08084	0.08084	0.08084	263
[4, 5, 4, 3, 10]	463	0	0.08084	0.08084	0	0.08084	0.08084	0.08084	263
[4, 5, 4, 2, 10]	463	0	0.11962	0.11962	0	0.11962	0.11962	0.11962	664
[4, 5, 4, 1, 10]	563	0	0.24938	0.24252	0	0.24938	0.24252	0.24252	263
[4, 5, 4, 2, 10]	563	0	0.24938	0.16168	0	0.24938	0.16168	0.16168	263
[4, 5, 4, 3, 10]	563	0	0.08084	0.08084	0	0.08084	0.08084	0.08084	263

Após a recolha de dados, podemos observar que, para k = 263, obtemos exatamente o mesmo resultado, porém para K = 463 com o nível de hierarquia [4, 5, 4, 3, 10], apesar de haver mais perdas de registos, o risco diminuiu. Por isso decidimos por optar atribuir a K o valor 463.

K-Anonymity + L-Diversity + T-Closeness

O próximo modelo a aplicar foi o T-Closeness, e experimentamos vários valores para T, porém nenhum valor fez alterações positivas no nosso dataset, pelo contrário, aumentou mais o risco. Posto isto procuramos outro algoritmo para substituir o T-Closeness, e decidimos aplicar o Differential Privacy.

		Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Valor de t	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[4, 5, 4, 3, 10]	[0.2, 1]	0	0.08084	0.08084	0	0.08084	0.08084	0.08084	263
[4, 5, 3, 3, 10]	[0.3, 1]	0	0.11013	0.11013	0	0.11013	0.11013	0.11013	592

K-Anonymity + L-Diversity + Differential Privacy

ε = 2		Prosecutor (%)			Journalist (%)			Marketer (%)	Registos apagados
Níveis de hierarquia	Valor de δ	Records at risk	Highest risk	Success rate	Records at risk	Highest risk	Success rate	Success rate	
[2, 5, 4, 4, 10]	0.001	0	0.20121	0.17525	0	0.16584	0.14609	0.14592	112

[2, 5, 4, 4, 10]	0.0001	0	0.20367	0.17986	0	0.16584	0.14613	0.14613	118
[2, 5, 4, 4, 10]	0.00001	0	0.19841	0.17513	0	0.16484	0.14612	0.14612	103
[2, 5, 4, 4, 10]	0.000001	0	0.20704	0.17762	0	0.16584	0.14609	0.14569	115

Analisando os resultados acima, concluímos que o valor ideal para δ seria 0.00001 uma vez que é o que apresenta o menor risco e menor número de registos suprimidos. Experimentamos também alterar o valor de ε e manter o de δ a 0.00001, mas não obtivemos valores relevantes quando comparados com os anteriores.

Conclusão

Com este trabalho foi possível perceber como alcançar a privacidade e segurança de um *dataset*, através de processos de anonimização.
Começamos por definir um objetivo para este trabalho, que seria uma plataforma de doação de esperma e análise de resultados. Para isso procuramos um *dataset* que fosse de encontro às nossas necessidades, mas rapidamente percebemos que existiam atributos que eram necessários não estavam presentes assim como havia alguns que não fazia sentido estarem presentes. Deste modo, fizemos o tratamento do *dataset*, adicionando também linhas para podermos ter uma maior quantidade de dados. Para se tratar de um *dataset* real, decidimos aplicar diferentes percentagens quando se tratava de adicionar valores, como por exemplo a orientação sexual: é mais provável haver heterossexuais do que homossexuais, bissexuais, etc.
Para a próxima fase aplicámos diferentes modelos de privacidade com diversos valores, chegando à conclusão que o mais benéfico, com 103 registos perdidos (6,86%) e 0.19% de risco de privacidade, com o K-Anonymity, L-Diversity e Differential Privacy juntos. Apostamos em comprometer um bocado a taxa de risco para a diminuição significativa dos registos suprimidos.
Concluindo, para o nosso trabalho, atribuímos a L valores entre 2 a 5, a K 463, a ε 2 e a δ 0.00001.

Referências

https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

https://en.wikipedia.org/wiki/Differential_privacy

https://www.k2view.com/blog/l-diversity/

https://www.linkedin.com/learning/privacy-governance-and-compliance-data-sharing/k-anonymity-vs-l-diversity

https://www.dataknobs.com/privacy/k-anonymity-vs-l-diversity.html

https://ilaydabeyreli.wixsite.com/website/post/k-anonymity-l-diversity-t-closeness