Anonymization of a Dataset with Utility and Risk Analysis

João Sousa - up202205238
Rui Santos - up202109728

CC2009: Segurança e Privacidade
Prof.º Manuel Correia
Prof.º João Vilela
Prof.º Henrique Faria
Maio de 2024

Índice

Introdução

O objetivo deste projeto é realizar a anonimização de um *dataset* de grande escala, utilizando como ferramenta auxiliar o ARX . Durante o processo de anonimização vamos descrever os diferentes passos que realizamos para obter o respetivo modelo final , incluindo : classificação dos atributos do *dataset* (indicando se são Identificadores , Quase-Identificadores , sensíveis ou insensíveis) , descrição dos riscos de privacidade que estão associados ao *dataset* original e por fim vamos aplicar e descrever os 2 modelos de anominação que concebemos sobre o *dataset* original .

Classificação dos Atributos

Contexto

Supondo que queremos uma base de dados anonimizada, que permita analisar a **faixa etária (Age)** das pessoas que tem um certo :

- **nível de educação** (Education)
- **sexo** (Sex)
- **estado civil** (Marital-status)
- **classe de trabalho** (Workclass)
- **ocupação** (Ocupation)
- **classe de salário** (Salary-class)

Os restantes atributos que possam ser fornecidas não interessam, neste contexto que definimos

Identificadores e Quase-Identificadores

Para a classificação de Identificadores e Quase-Identificadores, começamos por assumir que todos os atributos presentes no *dataset* são QID's. Utilizamos o separador *"Analyze risk"* e verificamos o valor da distinção e separação para cada atributo individual. Para além de ter em conta o contexto também tivemos em conta que os atributos que tem maior valor de distinção e separação tem maior probabilidade de ser QID's . Como tal , fizemos a seguinte classificação:

Atributo	QID?	Distinção (%)	Separação (%)
Sex	✓	0.03511	41.24421
Age	✓	0.73723	94.39219
Race	X	0.08777	21.50748

Marital status	✓	0.10532	56.8755
Education	✓	0.28085	82.05663
Native-Country	X	0.70212	15.24116
Workclass	✓	0.12287	57.92625
Ocupation	✓	0.22819	89.04712
Salary-class	X	0.03511	44.80915

É importante referir que:

- Nenhum atributo foi classificado como **identificador,** visto que nenhum permiti identificar explicitamente um individuo.
- Os valores de **distinção,** de forma geral , são muito baixos. Isto indica-nos que o dataset não tem grande variedade de valores.
    - **Distinção de um atributo** = n º de valores distintos (desse atributo) / n º de registos (tuplos)
        - Quanto maior for a distinção de valores de um certo atributo presente num dado dataset , maior vai ser o seu valor de distinção
- Os valores de **Separação,** de forma geral, são valores mais elevados e dispersos. Tendo isto em conta baseamo-nos mais nestes valores e no contexto para classificar os atributos como QID.
    - **Separação de um atributo** = n º de pares que podem ser separadas por esse atributo / n º de pares de tuplos distintos
        - Quanto maior for o número de pares que podem ser separados por um atributo, maior vai ser o seu valor de separação
- Não consideramos como QID os atributos **Race** e **Native-Country** por causa do contexto e por os valores baixos de distinção e separação. O atributo **Salary-class** não é QID porque consideramos que é um atributo sensível.

Sensíveis e não sensíveis

Para a classificação de atributos sensíveis e não sensíveis, consideramos todos os atributos que não são QID's ou Identificadores. A classificação feita foi a seguinte:

Race	Insensível
Native-Country	Insensível
Salary-class	Sensível

Apenas consideramos o atributo **Salary-class ,** como sensível porque consideramos que é o único que deve ser considerado privado , e não deve ser divulgado publicamente.

Analise dos riscos de privacidade sobre o *dataset* original

Para analisar os riscos de privacidade sobre o ***dataset* original** , vamos ter em conta os modelos de ataque: Prosecutor, Journalist e Marketer. É importante referir que o *dataset* original não tem qualquer tipo de anonimização associado, garantido assim uma utilidade dos dados máxima.

Measure	Value [%]
Lowest prosecutor risk	4%
Records affected by lowest risk	0.43921%
Average prosecutor risk	69.81729%
Highest prosecutor risk	100%
Records affected by highest risk	55.49895%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	69.81729%
Sample uniques	55.49895%
Population uniques	1.33524%
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figura 1 - Vista geral dos riscos do Dataset Original

Ataque Prosecutor

Neste ataque o respetivo **atacante** foca-se em apenas um individuo especifico, supondo previamente que esse individuo esta presente no *dataset* . Observando a *Figura 2* **,** *podemos retirar as seguintes conclusões:*

- A percentagem de **registos em risco** é aproximadamente 85%
- O **maior risco de re-identificação** , para todas as **classes de equivalência** , é de 100%.
- A **taxa de sucesso** associado a um ataque deste tipo é de aproximadamente 70%

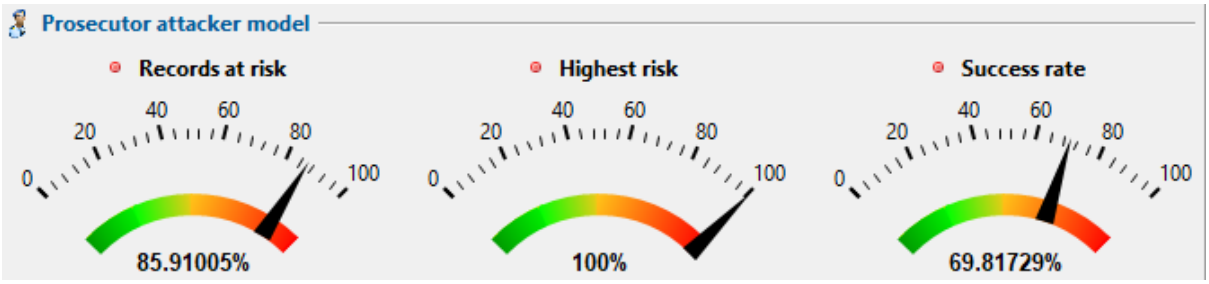

Figura 2 - Risco Prosecutor do dataset original

Ataque Journalist

Neste ataque o respetivo **atacante** foca-se em num individuo aleatório, que pode ou não estar presente no dataset.Assume que existe um dataset de **identificação** . Observando a Figura 3 , podemos retirar as seguintes conclusões:

- A percentagem de **registos em risco** é aproximadamente 85%
- O **maior risco** de re-identificação, para todas as **classes de equivalência**, é de 100% .
- A **taxa de sucesso** associado a um ataque deste tipo é de aproximadamente 69%

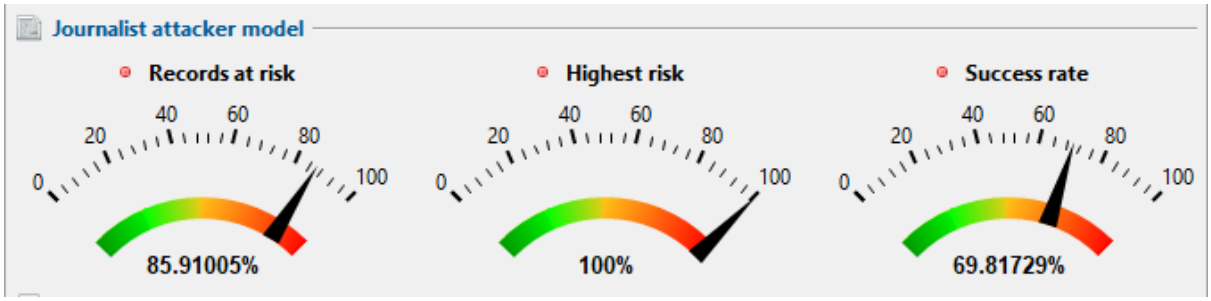

Figura 3 - Risco Attacker do dataset original

Ataque Marketer

Neste ataque o respetivo **atacante** foca-se em tantos indivíduos quanto possível, sendo esse ataque sucedido se uma grande porção dos registos pode ser re-identificado. Observando a Figura 4 , podemos retirar as seguintes conclusões:

- A **taxa de sucesso** associada a um ataque deste tipo é aproximadamente 70%

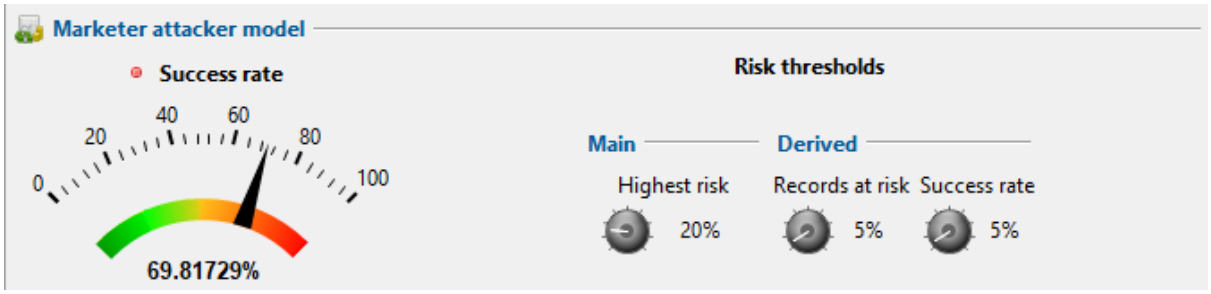

Figura 4 - Risco Marketer do dataset original

Análise dos Modelos

Modelo 1

A ideia principal que definimos ao conceber este modelo consiste em **favorecer a utilidade, mas manter a privacidade a um nível razoável**. De seguida temos a descrição do primeiro modelo de anonimização que aplicamos.

Hierarquias

Desenvolvemos Hierarquias para todos os atributos classificados como QID, exceto o atributo **sex** que na nossa opinião não tem grande ganho a definir uma hierarquia para este atributo.

Age

Ao observar o a distribuição do atributo age no *dataset* original (figura 5) reparamos que os valores tem uma distribuição muito dispersa , e idealmente para garantir maior privacidade esses devem-se juntar em intervalos.

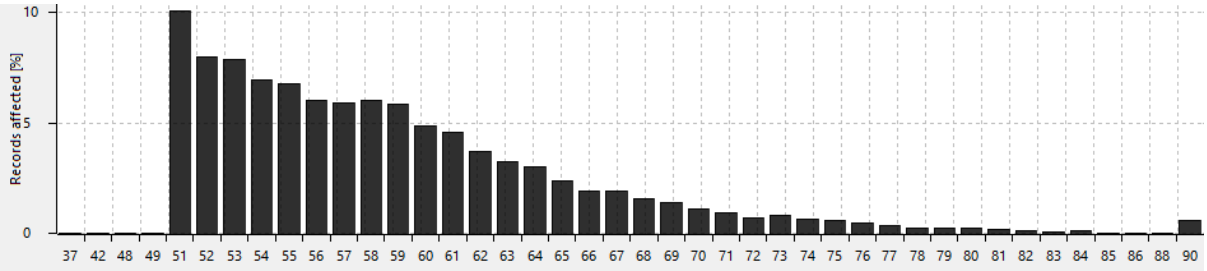

Figura 5 - Distribuição do atributo age no dataset original

Na primeira hierarquia definimos apenas um nível. Desta forma os valores ficariam menos dispersos do que no *dataset* original.

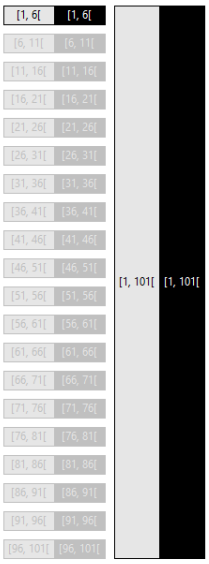

Figura 6 – Hierarquia criada e respetiva distribuição do atributo age (1ª iteração)

Após alguns testes, decidimos que necessitávamos de uma hierarquia com maior numero de níveis , para permitir que a distribuição de valores seja menos dispersa.Para tal definimos uma hierarquia com 5 níveis , em que os níveis podem ser descritos da seguinte forma:

- Nível 0
  - Contem o número em si
- Nível 1
  - Contem intervalos fechados de 5 em 5 (e.g. [1,5] , [6,10] , etc.)
- Nível 2
  - Contem intervalos fechados de 10 em 10 (e.g. [1,10] , [11,20] , etc.)
- Nível 3
  - Contem intervalos fechados de 20 em 20 (e.g. [1,20] , [21,40] , etc.)
- Nível 4
  - Dados que não se enquadram nos em qualquer um dos níveis anteriores intervalos definidos

Marital-status

Ao observar o a distribuição do atributo Marital-status no *dataset* original (figura 8) reparamos que os dados tem uma distribuição dispersa e neste caso trata-se de dados não numéricos. Portanto temos que definir uma hierarquia com conjuntos.

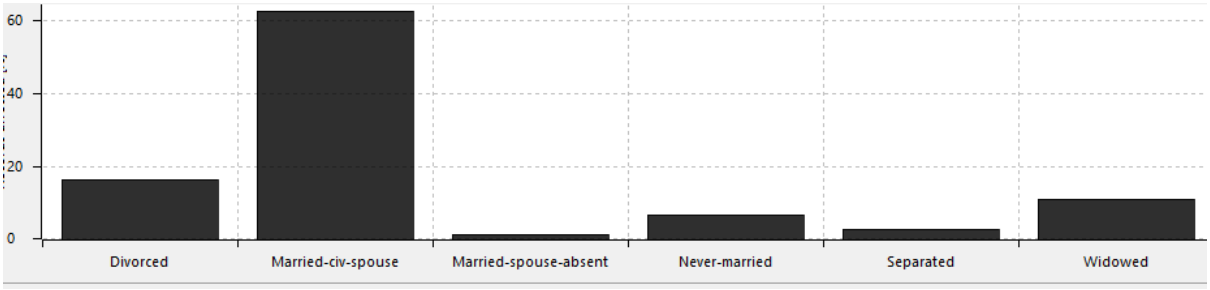

Figura 7 - Distribuição do atributo Marital-status no dataset original

No nosso caso decidimos definir os conjuntos **spouse not present** e **spouse presente** (figura 9). Desta forma podemos obter uma distribuição mais compacta.

Level-0	Level-1	Level-2
Divorced	Spouse not pres...	*
Never-married	Spouse not pres...	*
Separated	Spouse not pres...	*
Widowed	Spouse not pres...	*
Married-spouse-...	Spouse not pres...	*
Married-AF-spo...	Spouse present	*

Figura 8 - Hierarquia criada para o atributo Marital-status

Education

Ao observar o a distribuição do atributo Education no *dataset* original (figura 11) reparamos que os valores tem uma distribuição bastante dispersa e mais uma vez temos dados não numéricos, logo temos que definir uma hierarquia com conjuntos.

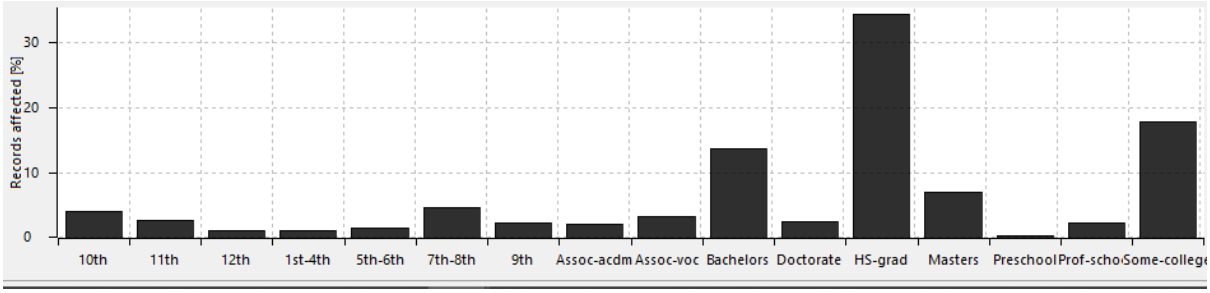

Figura 9 - Distribuição do atributo Education no dataset original

Decidimos definir uma hierarquia com quatro níveis (descrita na figura 12) . Com isto conseguimos melhorar a distribuição dos dados.

Level-0	Level-1	Level-2	Level-3
Some-college	Undergraduate	Higher education	*
11th	High School	Secondary educ...	*
HS-grad	High School	Secondary educ...	*
Prof-school	Professional Edu...	Higher education	*
Assoc-acdm	Professional Edu...	Higher education	*
Assoc-voc	Professional Edu...	Higher education	*
9th	High School	Secondary educ...	*
7th-8th	High School	Secondary educ...	*
12th	High School	Secondary educ...	*
Masters	Graduate	Higher education	*
1st-4th	Primary School	Primary education	*
10th	High School	Secondary educ...	*
Doctorate	Graduate	Higher education	*
5th-6th	Primary School	Primary education	*
Preschool	Primary School	Primary education	*

Figura 10 - Hierarquia criada para o atributo Education

Workclass

Ao observar o a distribuição do atributo Workclass no *dataset* original (figura 14) reparamos que os valores tem uma distribuição dispersa e novamente temos dados não numéricos, logo temos que definir uma hierarquia com conjuntos.

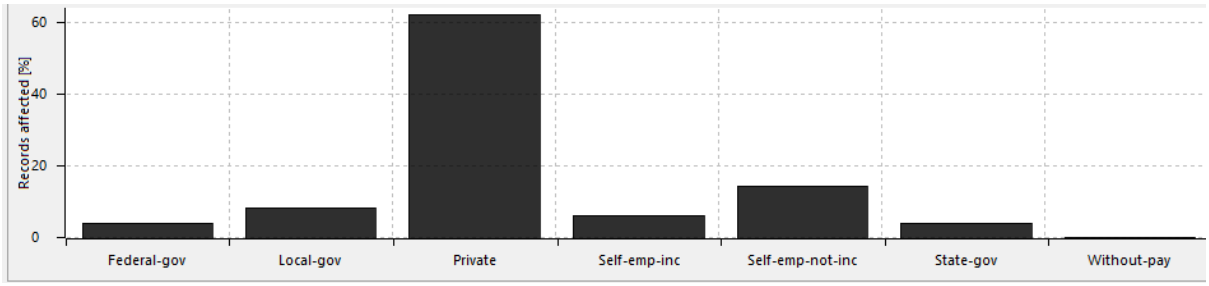

Figura 11 - Distribuição do atributo Workclass no dataset original

Concluímos que para a hierarquia deste atributo que 3 níveis seriam suficientes para obter uma distribuição mais compacta (descrição na figura 15).

Level-0	Level-1	Level-2
Self-emp-not-inc	Non-Government	*
Self-emp-inc	Non-Government	*
Federal-gov	Government	*
Local-gov	Government	*
State-gov	Government	*
Without-pay	Unemployed	*
Never-worked	Unemployed	*

Figura 12 - Hierarquia criada para o atributo Workclass

Occupation

Ao observar o a distribuição do atributo Occupation no dataset original (figura 17) verificamos que o dados tem uma distribuição dispersa e mais uma vez temos dados não numéricos, logo temos que definir uma hierarquia com conjuntos.

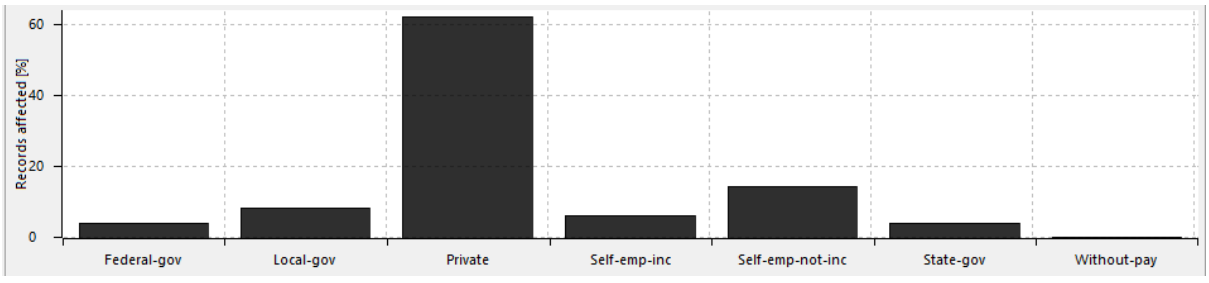

Figura 13 - Distribuição do atributo Ocupation no dataset original

Concluímos que para a hierarquia deste atributo seria suficiente 3 níveis. Com isto obtemos uma distribuição mais compacta (descrição na figura 18).

Level-0	Level-1	Level-2
Craft-repair	Technical	*
Other-service	Other	*
Sales	Nontechnical	*
Exec-managerial	Nontechnical	*
Prof-specialty	Technical	*
Handlers-cleaners	Nontechnical	*
Machine-op-ins...	Technical	*
Adm-clerical	Other	*
Farming-fishing	Other	*
Transport-moving	Other	*
Priv-house-serv	Other	*
Protective-serv	Other	*
Armed-Forces	Other	*

Figura 14 - Hierarquia criada para o atributo Ocupation

Atribuição de pesos aos atributos

Definimos os pesos associados a cada atributo (figura 15), por forma a garantir que o ARX irá tentar reduzir ao máximo a **perda de informação** nos atributos que consideramos relevante para o **contexto**:

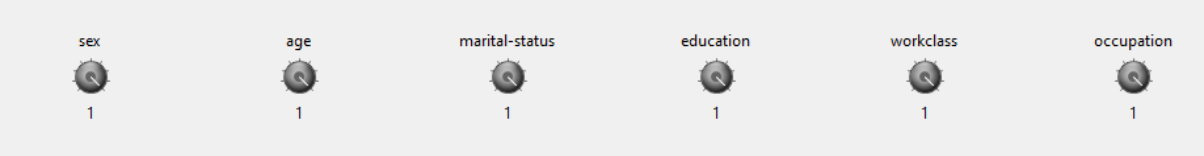

Figura 15 - Pesos atribuídos a cada atributo

É importante referir que testamos diferentes pesos e reparamos que o ideal é um atribuir um peso >= 0.5 , sendo que não há perdas ou ganhos significativos de utilidade ou privacidade se aumentarmos os pesos de qualquer combinação de atributos para cima de 0.5 .Mesmo assim , tendo em conta o contexto decidimos colocar o peso a 1 de todos os atributos .

Coding model

Utilizamos os modelo de privacidade definido mais a frente **(k=5 e l = 2),** para realizar as observações que se seguem.

Maior generalização

Podemos verificar que existe uma perda de aproximadamente 1% dos registos. Este valores de perda parecem um bom sinal, em termos de utilidade, contudo isto faz com que a distribuição do atributo **education** seja demasiado genérica (ver figura 18) , ficando a utilidade diminuída. Tendo isto em conta, este coding model não foi o escolhido.

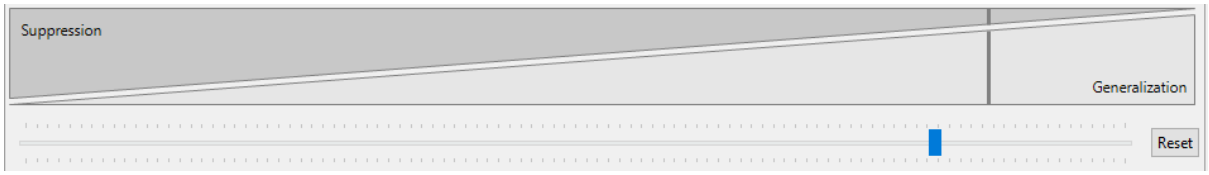

Figura 16 - Coding model de maior generalização

Parameter	Value
Scale of measure	Nominal scale
Number of measures	5697

Parameter	Value
Scale of measure	Nominal scale
Number of measures	5635

Figura 17 - Number of measures antes e depois da anonimização (com coding model da figura 16)

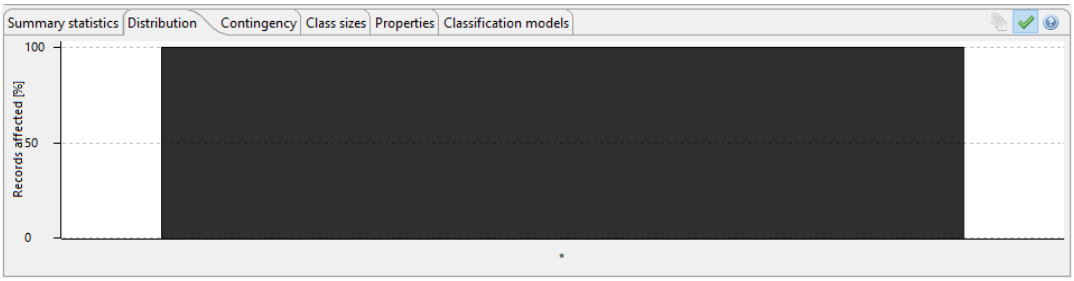

Figura 18 - distribuição do atributo education (utilizando o coding model definido na Figura 16)

Maior supressão

Com o coding model que favorece uma maior supressão, tal como esta indicado na figura 19, verificamos que há uma perda de 28% dos registos (o valor foi obtido com o campo number of measures que esta na figura 20) .Tendo isto em mente decidimos excluir este coding model.

Figura 19 – Coding model de maior supressão

Parameter	Value
Scale of measure	Nominal scale
Number of measures	5697

Parameter	Value
Scale of measure	Nominal scale
Number of measures	4075

Figura 20 - Number of measures antes e depois da anonimização (com coding model da figura 19)

"Intermedio" entre generalização e supressão (escolhido)

Com as observações feitas nos dois pontos anteriores, e após verificar o comportamento em termos de utilidade (tem distribuição dos atributos razoável e tem uma perda de 8% dos registos) e privacidade (tem níveis de risco aceitáveis) do **coding model default** , decidimos utiliza-lo como modelo.

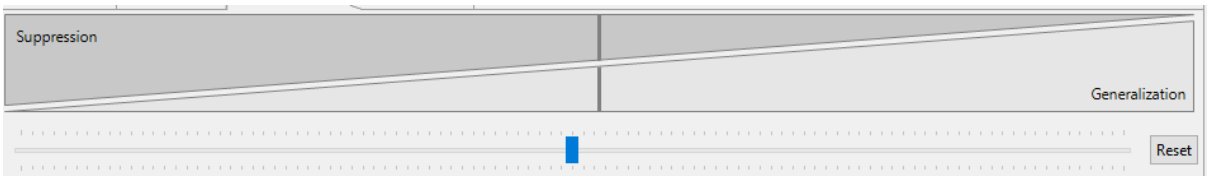

Figura 21 - coding model escolhido

Parameter	Value
Scale of measure	Nominal scale
Number of measures	5697

Parameter	Value
Scale of measure	Nominal scale
Number of measures	5255

Figura 22 - Number of measures antes e depois da anonimização (com coding model da figura 21)

Modelos de privacidade

Decidimos utilizar os modelos de privacidade **K-anonimity** e **L-diversity** , pelo o facto de que a junção de ambos os modelos permite a questão de ter classes de equivalência com atributos sensíveis diversos (importante referir que definimos **limite de supressão a 100** , tal como recomendado por o ARX).Para definir o valor de **K** (visto que o valor de **L** é igual a **2** , porque estamos a usar **l-diversity e** o atributo sensível **salary-class** apenas pode assumir dois valores) fizemos varias atribuições a **K** possíveis e destacamos as seguintes :

K=10

Com este valore podemos observar que , de forma geral , temos uma maior privacidade, mas em contrapartida uma menor utilidade de dados.

- Tendo como medida o Number of measures que existem no *dataset* podemos indicar que houve uma perda de aproximadamente 11% dos registos (que indica uma menor utilidade)
- Para verificar mais concretamente a utilidade verificamos a distribuição dos atributos e constatamos que a distribuição dos atributos é boa , ou seja, consegue compactar a distribuição original .
    - Decidimos não adicionar capturas de ecrã das distribuições , para não sobrecarregar o relatório
- A privacidade obtida com esta configuração é boa, podemos observar os respetivos valores na figura 23

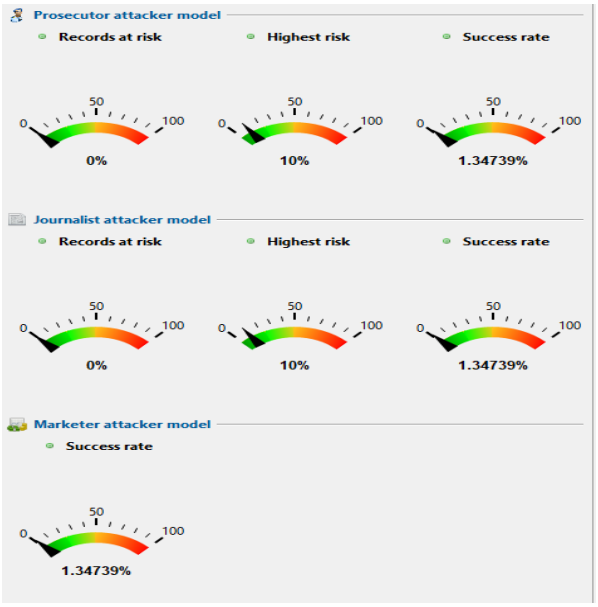

Figura 23 - Analise de risco para k=10 e l =2

K= 5 (valor escolhido)

Com este valor observamos que temos uma maior utilidade de dados, mas em contrapartida uma menor privacidade, contundo tal como foi referido anteriormente a ideia deste modelo consistiria em aplicações em que queremos favorecer a utilidade, mas queremos ter uma privacidade razoável.

- Em termos de utilidade temos uma perda de aproximadamente 8% registos.
- Para verificar mais concretamente a utilidade verificamos a distribuição e reparamos que não existe nenhuma diferença significativa a distribuição que temos no caso em que **K=10.**
- A privacidade obtida com esta configuração é razoável, mas menor do que aquela obtida no caso em que **K=10.** Podemos observar os resultados na aqui.
- Como o objetivo deste modelo é favorecer utilidade mantendo um nível de privacidade razoável, decidimos que este valor de K era o que vamos usar neste modelo.

k = 3

Por fim decidimos testar com o valor de **K=3** e verificamos que obtemos menos privacidade e menos utilidade do que k=5, visto que

- Temos uma perda de aproximadamente 12% dos registos, que é maior que a perde de registos de 8% em **K=5**
- A distribuição do atributo workclass é mais dispersa do que temos no caso em que **K=5** (logo temos uma menor utilidade)

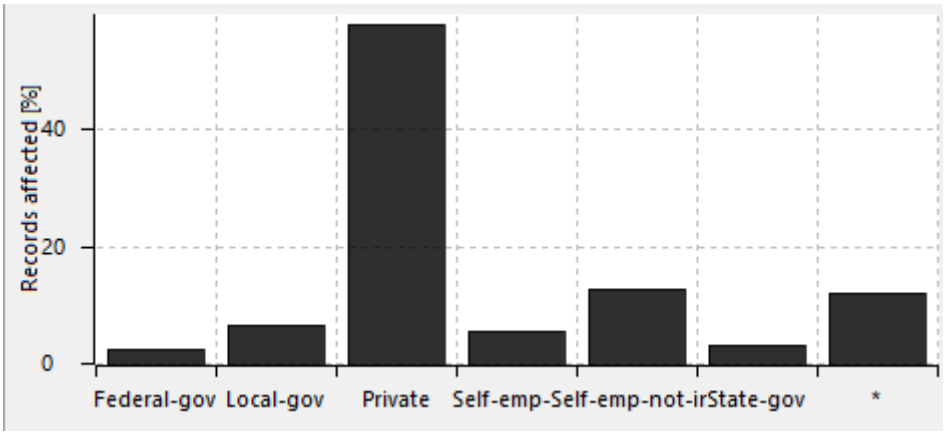

Figura 24 - distribuição do atributo workclass (quando k=3)

- Privacidade diminui , visto que os riscos de ataque são mais elevados

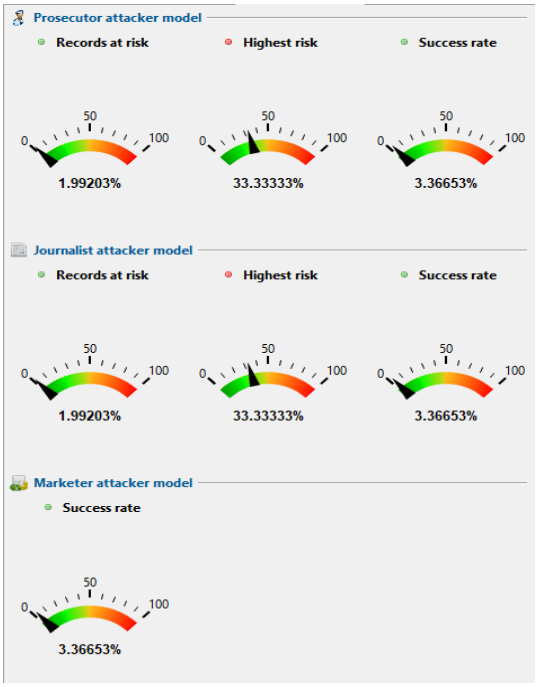

Figura 25 - Analise de risco para k=3 e l =2

Analise dos riscos

Observando as figuras 24 e 25 , podemos constatar que os riscos associados a cada um dos ataques foram reduzidos drasticamente .Contundo seria possível diminuir ainda mais os riscos associados , mas como o objetivo deste modelo é favorecer a utilidade ficamos satisfeitos com os resultados obtidos .

Figura 26 - Analise de risco antes e após aplicação de anonimização

Measure	Value [%]
Lowest prosecutor risk	4%
Records affected by lowest risk	0.43883%
Average prosecutor risk	69.84378%
Highest prosecutor risk	100%
Records affected by highest risk	55.538%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	69.84378%
Sample uniques	55.538%
Population uniques	1.34438%
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Measure	Value [%]
Lowest prosecutor risk	0.21552%
Records affected by lowest risk	8.82969%
Average prosecutor risk	1.6746%
Highest prosecutor risk	20%
Records affected by highest risk	0.38059%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	1.6746%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figura 27 - Analise de risco detalhada antes e após aplicação de anonimização

Analise da distribuição

Neste subtópico apenas vamos colocar imagem sobre a **distribuição** associada a alguns **atributos,** após a **aplicação do modelo de anonimização** (a distribuição original de maior parte dos atributos , exceto do atributo **Sex** , pode ser encontrada aqui).

- **Sex**

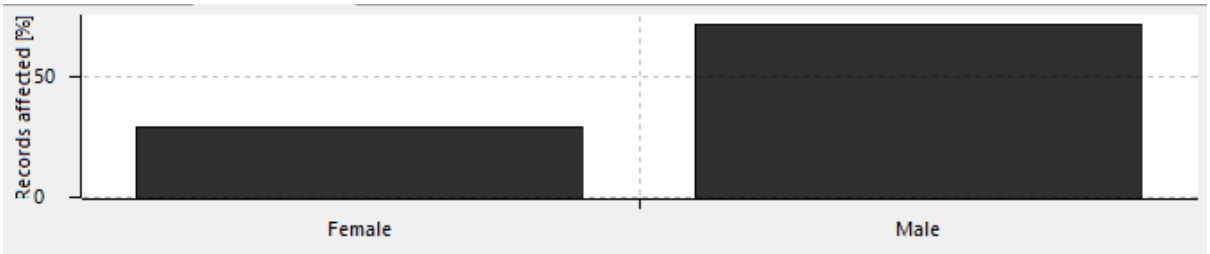

Figura 28 - Distribuição original do atributo Sex

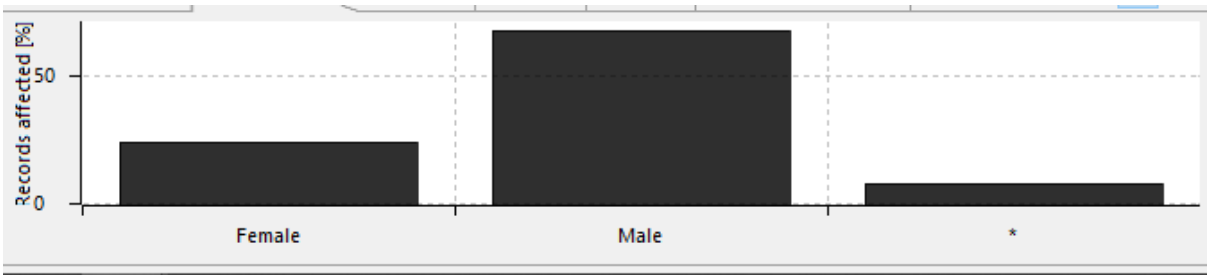

Figura 29 - Distribuição do atributo Sex (Modelo 1)

- **Age**

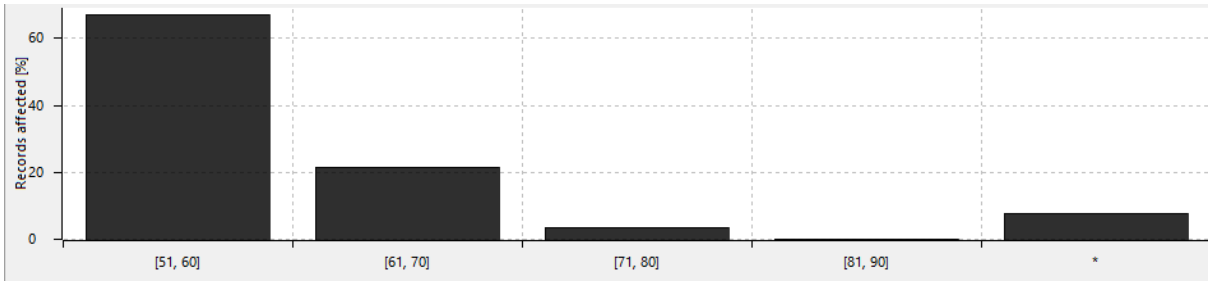

Figura 30 - Distribuição do atributo Age (Modelo 1)

- **Marital-status**

Figura 31 - Distribuição do atributo Marital-status (Modelo 1)

- **Education**

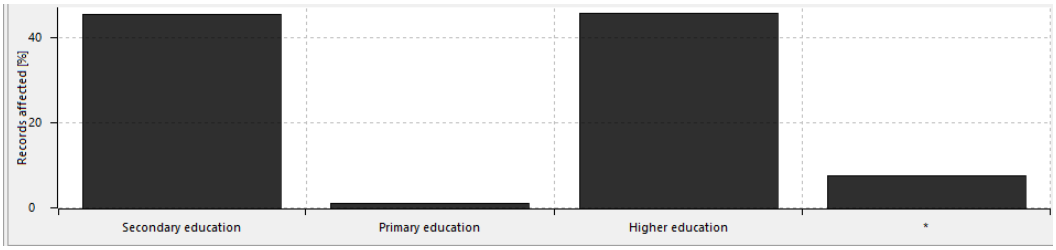

Figura 32 - Distribuição do atributo Education (Modelo 1)

- **Workclass**

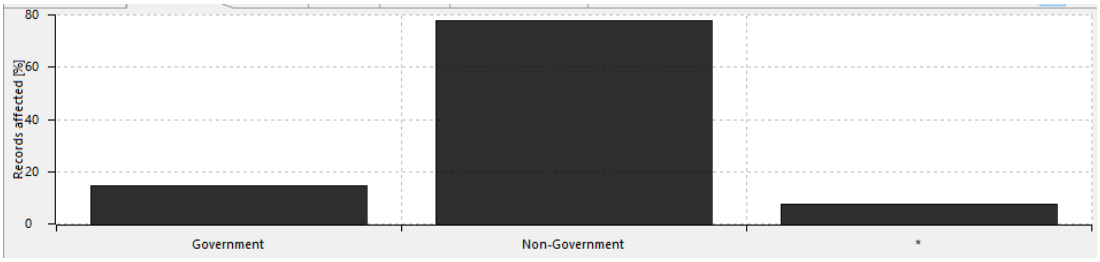

Figura 33 - Distribuição do atributo Workclass (Modelo 1)

- **Occupation**

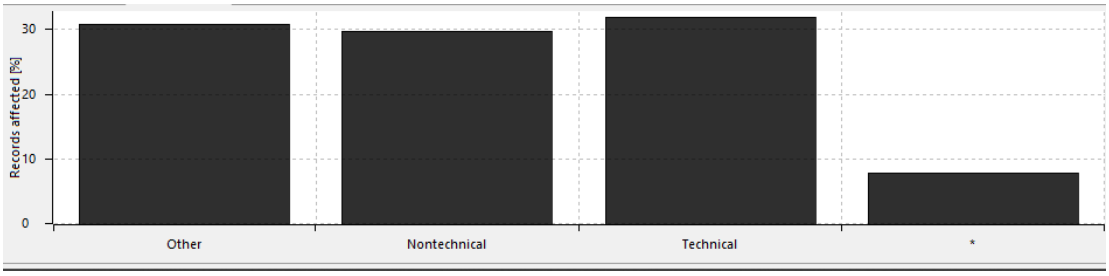

Figura 34 - Distribuição do atributo Occupation (Modelo 1)

Esquema dos resultados

O modelo que concebemos apresenta um score ARX de 0%, na figura 33 podemos verificar a transformação ótima a amarelo.

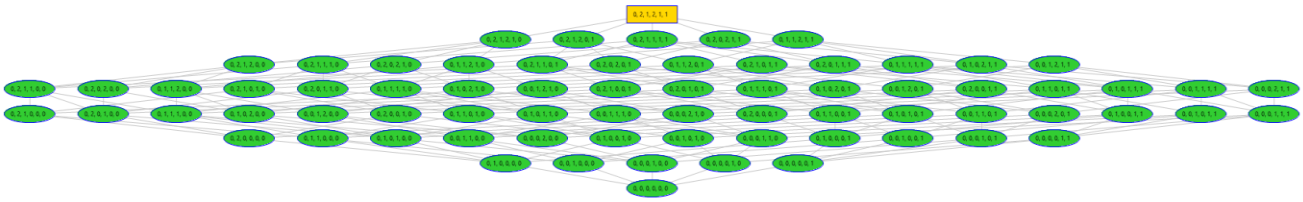

Figura 35 - resultados do modelo 1

Modelo 2

A ideia principal que definimos ao conceber este modelo consiste em **favorecer ainda mais a utilidade, mas manter a privacidade a um nível razoável**. De seguida temos a descrição do primeiro modelo de anonimização que aplicamos.

Hierarquias

As hierarquias deste modelo são as mesmas que no modelo 1, mas forem limitadas de modo a não ser possível anonimizar completamente uma coluna (o último nível das hierarquias foi excluído)

Atribuição de pesos aos atributos

Definimos os pesos associados a cada atributo (figura 34), por forma a garantir e reparamos que não é possível generalizar completamente uma coluna os pesos dos atributos estarem a 0 a 1 ou a qualquer outro valer não tem diferença, mas se os pesos forem diferentes entre os atributos a ordem de otimalidade é alterada preferindo modelo que minimizam a generalização em função do peso dos atributos(a arvore gerada contem os mesmos elementos só por uma ordem diferente) podendo o modelo ótimo não ser o ótimo global:

Nós escolhemos os pesos da figura de forma a manter a dispersão de cada classe sendo o peso maior quanto maior for a dispersão da classe

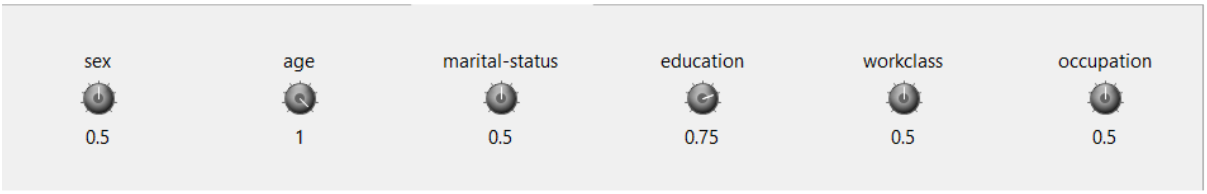

Figura 36 - Pesos dos atributos

Coding model

Utilizamos o modelo de privacidade definido mais a frente **(δ =5 e Γ = 2)** , e definimos a proporção de generalização supressão a 50% por não termos preferência em nenhum dos métodos.

Analise dos riscos

Observando a figura 35, podemos constatar que os riscos associados a cada um dos ataques foram reduzidos drasticamente. Contundo seria possível diminuir ainda mais os riscos associados, sendo ainda assim este modelo menos seguro que o modelo 1, mas o objetivo deste modelo é fornecer mais utilidade aos dados e ficamos satisfeitos com os resultados obtidos.

Modelos de privacidade

Decidimos utilizar os modelos de privacidade **δ -disclosure privacy** e **Γ - average reidentification risc**, pelo facto de serem modelos que permitir gerar bons resultados (importante referir que definimos **limite de supressão a 50**, o que resultou num menor número de modelos gerados mantendo apenas os modelos com uma percentagem baixa. Para definir o valor de **δ** testamos os seguintes valores 2,5,10, para δ igual a 2 o número de atributos "missing" aumenta significativa mais elevado que com **δ** igual a 5(que foi o escolhido) no caso de **δ** igual a 10 existe uma diminuição ligeira de atributos "missing" e um aumento ligeiro do número de registo com risco máximo sendo também uma possibilidade, estes testes foram feitos com Γ igual a 0.02. Para definir o valor de **Γ** testamos os seguintes valores .01,.02,.03, para Γ igual a .1 existe uma elevada generalização dos dados e o número de elemento "missing" é também elevado para Γ igual .02(que foi o valor escolhido) apresentando um bom equilíbrio entre privacidade e utilidade, para **Γ** igual a .03 o "highest risc" era 100% não havendo grande perda de utilidade dos dados, mas sem ser garantida a privacidade estes testes foram feitos com **δ** igual a 5.

Abaixo estão os resultados para o modelo escolhido:

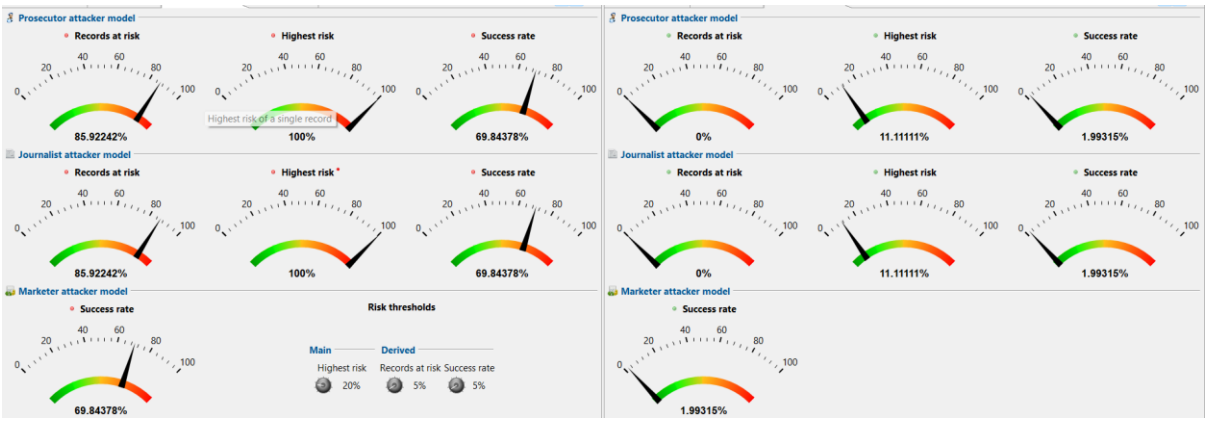

Figura 37 - Atack models antes e depois da aplicação das medidas de privacidade

Measure	Value [%]
Lowest prosecutor risk	4%
Records affected by lowest risk	0.43883%
Average prosecutor risk	69.84378%
Highest prosecutor risk	100%
Records affected by highest risk	55.538%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	69.84378%
Sample uniques	55.538%
Population uniques	1.34438%
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figura 38 - Performance do modelo antes da aplicação das medidas de privacidade

Measure	Value [%]
Lowest prosecutor risk	0.2584%
Records affected by lowest risk	7.79142%
Average prosecutor risk	1.99315%
Highest prosecutor risk	11.11111%
Records affected by highest risk	1.26837%
Estimated prosecutor risk	11.11111%
Estimated journalist risk	11.11111%
Estimated marketer risk	1.99315%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, occupation, sex, workclass

Figura 39 - Performance do modelo depois da aplicação das medidas de privacidade

Analise da distribuição

Neste subtópico vamos colocar imagem sobre a **distribuição** associada a alguns **atributos,** após a **aplicação do modelo de anonimização** (a distribuição original destes atributos pode ser encontrada aqui) e comparar com o modelo 1.

- **Sex**

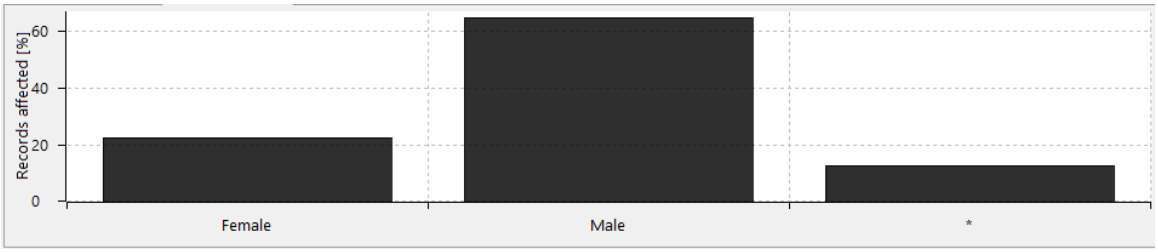

Figura 40 - Distribuição da classe Sex

- **Age**

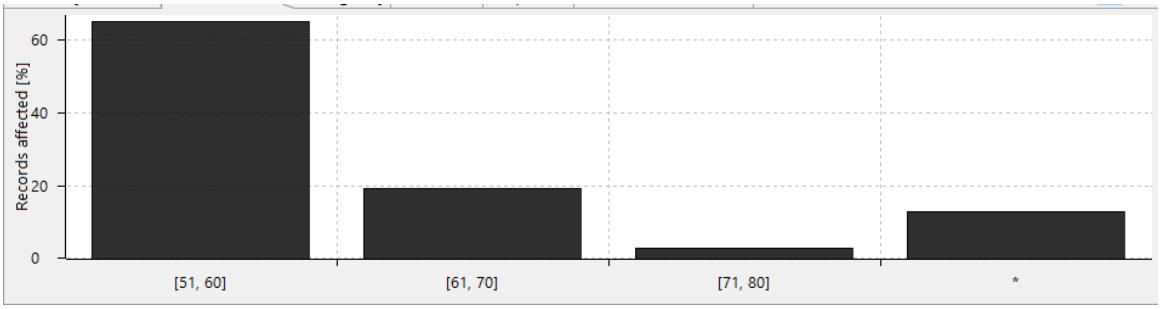

Figura 41 - Distribuição da classe Age

- **Education**

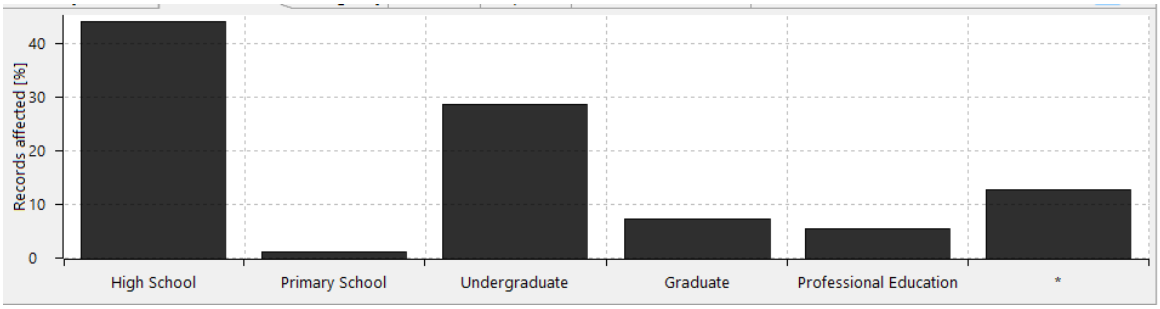

Figura 42 - Distribuição da classe Education

- **Marital status**

Figura 43 - Distribuição da classe Marital Status

- **Workclass**

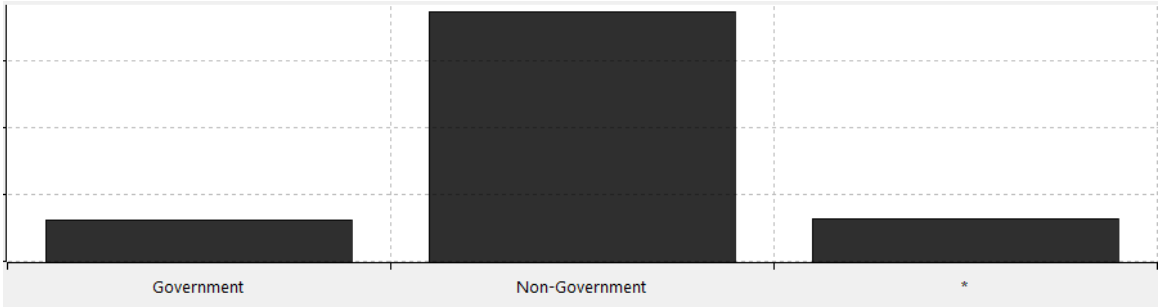

Figura 44 - Distribuição da classe Workclass

- **Ocupation**

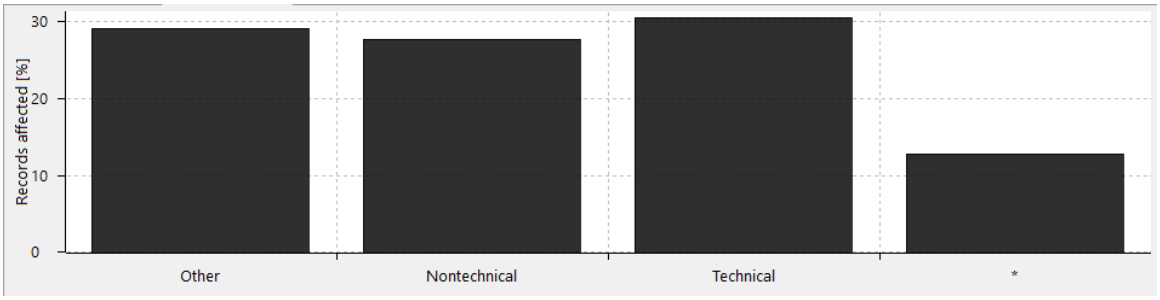

Figura 45 - Distribuição da classe Ocupation

As proporção e distribuições das classes são muito semelhantes ao modelo 1 exceto a classe “Education” que se encontra menos generalizada e apresenta uma maior retenção da proporção original.

Esquema dos resultados

Encontra-se abaixo na figura 44 a árvore gerada pelo arx sendo que na figura 45 se encontra todos os modelos ponderados e no formato de caixa o modelo escolhido

Figura 46 - Árvore gerada pelo ARX

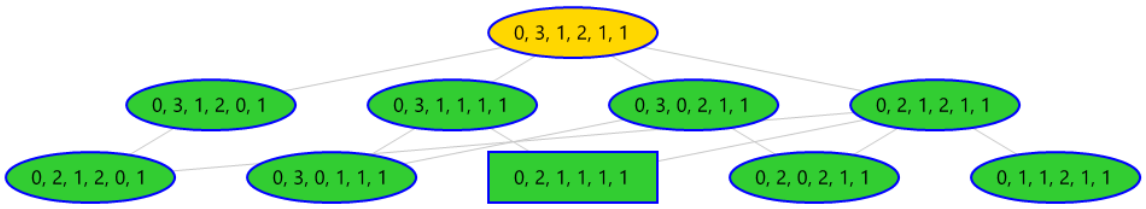

Figura 47 - Árvore de tamanho reduzido limitada pelo score

Conclusão

Concluímos que conseguimos gerar 2 modelos de privacidade que conseguem garantir a privacidade dos dados a um nível aceitável sem comprometer a utilidade dos dados, ambos os modelos apresentam um risco muito baixo, e em modelos futuros caso seja necessária maior utilidade dos dados deve ser criado um modelo que comprometa um pouco mais a privacidade de modo a garantir essa maior utilidade mas para nós os modelos apresentados apresentam o melhor compromisso entre utilidade e privacidade sendo por isso os modelos apresentados.