U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Professor João Vilela

Segurança e Privacidade

Quarto Trabalho: Anonimização de um Dataset com Análise de Risco e Utilidade

Maio de 2022

Trabalho realizado pelo Grupo S, constituído por:

Pedro Leite - 201906697

Pedro Carvalho - 201906291

Índice

1. Introdução

O objetivo deste projeto é anonimizar um Dataset de grande escala, utilizando a ferramenta ARX. Ao longo do processo de anonimização descrevemos os diferentes passos que nos levaram ao modelo final: classificação dos atributos do Dataset, dizendo se são Identifiers, Quasi-Identifiers, Sensitive ou Insensitive, descrever os riscos de privacidade do Dataset original, aplicar dois modelos de anonimização ao Dataset original e descreve-los.

2. Classificação dos Atributos

2.1. Identifiers e Quasi-Identifiers

Para a classificação dos Identifiers e dos Quasi-Identifiers (QID), começamos por assumir vários atributos como QIDs, e fomos retirando os atributos um a um, vendo várias combinações possíveis. Fizemos uma análise e determinamos quais tinham maior impacto na percentagem de distribuição sendo esses os QIDs. Nenhum atributo demonstrou ter um impacto suficientemente grande, para ser Identifier. O "Education" e o "Education-num" demonstraram o mesmo valor de distribuição, logo vão ter a mesma classificação, neste caso QID.

Age	QID
Work-class	QID
Education	QID
Education-num	QID
Ocupation	QID
Relationship	QID
Hours-per-week	QID

2.2. Sensitive e Insensitive

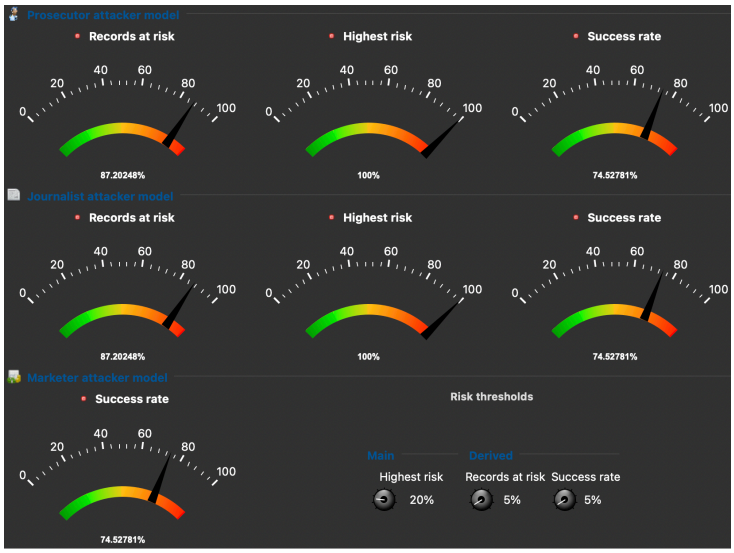
Para a classificação dos Sensitive, consideramos apenas os atributos que não são QIDs ou Identifiers. Classificamos como Sensitive, os atributos que não devem ser expostos publicamente. Sendo estes o "Capital-gain" e o "Capital-loss". Não

colocamos o "Salary", como um atributo Sensitive, por não ser específico suficiente já que só coloca os registos num intervalo de salário e não contém o salário concreto. Classificamos como Insensitive, os atributos que sobraram.

Fnlwgt	Insensitive
Martial-status	Insensitive
Race	Insensitive
Sex	Insensitive
Capittal-gain	Sensitive
Capital-loss	Sensitive
Native-country	Insensitive
Salary	Insensitive

3. Análise dos Riscos de Privacidade do Dataset Original

Na forma original, o Dataset demonstrou apresentar riscos muito elevados a todo tipos de ataques ("Prosecutor", que ataca um indivíduo em específico, "Journalist", que ataca qualquer indivíduo e "Marketer", que ataca o maior número de indivíduos possível). Já que este Dataset não está minimamente anonimizado, por outro lado a utilidade vai ser muito alta.



Risco Dataset Original

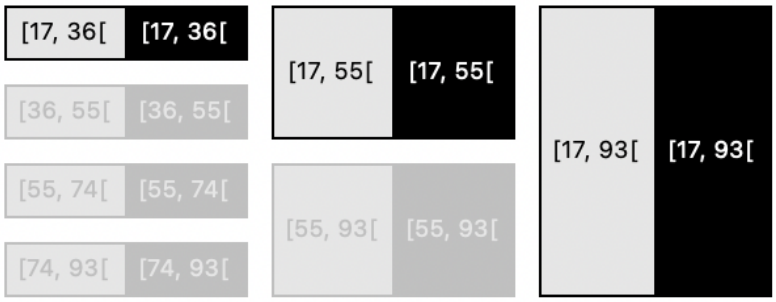
Measure	Value [%]
Lowest prosecutor risk	4.16667%
Records affected by lowest risk	0.07371%
Average prosecutor risk	74.52781%
Highest prosecutor risk	100%
Records affected by highest risk	62.67314%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	74.52781%
Sample uniques	62.67314%

Risco Dataset Original 2

4. Análise dos Modelos

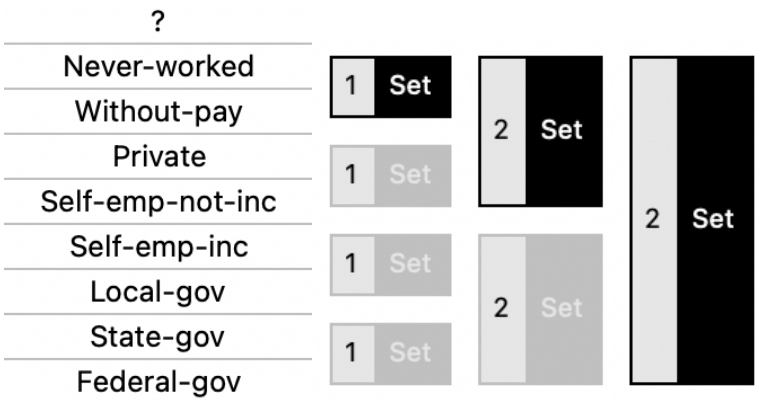
4.1. Modelo 1

Depois de escolhidos os QIDs, críamos uma hierarquia para cada um deles, com um intervalo adequado, para os atributos numéricos. Por exemplo no atributo ”age”, a hierarquia é a seguinte:



Neste caso, definimos o intervalo máximo como [17,93[, porque todos os registos do Dataset se encontram neste intervalo e este não é demasiado grande.

Para os atributos não numéricos, utilizamos ordenação. Por exemplo, para o atributo “workclass”, a hierarquia é a seguinte:



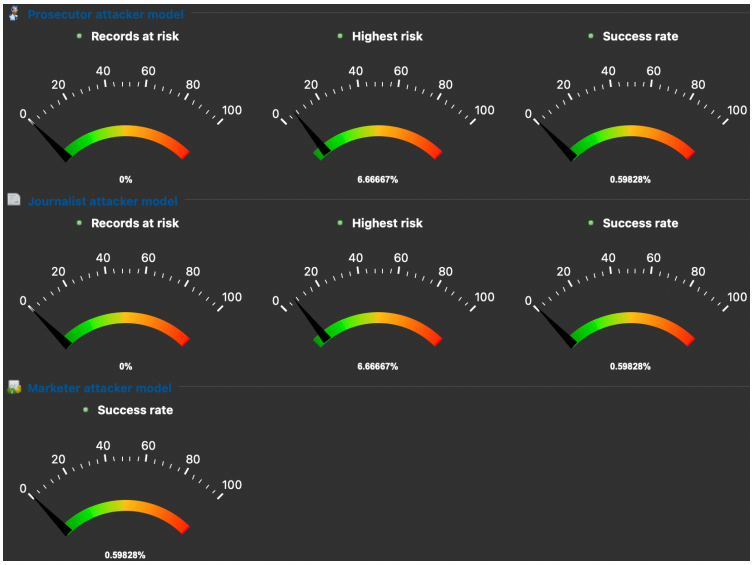
O primeiro modelo que escolhemos foi o "15-anonymity", com "2-diversity" nos atributos sensíveis e colocamos a "suppression" a 100%. Inicialmente tínhamos escolhido um modelo com "2-anonymity", mas este não tinha muito utilidade. Por isso fomos subindo o valor do "k" até encontrarmos um modelo que considerássemos que tivesse um equilíbrio entre segurança e utilidade. Para medir este equilíbrio, comparamos o valor do "number of measures" do Dataset original, com o modelo atual e verificamos que apenas houve uma perda de aproximadamente 12% dos registos. Mas em contrapartida, o risco diminui significativamente.

Parameter	Value
Scale of measure	Ratio scale
Number of measures	32561
Number of distinct values	73
Mode	36
Median	37
Min	17
Max	90
Arithmetic mean	39
Sample variance	186

Number of Measures Dataset Original

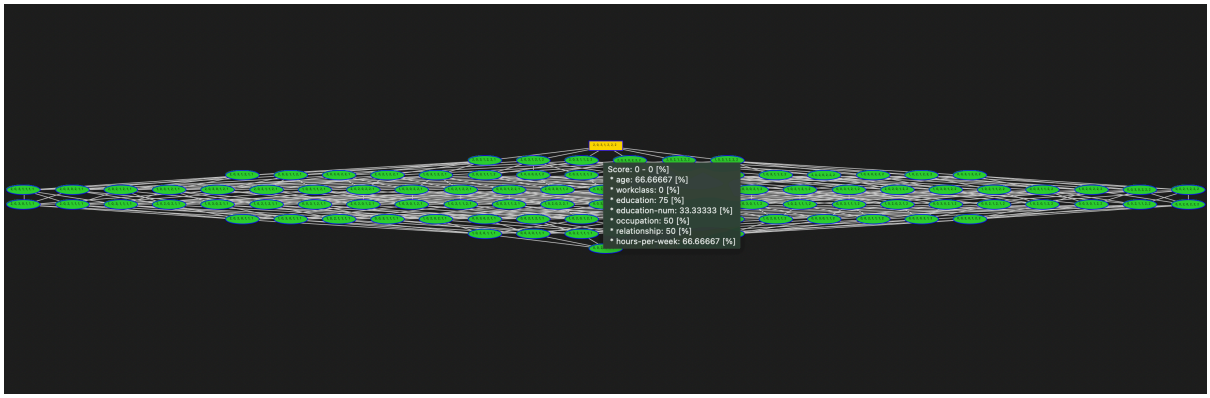
Parameter	Value
Scale of measure	Ordinal scale
Number of measures	28916
Number of distinct values	2
Mode	[17, 55[
Median	[17, 55[
Min	[17, 55[
Max	[55, 91[

Number of Measures Modelo 1



Risco Modelo 1

O modelo que escolhemos apresenta um score ARX de 0%, como pretendido (a amarelo está a transformação ótima).



Níveis Modelo 1

4.2. Modelo 2

No segundo modelo as únicas diferenças em relação ao modelo 1, é que em vez de utilizarmos "l-diversity", utilizamos "t-closeness", com t=0.4. Para chegarmos a este valor, fomos aumentando t até chegar a um valor a partir do qual o "number of measures" não se alterava. Mantivemos o valor de k no "k-anonymity" como 15, porque depois de analisar vários valores, a relação de utilidade e privacidade, para vários valores de k, 15 pareceu-nos o valor mais adequado.

Parameter	Value
Scale of measure	Ordinal scale
Number of measures	29607
Number of distinct values	4
Mode	[9, 13[
Median	[9, 13[
Min	[1, 5[
Max	[13, 17[

Number of Measures Modelo 2

Risco Modelo 2

Neste modelo a "success rate" aumentou aproximadamente 0.3% nos três tipos de ataques, portanto o Dataset acaba por estar mais vulnerável a ataques em comparação com o modelo 1.

5. Conclusão

Os resultados obtidos do estudo dos dois modelos são semelhantes. O modelo 1 demonstrou ser mais seguro, enquanto o modelo 2 demonstrou ser mais útil. Já que o "success rate" do modelo 1 é mais baixo e o "number of measures" do modelo 2 é mais alto. Dependendo do objetivo de quem está a gerir o Dataset, pode ser preferível o modelo 1 ou modelo 2.