**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)

1. **Índice (copiar índice do trabalho anterior)**
2. Introdução
3. Classificação de Atributos
4. Riscos de privacidade do dataset original
5. Modelos de Privacidade
6. Conclusão
7. Referências
8. **Introdução (samuel)-> introduzir ARX, exemplo dado, objetivo da anonimização do dataset...**
9. **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.

|  |  |  |
| --- | --- | --- |
| **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 discriminatórios. |
| 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 discriminatórios. |

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

1. **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**

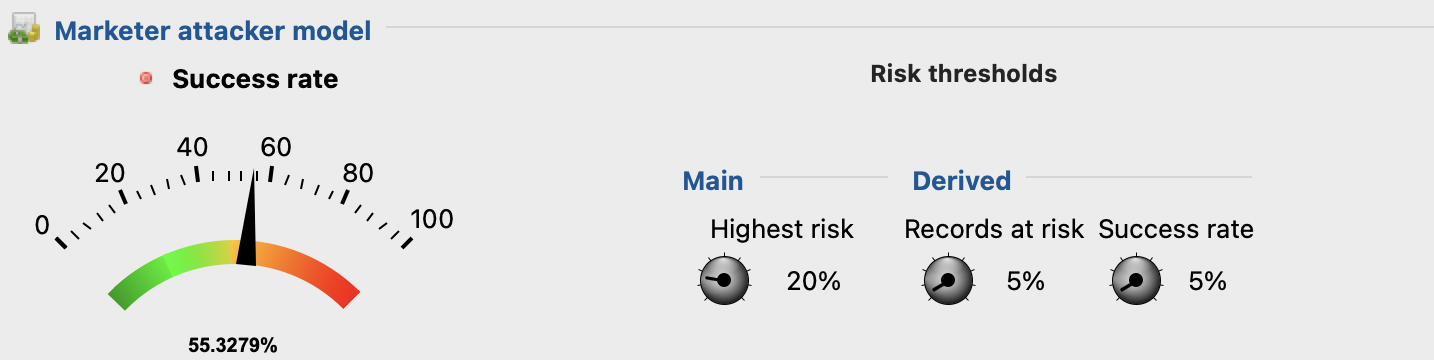
Neste modelo, o agressor tem como alvo um indivíduo específico, mas não se espera que ele tenha conhecimento prévio sobre a filiação.



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.

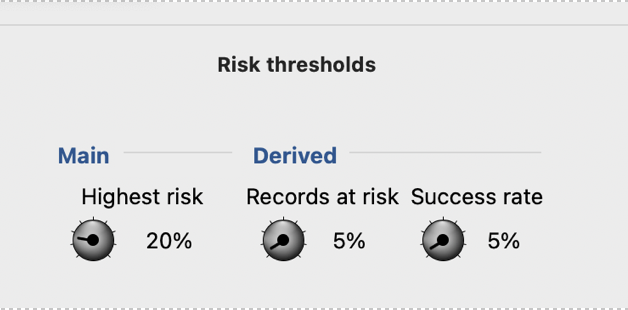
Como está apresentado na figura 2, cerca de 68.4% dos registos podem ser reidentificados por um jornalista, mas 100% dos registos atingem o risco máximo de reidentificação sob esse modelo. Assim, tal como no Modelo do Promotor, a taxa de sucesso na reidentificação de registos é de 55.33%.

**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.

1. **Modelos de Privacidade**

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*).

**4.1. 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*).

4.1.1. Variação de *k* e *l:* análise de riscos e de utilidade

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;

Uma imagem com texto, captura de ecrã, file, diagrama

Os conteúdos gerados por IA podem estar incorretos.- k = 10 e l = 2.

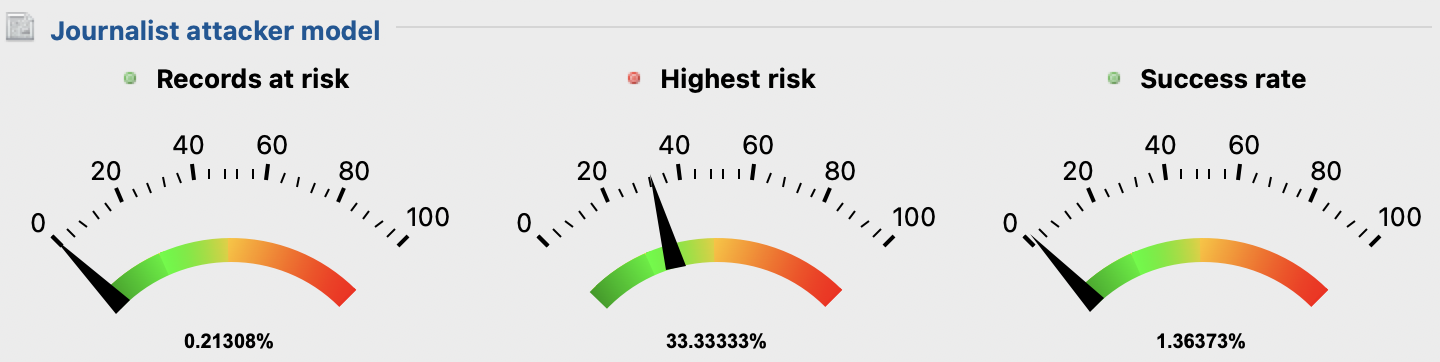
Uma imagem com texto, file, captura de ecrã, diagrama

Os conteúdos gerados por IA podem estar incorretos.Uma imagem com texto, file, captura de ecrã, diagrama

Os conteúdos gerados por IA podem estar incorretos.

3 primeiros com utility measure Loss e tudo padrão (sup 100%)

**K-ANONIMITY (3) + L\_DIVERSITY (2)**

**ANALISE DE RISCO ATTACKER MODELS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.22676% |
| Registos afetados pelo menor risco | 9.39697% |
| Risco Médio (Promotor) | 1.36373% |
| Maior risco (Promotor) | 33,33333% |
| Registos afetados pelo maior risco | 0.12785% |
| Risco estimado (Promotor) | 33.33333% |
| Risco estimado (Jornalista) | 33.33333% |
| Risco estimado (Marketing) | 1.36373% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Análise de Utilidade:**

**Attribute-level quality**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 84.44069% | 15.55931% | 15.55931% | 14.20482% | 15.55931% |
| age | String | 84.44069% | 3,88983% | 12.57318% | 6.89859% | 8.17171% |
| Marital-status | String | 84.44069% | 7.77966% | 10.29717% | 8.99834% | 1.58099  % |
| education | String | 84.44069% | 5.18644% | 9.8844% | 5.66722% | 4.78501  % |
| Native-country | String | 84.44069% | 15.55931% | 15.55931% | 3.00695% | 16.25901% |
| workclass | String | 84.44069% | 7.77966% | 11.1138% | 7.29163% | 12.96432% |
| occupation | String | 84.44069% | 7.77966% | 11.12531% | 7.2774% | 9.40166% |

**Dataset-level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 9.07627% |
| Granularity | 12.13012% |
| N. -U. entropy | 7.22233% |
| DIscerniblity | 15.46884% |
| Average Class Size | 99.76286% |
| Record-level squared error | 7.62272% |
| Attribute-level squared error | 10.64182% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + L\_DIVERSITY (4)**







**ANALISE DE RISCOS ATTACKER MODELS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0% |
| Registos afetados pelo menor risco | 0% |
| Risco Médio (Promotor) | 0% |
| Maior risco (Promotor) | 0% |
| Registos afetados pelo maior risco | 0% |
| Risco estimado (Promotor) | 0% |
| Risco estimado (Jornalista) | 0% |
| Risco estimado (Marketing) | 0% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 100% | 0% | 0% | 0% | 0% |
| occupation | String | 100% | 0% | 0% | 0% | 0% |

**Dataset-level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 0% |
| Granularity | 0% |
| N. -U. entropy | 0% |
| DIscerniblity | N/A |
| Average Class Size | N/A |
| Record-level squared error | 0% |
| Attribute-level squared error | 0% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (10) + L\_DIVERSITY (2)**

**ANALISE DE RISCOS ATTACKER MODELS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.05476% |
| Registos afetados pelo menor risco | 7.09429% |
| Risco Médio (Promotor) | 0.65659% |
| Maior risco (Promotor) | 10% |
| Registos afetados pelo maior risco | 0.11655% |
| Risco estimado (Promotor) | 10% |
| Risco estimado (Jornalista) | 10% |
| Risco estimado (Marketing) | 0.65659% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 14.66415% | 85.33585% | 85.33585% | 83.86806% | 85.33585% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 14.66415% | 85.33585% | 85.33585% | 81.46696% | 87.56813% |
| education | String | 14.66415% | 28.44528% | 53.67062% | 30.84719% | 27.11003% |
| Native-country | String | 14.66415% | 85.33585% | 85.33585% | 36.97475% | 88.60237% |
| workclass | String | 14.66415% | 85.33585% | 85.33585% | 81.06505% | 86.29174% |
| occupation | String | 14.66415% | 42.66793% | 60.86925% | 39.67725% | 49.54891% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 58.92237% |
| Granularity | 62.76138% |
| N. -U. entropy | 35.10996% |
| DIscerniblity | 83.39336% |
| Average Class Size | 99.50102% |
| Record-level squared error | 41.7976% |
| Attribute-level squared error | 30.98822% |
| Aggregation-specific squared error | 0% |

**Outros parametros (utility measure, coding model e attribute weights)**

**K-ANONIMITY (5) + L\_DIVERSITY (2) utility measure Discernibility, supression limit 100%**



**ANALISE DE RISCOS ATTACKER MODELS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.06423% |
| Registos afetados pelo menor risco | 5.24613% |
| Risco Médio (Promotor) | 0.30998% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 0.01685% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 0.30998% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 1.60135% | 55.97275% | 55.97275% | 54.10872% | 14.65291% |
| education | String | 1.60135% | 79.6742% | 79.6742% | 58.4584% | 76.38503% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 100% | 0% | 0% | 0.% | 0% |
| occupation | String | 1.60135% | 98.39865% | 98.39865% | 98.65089% | 98.20246% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 30.45672% |
| Granularity | 35.29809% |
| N. -U. entropy | 35.67446% |
| DIscerniblity | 96.00528% |
| Average Class Size | 98.93638% |
| Record-level squared error | 8.79876% |
| Attribute-level squared error | 4.51125% |
| Aggregation-specific squared error | 0% |

**Com sup limit 50%**

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 1.60135% | 49.19932% | 55.97275% | 54.10872% | 14.65291% |
| education | String | 1.60135% | 65.5991% | 79.6742% | 58.4584% | 76.38503% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 100% | 0% | 0% | 0.% | 0% |
| occupation | String | 1.60135% | 98.39865% | 98.39865% | 98.65089% | 98.20246% |

**Dataset level quality**

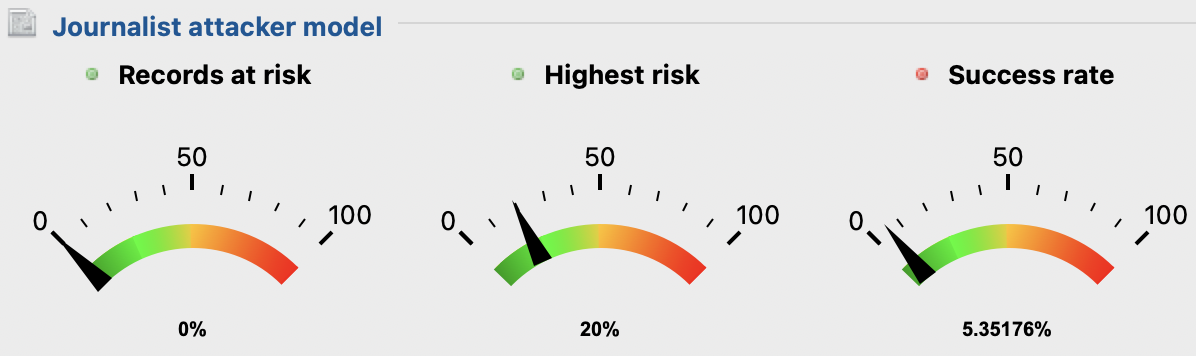
|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 30.45672% |
| Granularity | 35.29809% |
| N. -U. entropy | 35.67446% |
| DIscerniblity | 96.00528% |
| Average Class Size | 98.93638% |
| Record-level squared error | 8.79876% |
| Attribute-level squared error | 4.51125% |
| Aggregation-specific squared error | 0% |

Sup Limit 10% utility measure discernibility  
(riscos iguais) e utilidade igual

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 1.60135% | 49.19932% | 55.97275% | 54.10872% | 14.65291% |
| education | String | 1.60135% | 65.5991% | 79.6742% | 58.4584% | 76.38503% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 100% | 0% | 0% | 0.% | 0% |
| occupation | String | 1.60135% | 98.39865% | 98.39865% | 98.65089% | 98.20246% |

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 30.45672% |
| Granularity | 35.29809% |
| N. -U. entropy | 35.67446% |
| DIscerniblity | 96.00528% |
| Average Class Size | 98.93638% |
| Record-level squared error | 8.79876% |
| Attribute-level squared error | 4.51125% |
| Aggregation-specific squared error | 0% |

**Sup limit 100% utility measure average equivalence class size**



|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 1.1236% |
| Registos afetados pelo menor risco | 0.52573% |
| Risco Médio (Promotor) | 5.25176% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 1.24047% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 5.35176% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 43.87309% | 56.12691% | 56.12691% | 52.66636% | 56.12691% |
| age | String | 43.87309% | 56.12691% | 56.12691% | 55.746% | 51.71774% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 43.87309% | 37.41794 | 45.82963% | 31.24988% | 46.3936% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 100% | 0% | 0% | 0.% | 0% |
| occupation | String | 43.87309% | 56.12691% | 56.12691% | 55.01472% | 55.76918% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 29.39981% |
| Granularity | 29.47187% |
| N. -U. entropy | 37.83696% |
| DIscerniblity | 56.08446% |
| Average Class Size | 99.94404% |
| Record-level squared error | 10.90012% |
| Attribute-level squared error | 34.61386% |
| Aggregation-specific squared error | 0% |

**Utility Measure LOSS**

**Uma imagem com texto, captura de ecrã, Tipo de letra, file

Os conteúdos gerados por IA podem estar incorretos.**

**Uma imagem com texto, captura de ecrã, Tipo de letra, file

Os conteúdos gerados por IA podem estar incorretos.**

**Uma imagem com texto, captura de ecrã, Tipo de letra, Instrumento de medição

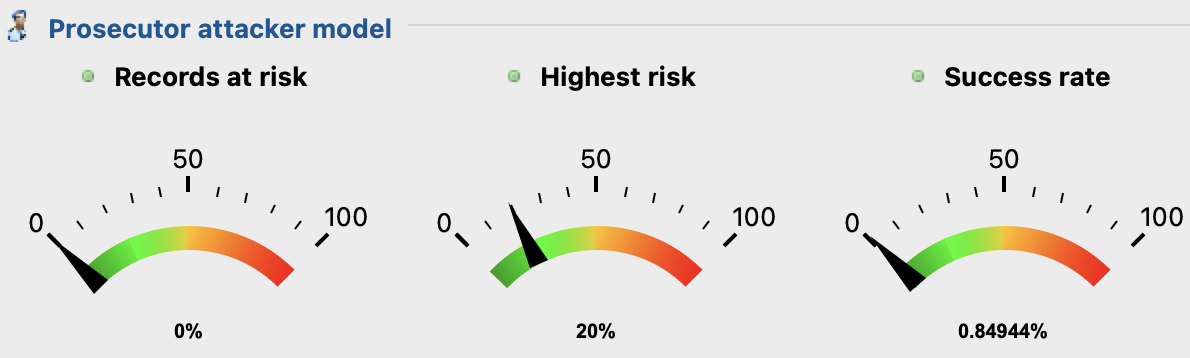
Os conteúdos gerados por IA podem estar incorretos.**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.05476% |
| Registos afetados pelo menor risco | 7.03552% |
| Risco Médio (Promotor) | 0.7783% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 0.15412% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 0. 7783% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 13.95133% | 86.04867% | 86.04867% | 84.66006% | 86.04867% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 13.95133% | 86.04867% | 86.04867% | 82.37882% | 88.30018% |
| education | String | 13.95133% | 28.68289% | 54.12395% | 31.15165% | 27.33838% |
| Native-country | String | 13.95133% | 86.04867% | 86.04867% | 39.53836% | 89.19385% |
| workclass | String | 13.95133% | 86.04867% | 86.04867% | 82.27332% | 86.86455% |
| occupation | String | 13.95133% | 43.02434% | 61.41579% | 40.01922% | 49.99781% |

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 59.41456% |
| Granularity | 63.29224% |
| N. -U. entropy | 35.59828% |
| DIscerniblity | 84.10614% |
| Average Class Size | 99.57998% |
| Record-level squared error | 42.1172% |
| Attribute-level squared error | 31.1998% |
| Aggregation-specific squared error | 0% |

**Sup limit 100% utility measure Loss Attribute Weights age 0.8, sex 0.2**





|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.08518% |
| Registos afetados pelo menor risco | 4.8176% |
| Risco Médio (Promotor) | 0.84944% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 0.20518% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 0.84944% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**Analise de utilidade:**  
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 19.20629% | 80.79371% | 80.79371% | 78.88615% | 80.79371% |
| age | String | 19.20629% | 20.19843% | 65.28785% | 20.56381% | 71.45296% |
| Marital-status | String | 19.20629% | 80.79371% | 80.79371% | 77.6066% | 84.89527% |
| education | String | 19.20629% | 26.93124% | 50.77316% | 29.05622% | 25.5293% |
| Native-country | String | 19.20629% | 80.79371% | 80.79371% | 32.60047% | 83.86574% |
| workclass | String | 19.20629% | 40.39686% | 57.7098% | 33.72721.% | 71.48557% |
| occupation | String | 19.20629% | 40.39686% | 57.87976% | 37.64467% | 47.95735% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 52.90065% |
| Granularity | 66.78098% |
| N. -U. entropy | 36.40676% |
| DIscerniblity | 79.53727% |
| Average Class Size | 99.61566% |
| Record-level squared error | 53.87927% |
| Attribute-level squared error | 73.48319% |
| Aggregation-specific squared error | 10.60074% |

**Sup limit 100% utility measure Loss Attribute Weights age 0.2, sex 0.8**

Analise de riscos igual a loss padrão , e utilidade tambem

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 13.95133% | 86.04867% | 86.04867% | 84.66006% | 86.04867% |
| age | String | 100% | 0% | 0% | 0% | 0 % |
| Marital-status | String | 13.95133% | 86.04867% | 86.04867% | 82.27332% | 88.30018% |
| education | String | 13.95133% | 28.68289% | 54.12395% | 31.15165% | 27.33838% |
| Native-country | String | 13.95133% | 86.04867% | 86.04867% | 39.53836% | 89.19385% |
| workclass | String | 13.95133% | 86.04867% | 86.04867% | 82.27332.% | 86.86455% |
| occupation | String | 13.95133% | 43.02434% | 61.41579% | 40.01922% | 49.99781% |

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 59.41456% |
| Granularity | 63.29224% |
| N. -U. entropy | 35.59828% |
| DIscerniblity | 84.10614% |
| Average Class Size | 99.57998% |
| Record-level squared error | 42.1172% |
| Attribute-level squared error | 31.1998% |
| Aggregation-specific squared error | 0% |

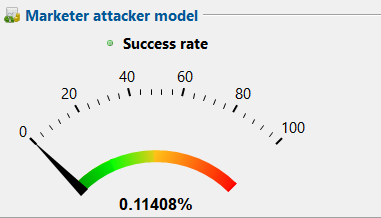
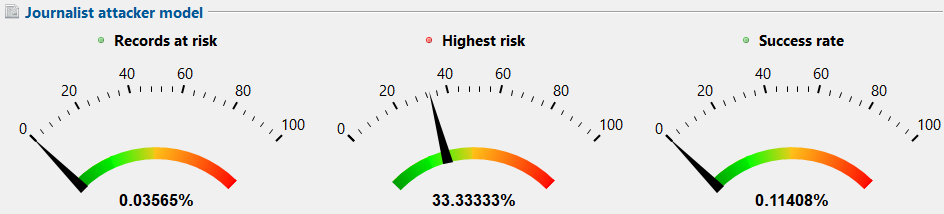
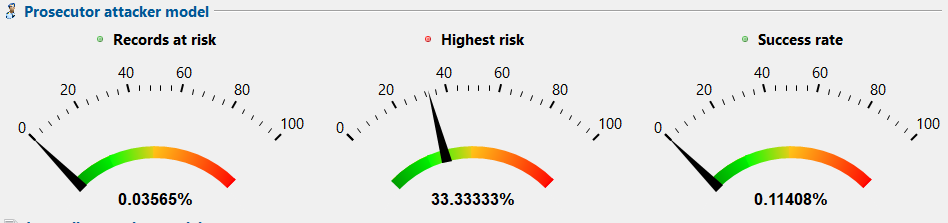
PARTE DO MAX

**Primeira parte, variação de K e T**

**K ANONIMITY (K = 3) + T-CLOSENESS(T = 0.15)**

**Supression Limit 100%**  
**Utility Measure: Loss**

**Attribute Weights: padrão (0.5 for all)**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.00623% |
| Registos afetados pelo menor risco | 57.21212% |
| Risco Médio (Promotor) | 0.11408% |
| Maior risco (Promotor) | 33.33333% |
| Registos afetados pelo maior risco | 0.02139% |
| Risco estimado (Promotor) | 33.33333% |
| Risco estimado (Jornalista) | 33.33333% |
| Risco estimado (Marketing) | 0.11408% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 7.00219% | 92.99781% | 92.99781% | 93.069% | 92.99781% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 7.00219% | 92.99781% | 92.99781% | 64.4205% | 96.48339% |
| workclass | String | 7.00219% | 46.49891% | 66.42701% | 39.72671% | 82.31514% |
| occupation | String | 100% | 0% | 0% | 0% | 0% |

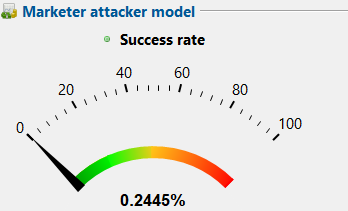
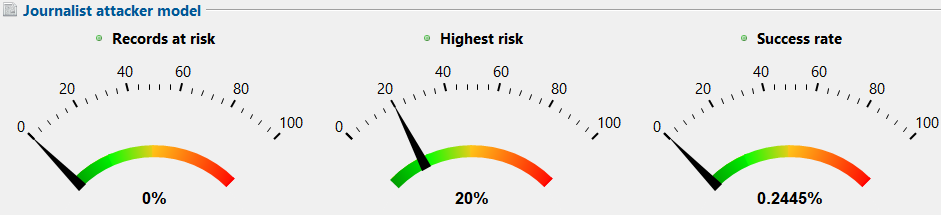
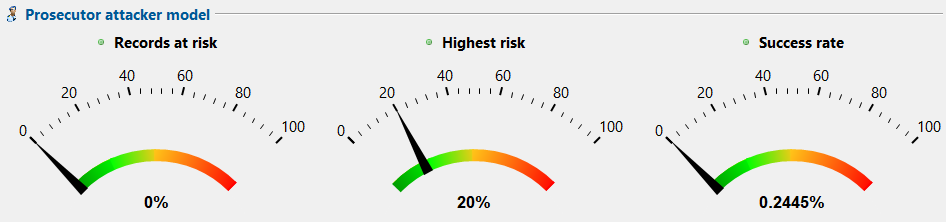
**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 33.2135% |
| Granularity | 32.21599% |
| N. -U. entropy | 11.47625% |
| DIscerniblity | 57.73714% |
| Average Class Size | 97.09964% |
| Record-level squared error | 28.17757% |
| Attribute-level squared error | 31.12375% |
| Aggregation-specific squared error | 0% |

**K ANONIMITY (K = 5) + T-CLOSENESS(T = 0.2)**

**Supression Limit: 100%**  
**Utility Measure: Loss**

**Attribute Weights: padrão (0.5 for all)**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.01795% |
| Registos afetados pelo menor risco | 21.27903% |
| Risco Médio (Promotor) | 0.2445% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 0.0191% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 0.2445% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 13.2153% | 86.7847% | 86.7847% | 87.33269% | 86.7847% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 13.2153% | 86.7847% | 86.7847% | 45.83333% | 89.74524% |
| workclass | String | 13.2153% | 86.7847% | 86.7847% | 78.71717% | 89.88957% |
| occupation | String | 13.2153% | 43.39235% | 61.71877% | 40.39283% | 49.89006% |

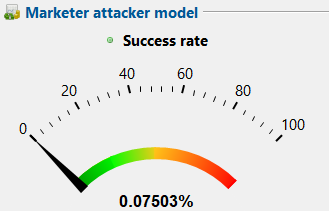
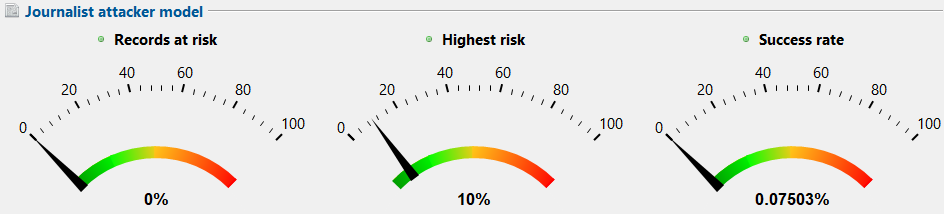
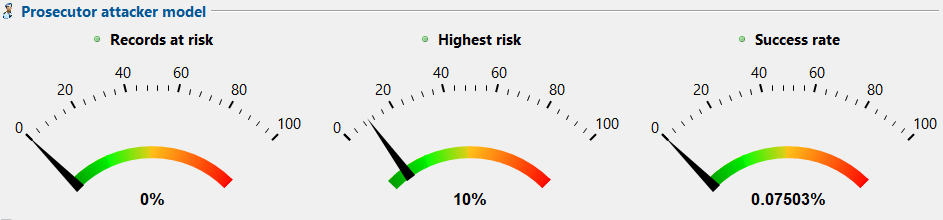
**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 43.39235% |
| Granularity | 43.09281% |
| N. -U. entropy | 21.61159% |
| DIscerniblity | 77.92549% |
| Average Class Size | 98.6499% |
| Record-level squared error | 29.95506% |
| Attribute-level squared error | 30.00089% |
| Aggregation-specific squared error | 0% |

**K ANONIMITY (K = 10) + T-CLOSENESS(T = 0.15)**

**Supression Limit: 100%**  
**Utility Measure: Loss**

**Attribute Weights: padrão (0.5 for all)**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.00623% |
| Registos afetados pelo menor risco | 57.33886% |
| Risco Médio (Promotor) | 0.07503% |
| Maior risco (Promotor) | 10% |
| Registos afetados pelo maior risco | 0.03573% |
| Risco estimado (Promotor) | 10% |
| Risco estimado (Jornalista) | 10% |
| Risco estimado (Marketing) | 0.07503% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

**Attribute-level quality:**

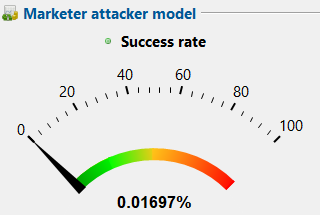
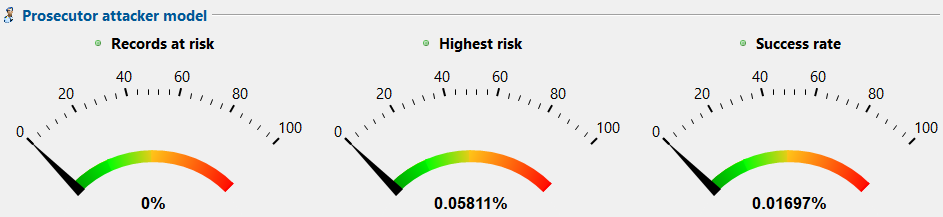
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 7.20774% | 92.79226% | 92.79226% | 92.79226% | 92.79226% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 7.20774% | 92.79226% | 92.79226% | 62.55056% | 96.37% |
| workclass | String | 7.20774% | 46.39613% | 66.28018% | 39.56702% | 82.1597% |
| occupation | String | 100% | 0% | 0% | 0% | 0% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 33.14009% |
| Granularity | 32.14478% |
| N. -U. entropy | 11.3564% |
| DIscerniblity | 57.53159% |
| Average Class Size | 95.58705% |
| Record-level squared error | 28.1346% |
| Attribute-level squared error | 31.08676% |
| Aggregation-specific squared error | 0% |

**Segunda parte, variação de supression limit, utility measure and attribute weights com parametros especificos -> esses nao vao ser alterados!**

**K-ANONIMITY (5) + T-CLOSENESS (0.15) utility measure Discernibility, supression limit 100%**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.01084% |
| Registos afetados pelo menor risco | 31.30473% |
| Risco Médio (Promotor) | 0.01697% |
| Maior risco (Promotor) | 0.05811% |
| Registos afetados pelo maior risco | 5.84142% |
| Risco estimado (Promotor) | 0.05811% |
| Risco estimado (Jornalista) | 0.05811% |
| Risco estimado (Marketing) | 0.01697% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

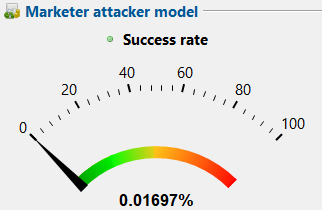
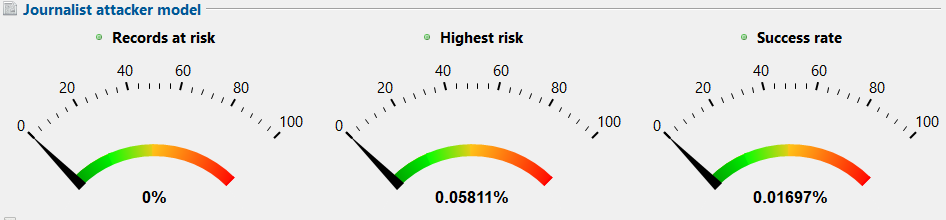
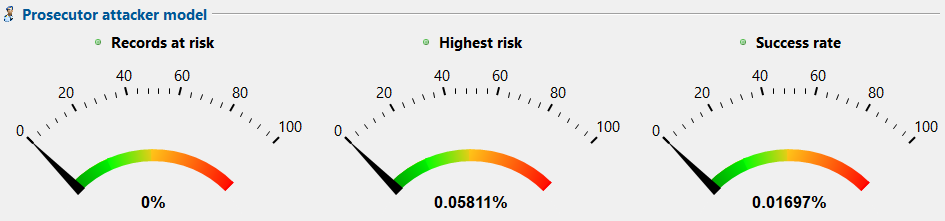
**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 2.3208% | 48.8396% | 67.77086% | 41.97076% | 87.78506% |
| occupation | String | 2.3208% | 48.8396% | 69.37461% | 46.41364% | 55.66466% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 13.95417% |
| Granularity | 20.69136% |
| N. -U. entropy | 12.81342% |
| DIscerniblity | 72.41476% |
| Average Class Size | 80.46898% |
| Record-level squared error | 8.94323% |
| Attribute-level squared error | 1.65082% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + T-CLOSENESS (1.5) utility measure Discernibility, supression limit 50%**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.01084% |
| Registos afetados pelo menor risco | 31.30473% |
| Risco Médio (Promotor) | 0.01697% |
| Maior risco (Promotor) | 0.05811% |
| Registos afetados pelo maior risco | 5.84142% |
| Risco estimado (Promotor) | 0.05811% |
| Risco estimado (Jornalista) | 0.05811% |
| Risco estimado (Marketing) | 0.01697% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100% | 0% | 0% | 0% | 0% |
| Native-country | String | 100% | 0% | 0% | 0% | 0% |
| workclass | String | 2.3208% | 48.8396% | 67.77086% | 41.97076% | 87.78506% |
| occupation | String | 2.3208% | 48.8396% | 69.37461% | 46.41364% | 55.66466% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 13.95417% |
| Granularity | 20.69136% |
| N. -U. entropy | 12.81342% |
| DIscerniblity | 72.41476% |
| Average Class Size | 80.46898% |
| Record-level squared error | 8.94323% |
| Attribute-level squared error | 1.65082% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + T-CLOSENESS (0.15) utility measure Discernibility, supression limit 10%**

Analise de riscos igual a 50%

Analise de utilidade igual a 50%

**K-ANONIMITY (5) + T-CLOSENESS (0.15) utility measure Avg equi class size, supression limit 100%**

**Uma imagem com texto, Tipo de letra, diagrama, Gráfico

Os conteúdos gerados por IA podem estar incorretos.**

**Uma imagem com texto, Tipo de letra, diagrama, Gráfico

Os conteúdos gerados por IA podem estar incorretos.**

**Uma imagem com texto, captura de ecrã, Tipo de letra, Instrumento de medição

Os conteúdos gerados por IA podem estar incorretos.**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 1.14943% |
| Registos afetados pelo menor risco | 2.19697% |
| Risco Médio (Promotor) | 6.11111% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 2.08333% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 6.11111% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 73.74179% | 26.25821% | 26.25821% | 24.54885% | 26.25821% |
| age | String | 73.74179% | 26.25821% | 26.25821% | 25.82693% | 24.02812% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 73.74179% | 17.90947% | 20.60783% | 13.46796% | 19.20663% |
| Native-country | String | 73.74179% | 13.1291% | 19.0372% | 2.45111% | 19.20663% |
| workclass | String | 100 % | 0% | 0% | 0% | 0% |
| occupation | String | 73.74179% | 26.25821% | 26.25821% | 26.20568% | 27.87469% |

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 15.62988% |
| Granularity | 16.72213% |
| N. -U. entropy | 17.62831% |
| DIscerniblity | 26.23909% |
| Average Class Size | 99.95174% |
| Record-level squared error | 6.09655% |
| Attribute-level squared error | 17.66116% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + T-CLOSENESS (0.15) utility measure Loss, supression limit 100%**

**Uma imagem com texto, Tipo de letra, file, Gráfico

Os conteúdos gerados por IA podem estar incorretos.Uma imagem com texto, Tipo de letra, file, Gráfico

Os conteúdos gerados por IA podem estar incorretos.**

**Uma imagem com texto, captura de ecrã, Tipo de letra, número

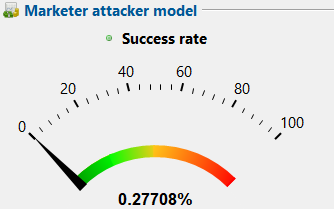
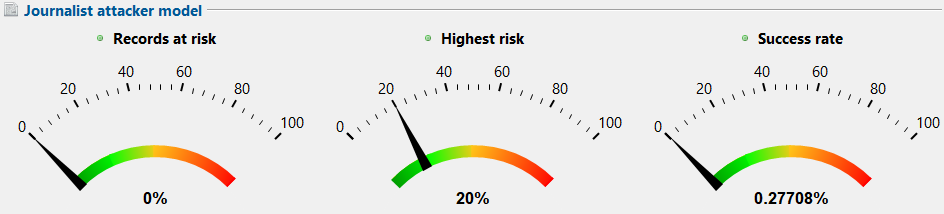
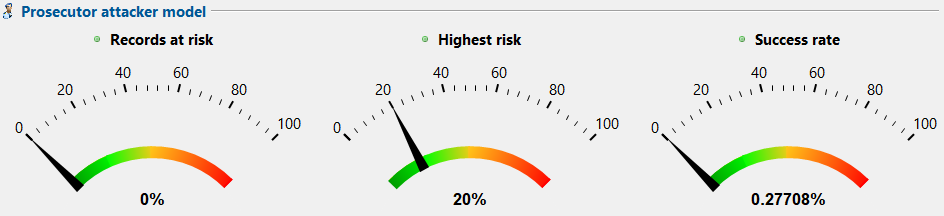
Os conteúdos gerados por IA podem estar incorretos.**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.00623% |
| Registos afetados pelo menor risco | 57.23252% |
| Risco Médio (Promotor) | 0.10342% |
| Maior risco (Promotor) | 16.66667% |
| Registos afetados pelo maior risco | 0.10699% |
| Risco estimado (Promotor) | 16.6667% |
| Risco estimado (Jornalista) | 16.6667% |
| Risco estimado (Marketing) | 0.10342% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 7.03534% | 92.96466% | 92.96466% | 93.93276 | 92.96466% |
| age | String | 100% | 0% | 0% | 0% | 0% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 100 % | 0% | 0% | 0% | 0% |
| Native-country | String | 7.03534% | 92.96466% | 92.96466% | 64.154% | 92.4692% |
| workclass | String | 7.03534 % | 46.48233% | 66.40333% | 39.67701% | 82.29841% |
| occupation | String | 100% | 0% | 0% | 0% | 0% |

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 33.20166% |
| Granularity | 32.20451% |
| N. -U. entropy | 11.45706% |
| DIscerniblity | 57.70298 % |
| Average Class Size | 96.80012% |
| Record-level squared error | 28.17212% |
| Attribute-level squared error | 31.11913% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + T-CLOSENESS (1.5) Sup limit 100% utility measure Loss Attribute Weights age 0.8, sex 0.2**



**ANALISE DE RISCOS**

|  |  |
| --- | --- |
| Medida | Valor [%] |
| Menor risco (Promotor) | 0.02236% |
| Registos afetados pelo menor risco | 22.13151% |
| Risco Médio (Promotor) | 0.27708% |
| Maior risco (Promotor) | 20% |
| Registos afetados pelo maior risco | 0.07422% |
| Risco estimado (Promotor) | 20% |
| Risco estimado (Jornalista) | 20% |
| Risco estimado (Marketing) | 0.27708% |
| Unicidade na amostra | 0% |
| Unicidade na população | 0% |
| Modelo populacional | DANKAR |
| QIDs | Age, education, marital-status, native-country, occupation, sex, workclass |

**ANALISE DE UTILIDADE**

**Attribute-level quality:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atributo | Tipo de Dados | Missings | Gen. Intensity | Granularity | N. -U. entropy | Squared Error |
| sex | String | 100% | 0% | 0% | 0% | 0% |
| age | String | 32.99184% | 16.75204% | 54.148% | 16.5191% | 61.17844% |
| Marital-status | String | 100% | 0% | 0% | 0% | 0% |
| education | String | 32.99184% | 67.00816% | 67.00816% | 60.79881% | 68.25362% |
| Native-country | String | 32.99184% | 67.00816% | 67.00816% | 24.17463% | 69.73388% |
| workclass | String | 32.99184% | 33.50408% | 47.86297% | 27.78455% | 59.73014% |
| occupation | String | 2.3208% | 0% | 0% | 0% | 0% |

**Dataset level quality**

|  |  |
| --- | --- |
| Modelo | Qualidade |
| Gen- Intensity | 26.32463% |
| Granularity | 36.88771% |
| N. -U. entropy | 19.47787% |
| DIscerniblity | 61.97154% |
| Average Class Size | 98.80935% |
| Record-level squared error | 33.09458% |
| Attribute-level squared error | 62.58453% |
| Aggregation-specific squared error | 0% |

**K-ANONIMITY (5) + T-CLOSENESS (1.5) Sup limit 100% utility measure Loss Attribute Weights age 0.2, sex 0.8**

**IGUAL AO LOSS PADRAO**

1. **Conclusão**
2. **Referências**

<https://arx.deidentifier.org/overview/privacy-criteria/>

[https://chatgpt.com](http://www.chatgpt.com)

<https://ics.uci.edu/~projects/295d/presentations/295d-tcloseness>

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

<https://www.k2view.com/blog/what-is-k-anonymity>

<https://en.wikipedia.org/wiki/T-closeness>

<https://dicionario.priberam.org>