Tecnologias de Reforço da Privacidade

Anonimização de Datasets

- com Privacidade, Utilidade e Análise de Risco -

Maria Sousa Carreira up202408787
Matilde Isabel da Silva Simões up202108782

Abril 2025

Conteúdo

1 Seleção, Importação e Objetivo do *Dataset*

1.1 *Dataset* Escolhido

Para realizar este projeto, foi escolhido o *dataset* ***Predicting Churn for Bank Customers***, disponível no *Kaggle*.

Este *dataset* contém dados pessoais e financeiros sobre clientes de um banco, portanto torna-se apropriado para aplicar técnicas de anonimização devido à natureza das informações presentes e à necessidade de proteção de dados sensíveis.

O *dataset* possui as seguintes características:

- **Número de registos:** 10.000 clientes.
- **Número de atributos:** 14.
- **Atributos detalhados:**
  - `RowNumber`: Número sequencial da linha.
  - `CustomerId`: Identificador único do cliente.
  - `Surname`: Apelido do cliente.
  - `CreditScore`: "Pontuação financeira"do cliente.
  - `Geography`: País de origem do cliente.
  - `Gender`: Género do cliente.
  - `Age`: Idade do cliente (em anos).
  - `Tenure`: Tempo de ligação do cliente com o banco (em anos).
  - `Balance`: Saldo bancário do cliente (em euros).
  - `NumOfProducts`: Número de produtos do banco que o cliente possui.
  - `HasCrCard`: Indica se o cliente tem cartão de crédito (1 - tem cartão de crédito; 0 - não tem cartão de crédito).
  - `IsActiveMember`: Indica se o cliente é um membro ativo (1 - é um membro ativo; 0 - não é um membro ativo).
  - `EstimatedSalary`: Salário anual estimado do cliente (em euros).
  - `Exited`: Indica se o cliente abandonou o banco (1 - abandonou o banco; 0 - não abandonou o banco).

1.1.1 Distribuição do *Dataset*

A distribuição do *dataset* foi realizada em `Python` e pode ser consultada na Sec.4. O objetivo deste passo foi analisar a distribuição geral do *dataset* para verificar se os dados seguiam padrões esperados e se iam ao encontro dos requisitos de uma *base de dados* realista.

- `CreditScore`: Segue uma distribuição aproximadamente normal entre 400 e 850.
- `Geography`: A maioria dos clientes vive na França, seguida pela Espanha e pela Alemanha.
- `Gender`: A variável está balanceada entre homens e mulheres.
- `Age`: Apresenta uma distribuição assimétrica, com maior concentração de clientes entre os 25 e 50 anos, aproximadamente.

- `Tenure`: Está distribuído de forma uniforme, sem grandes picos.
- `Balance`: Um número significativo de clientes tem saldo nulo, enquanto os restantes possuem valores distribuídos.
- `NumOfProducts`: A maioria dos clientes possui 1 ou 2 produtos, enquanto valores mais elevados, como 3 ou 4 produtos, são menos frequentes.
- `EstimatedSalary`: É uniforme, indicando ampla diversidade entre os clientes.

1.1.2 Sanitização

A análise da conformidade e da qualidade dos dados foi, também, realizada em `Python` e pode ser consultada na Sec.4. Desta forma, verificamos que o *dataset* original não contém valores em falta e que todos os valores estão em conformidade com o formato esperado. Também foi verificado que não existem registos duplicados, logo cada cliente só aparece uma vez no *dataset*. Assim, concluímos que o *dataset* está pronto para ser processado no `ARX` sem ser necessário realizar sanitização.

1.2 Objetivo de Divulgação do *Dataset*

O objetivo definido para divulgar o *dataset* anonimizado é permitir a análise da relação entre a faixa etária e o salário, sem comprometer a privacidade dos clientes. Na prática, deseja-se possibilitar estudos sobre como a idade influencia a média salarial dos clientes, garantindo que possíveis atacantes não consigam reidentificar indivíduos específicos.

A análise estatística inicial dos dados, também, foi realizada em `Python` e pode ser consultada na Sec.4. A Tab.1 apresenta os resultados obtidos antes da anonimização:

Faixa Etária	**Média do Salário**
18 - 25	102,093.90
25 - 30	100,637.99
30 - 35	99,171.20
35 - 40	99,540.70
40 - 45	101,615.20
45 - 50	102,793.88
50 - 55	99,767.00
55 - 93	96,209.07

Tabela 1: Média dos salários por faixa etária.

Os intervalos escolhidos têm em conta a distribuição de `Age` e correspondem aos intervalos iniciais escolhidos para a hierarquia deste atributo, na Sec.2.4.

1.3 Requisitos de Privacidade

Nesta secção definem-se os requisitos de privacidade necessários para **garantir a proteção dos dados** dos clientes enquanto se **mantém a utilidade** do *dataset*.

1.3.1 Limite de Supressão

Inicialmente, decidiu-se que será aplicado um **limite de supressão** de aproximadamente **3%** dos registos. Este valor foi definido com base na necessidade de equilibrar a proteção da privacidade com a preservação da utilidade dos dados, uma vez que a supressão atua como uma **medida eficaz** para **eliminar *outliers*** ou **combinações de atributos** que possam expor os indivíduos. Na Sec.3, durante a anonimização, ir-se-á tentar cumprir esse limite e perceber se o mesmo é razoável.

1.3.2 *Coding Model*

Adicionalmente, é importante ter em conta o ***trade-off*** entre generalização e supressão ao aplicar técnicas de anonimização. A **generalização** permite **preservar uma maior quantidade de informação útil** nos dados. No entanto, essa técnica pode **não ser suficiente** para proteger registos que continuam a ser identificáveis mesmo após a generalização. Nestes casos, **a supressão torna-se necessária** para garantir a privacidade.

Dá-se inicialmente preferência à generalização, por preservar melhor a utilidade dos dados, recorrendo-se à supressão apenas quando esta não é suficiente para atingir os requisitos de privacidade definidos.

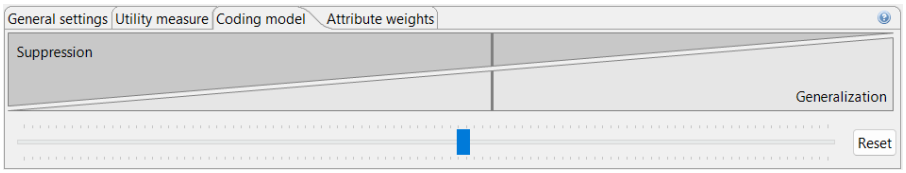


Figura 1: *Trade-off* entre generalização e supressão.

1.3.3 Peso de Atributos e Métrica de Utilidade

A métrica de utilidade escolhida para avaliar a qualidade da informação após anonimazação é a ***Loss***.

Para além disso, durante o processo de anonimização, será necessário atribuir diferentes pesos aos atributos. A **idade**, em particular, terá de receber um **peso mais elevado**, dado que está diretamente relacionada com o **objetivo principal** da divulgação do *dataset* e pretende-se que perda desta informação (*loss*) seja minimizada.

1.3.4 Modelos Selecionados

Os modelos a utilizar consistem em: ***k-Anonymity***, ***ℓ-Diversity*** e ***t-Closeness***. Os **parâmetros** de cada um **serão ajustados** conforme o necessário, mais uma vez, de modo a garantir equilíbrio entre proteção da privacidade e preservação da estrutura dos dados.

2 Caracterização do *Dataset* e *Coding Models*

2.1 Caracterização de Atributos

De modo a caracterizar cada um dos 14 atributos, inicialmente considerou-se as seguintes categorias:

- **Identificadores:** Atributos que identificam diretamente um indivíduo.

  Neste categoria, estão inseridos os atributos `CustomerId` e `Surname`.

- **Sensíveis:** Atributos que podem revelar informações privadas acerca de um indivíduo.

  Incluiu-se, nesta categoria, os atributos `CreditScore`, `Balance`, `NumOfProducts` e `EstimatedSalary`. Desta forma, atributos que revelem um *score* pessoal, valores de saldo bancário ou de salário e um número de produtos de um cliente são referentes a dados que um indivíduo desejaria manter como privados.

De seguida, as próximas categorias a considerar seriam atributos **Quasi-identificadores (QID)** e **Insensíveis**. Atributos que, de forma individual ou combinada, podem reidentificar um indivíduo e atributos que não comprometem a privacidade de um indivíduo, respetivamente.

Recorreu-se ao `ARX` para determinar os melhores candidatos a **QID** e, consequentemente, classificar os restantes como **Insensíveis**.

Na Fig.2, apresentam-se os resultados de **distinção** e **separação** de cada atributo. É notório que todos os valores de distinção são extremamente baixos, o que se deve ao baixo número de valores únicos de cada atributo no *dataset*. Posto isto, decidiu-se destacar inicialmente, com base nos altos valores de separação, os atributos `Tenure` e `Age`.

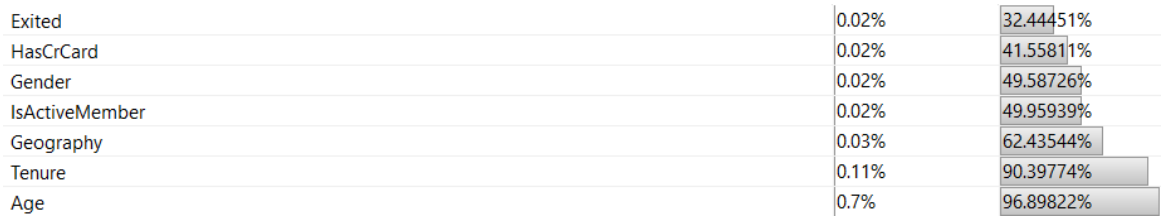


Figura 2: Análise individual dos valores de distinção e separação.

Na Fig.3, são observáveis os valores obtidos para atributos combinados em pares. Mais uma vez, os valores de distinção mantiveram-se reduzidos, no entanto, aumentaram nos pares destacados, principalmente para a combinação `Age` e `Tenure`.

Para além do par anterior, destacam-se os pares `Age` e `IsActiveMember`, `Age` e `HasCrCard`, `Gender` e `Age` e, por fim, `Geography` e `Age`, pois são os conjuntos de atributos com os maiores valores de distinção e separação conjugados. Para serem considerados, é importante observar o seu comportamento em combinações maiores e diferentes, como na Fig.4.

Figura 3: Análise dos valores de distinção e separação para pares de atributos.

Por fim, como se demonstra na Fig.4, ainda se consideraram conjuntos de três atributos.

Neste caso, a distinção apresenta valores mais elevados, pelo que se decidiu tomar estes valores como pontos decisivos na classificação dos atributos `Age`, `Tenure` e `Geography` como QID. Isto justifica-se pelos resultados anteriores de distinção e separação dos dois primeiros atributos e pelo aumento significativo de distinção do conjunto, quando é adicionado o atributo `Geography`. Mais ainda, é razoável considerar `Geography` como QID, pois, na Fig.3, ele está presente no segundo conjunto com maiores valores.

Age, Tenure, Exited	10.92%	99.77542%
Age, Tenure, IsActiveMember	11.48%	99.85126%
Age, Tenure, HasCrCard	11.52%	99.82574%
Gender, Age, Tenure	11.72%	99.84915%
Geography, Age, Tenure	15.95%	99.88705%

Figura 4: Análise dos valores de distinção e separação para trios de atributos.

Contudo, em relação aos atributos `Gender`, `IsActiveMember` e `HasCrCard`, foi necessário aprofundar a análise tendo em conta diferentes combinações.

Em primeiro lugar, considerou-se combinações com o atributo `Exited`, já que é o único a ser excluído dos candidatos a QID. Na Fig.5, é observável que a separação de qualquer combinação com os atributos `Gender` e `IsActiveMember` aumenta significativamente quando comparada com a do atributo `Exited` individualmente.

Exited	0.02%	32.44451%
HasCrCard	0.02%	41.55811%
Gender	0.02%	49.58726%
IsActiveMember	0.02%	49.95939%
HasCrCard, Exited	0.04%	60.45612%
Gender, Exited	0.04%	65.53113%
IsActiveMember, Exited	0.04%	65.68887%

Figura 5: Combinações dos candidatos a QID com o atributo `Exited`.

Para além disso, olhando para os resultados obtidos na Fig.6 através de combinações com um QID (com valores de distinção e separação não muito altos, inicialmente), o atributo `Geography`, conclui-se que os candidatos `Gender` e `IsActiveMember` podem trazer uma diferença significativa em comparação com um não QID.

Geography, Exited	0.06%	73.71137%
Geography, HasCrCard	0.06%	78.03453%
Geography, Gender	0.06%	81.0494%
Geography, IsActiveMember	0.06%	81.19772%

Figura 6: Combinações dos candidatos a QID com o atributo `Geography`.

Em suma, voltando a referir os resultados da Fig.4, em combinações com dois QID "fortes"(`Age` e `Tenure`), ainda se observam diferenças entre os cadidatos a considerar e o atributo que não é um QID.

Decidiu-se não se classificar o atributo `HasCrCard` como QID, dado o destaque dos restantes candidatos em relação ao mesmo e, no geral, tendo em conta que não se evidencia nas várias combinações.

Concluiu-se a análise com a classificação dos seguintes atributos como QID: `Geography`, `Gender`, `Age`, `Tenure`, `HasCrCard` e `IsActiveMember`.

Como já referido, os restantes atributos `RowNumber` e `Exited` foram classificados como Insensíveis.

2.2 Riscos de Privacidade do *Dataset*

Nesta subsecção, serão analisados os **riscos de privacidade** do *dataset*, na sua forma original.

Na Fig.7, o risco de reidentificação é calculado sob três modelos distintos de ataque – ***Prosecutor Attacker Model***, ***Journalist Attacker Model*** e ***Marketer Attacker Model***.

Cada modelo é avaliado em relação a três métricas principais – ***Records at risk***, ***Highest risk*** e ***Success rate***.

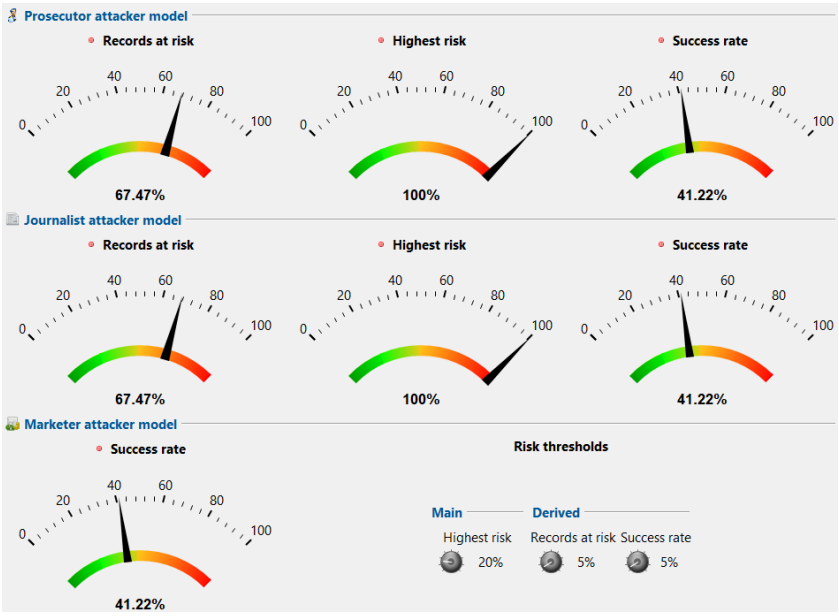


Figura 7: Análise dos riscos de privacidade do *dataset* original.

Os valores apresentados mostram que os riscos não variam entre modelos, com maior vulnerabilidade no cenário de ataque de ***Prosecutor*** e ***Journalist***, onde os registos em risco atingem 67.47% e o o risco máximo 100%.

2.3 Distribuição de Atributos

De forma a se decidir acerca da criação de hierarquias e pesos a atribuir a cada atributo, teve-se em conta a distribuição dos mesmos.

Excluindo atributos binários e um atributo com três valores únicos, foi importante verificar a distribuição de `Age` e `Tenure`.

- `Age`: Tal como referido em 1.1.1 e como se observa na Fig.8, este atributo concentra a maioria dos seus valores entre os 25 e 50 anos.

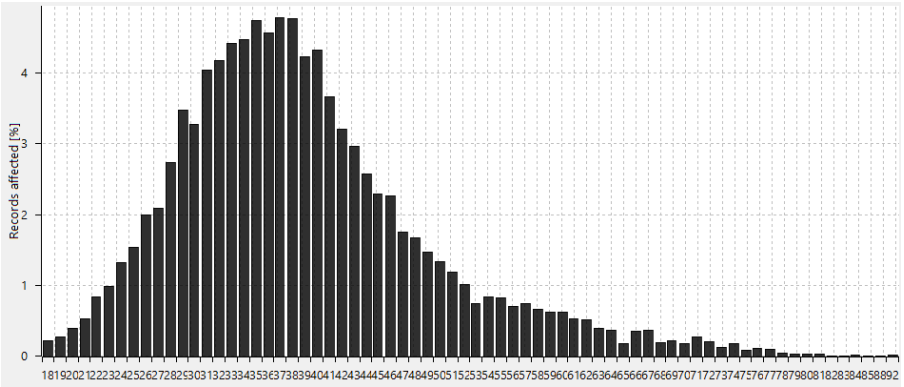


Figura 8: Distribuição do atributo `Age` (`ARX`).

- `Tenure`: Por sua vez, refere-se novamente que a distribuição do atributo `Tenure`, observável na Fig.9, apresenta um comportamento uniforme.

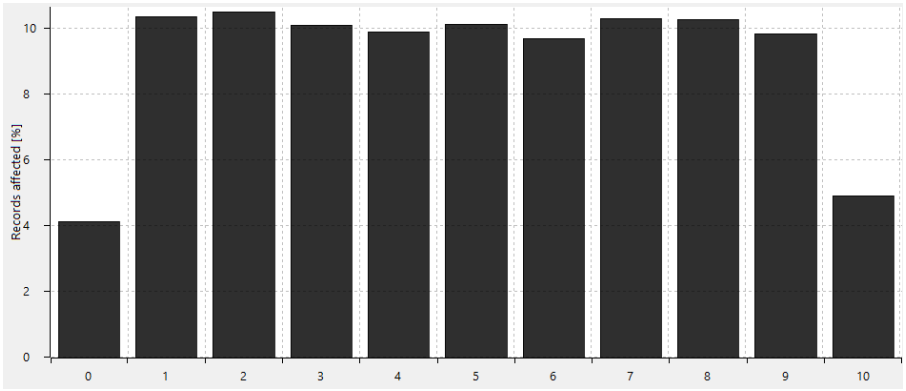


Figura 9: Distribuição do atributo `Tenure` (`ARX`).

2.4 Criação de Hierarquias

Através da subsecção 2.3, criaram-se as diferentes hierarquias para todos os QID.

Em primeiro lugar, para todos os atributos binários, ou seja, `Gender` e `IsActiveMember`, só existe uma forma de criar a hierarquia. Isto é, com um nível de generalização, que leva à supressão. As hierarquias criadas estão ilustradas na Fig.10.

Level-0	Level-1
Male	*
Female	*

Level-0	Level-1
0	*
1	*

Figura 10: Hierarquias criadas para os atributos `Gender` e `IsActiveMember`.

Para o atributo `Geography`, cujos valores únicos são *Spain*, *France* e *Germany*, agrupou-se, primeiramente, em *Southern Europe* e *Western Europe* e, finalmente, em *Europe*.

Level-0	Level-1	Level-2
France	Western Europe	Europe
Germany	Western Europe	Europe
Spain	Southern Europe	Europe

Figura 11: Hierarquia criada para o atributo `Geography`.

Para o atributo `Tenure`, dado que a sua distribuição é normal, decidiu-se utilizar uma hierarquia por intervalos, como ilustrada na Fig.12.

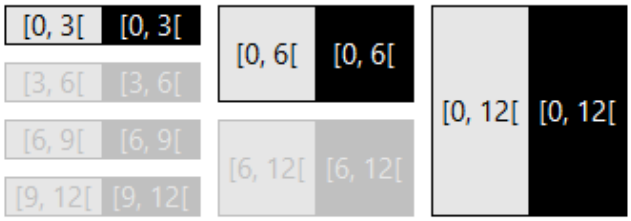


Figura 12: Hierarquia criada para o atributo `Tenure`.

Por fim, a hierarquia criada para o atributo `Age` teve em conta a sua distribuição não uniforme. Achou-se razoável iniciar com um intervalo que contivesse as idades até aos 25 anos,

alguns intervalos que contivessem as idades compreendidas entre os 25 anos e os 50 anos, um intervalo dos 50 aos 55 anos e, finalmente, um intervalo que englobasse as idades acima de 55. Para os seguintes níveis, juntaram-se intervalos dois a dois, como ilustra a Fig.13.

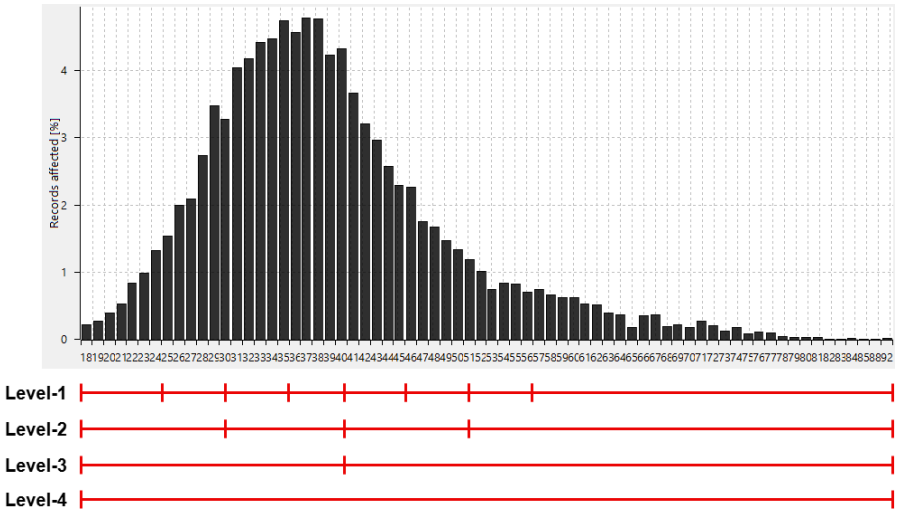


Figura 13: Hierarquia criada para o atributo `Age`.

2.5 Peso de Atributos

Como referido em 1.3.3, de modo a garantir uma boa utilidade dos dados após a anonimização, e considerado o objetivo proposto, decidiu-se atribuir um peso igual a 1 ao atributo `Age`. Esta decisão deve-se ao facto de se tratar de um forte QID, sendo muito provável que todos os modelos venham a aplicar algum nível de generalização sobre este atributo.

3 Modelos de Privacidade: Utilidade, Privacidade e Avaliação de Risco

3.1 *k-Anonimity* e ℓ*-Diversity*

A primeira abordagem que se tomou combina os modelos ***k-Anonimity*** e ℓ***-Diversity***. Os principais parâmetros a ajustar são os valores de $\boldsymbol{k}$ e ℓ.

3.1.1 Escolha de k (para $\ell = 2$)

Dado que as combinações de possíveis valores de ℓ são bastantes, iniciou-se a análise com um ℓ **fixo** ($\ell = 2$), de modo a **interpretar o comportamento de diferentes valores de** $\boldsymbol{k}$.

Para os valores iniciais, $\boldsymbol{k = 5}$, **10** e **15**, como se observa na Fig.14, foram consideradas as seguintes métricas:

- ***Highest Risk*** e ***Success Rate***;
- ***Loss Score***: Perda (*loss*) de informação;
- ***N.-U. Entropy***: Qualidade dos dados tendo em conta a semelhança entre a distribuição original e após anonimização;
- ***Discernibility***: Qualidade dos dados tendo em conta o tamanho das classes de equivalência, penalizando a supressão;

- Qualidade dos dados tendo em conta ***Attribute/Record-level Squared Error***;

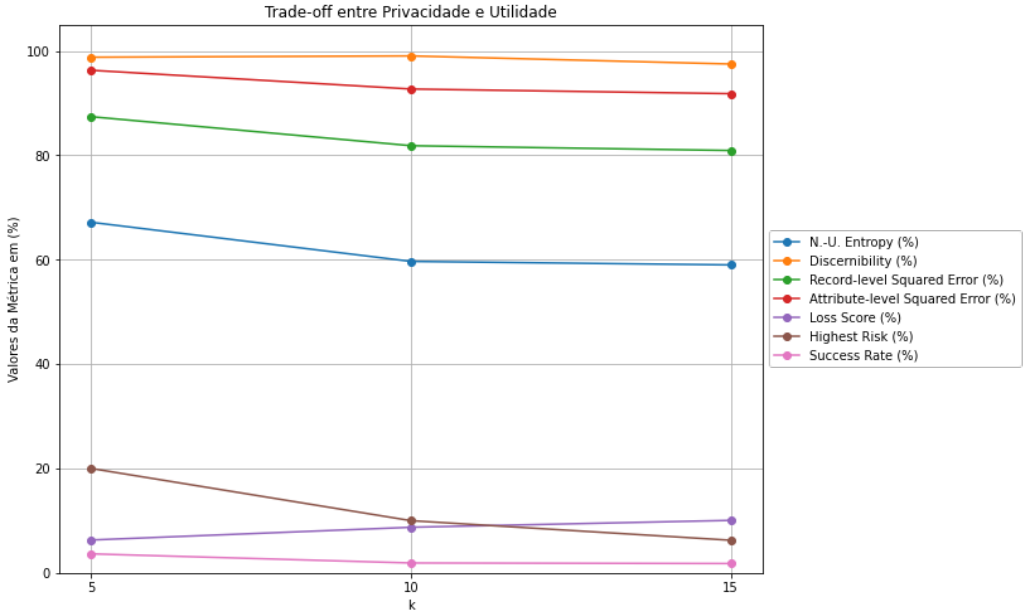


Figura 14: Análise do valor de k, considerando a transformação ótima.

Em relação a todas as métricas de utilidade, é visível que apresentam **valores piores** quando se passa de $\boldsymbol{k = 5}$ para $\boldsymbol{k = 10}$. No entanto, a **descida dos riscos de reidentificação** é muito relevante.

Já na passagem de $\boldsymbol{k = 10}$ para $\boldsymbol{k = 15}$, as únicas alterações que podemos destacar são a **descida considerável** de ***Highest Risk*** (desce de 10% para 6.25%) e um **pequeno aumento** de ***Loss Score*** (de 8.74% para 10.06%).

Posto isto, decidiu-se que poderia ser vantajoso continuar a analisar os valores de k acima de $\boldsymbol{k = 15}$, nomeadamente até $\boldsymbol{k = 20}$, de modo a se perceber como se comportam os riscos de reidentificação e as métricas de utilidade.

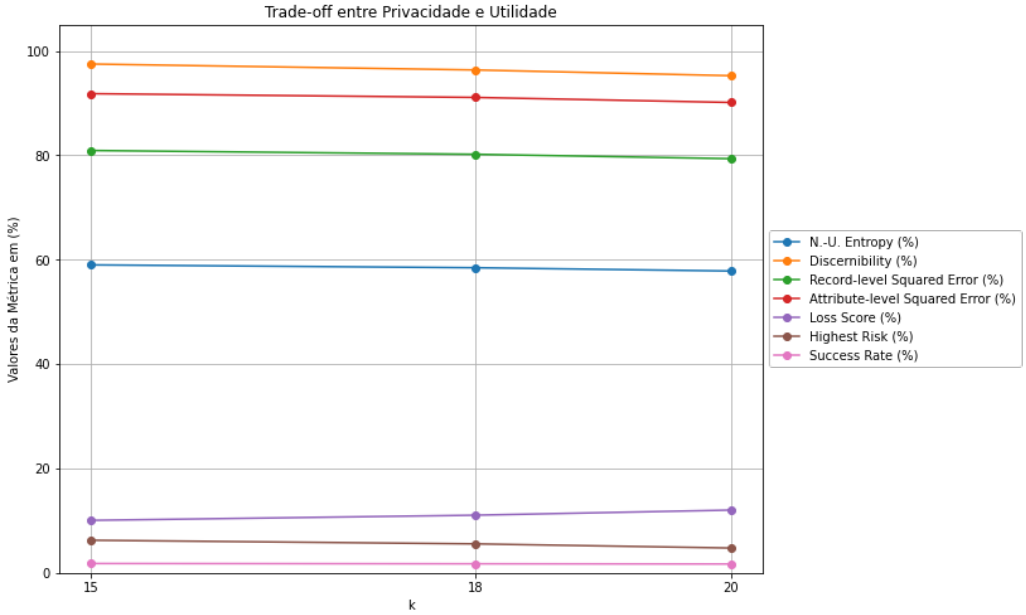


Figura 15: Análise de outros valores de k, considerando a transformação ótima.

Na Fig.15, estão representados os resultados obtidos para $\boldsymbol{k = 15, 18}$ e **20**. Como já era expectável, com o **aumento do valor de $\boldsymbol{k}$**, observa-se o **aumento ligeiro** de ***Loss Score*** e alguma **diminiução** de ***Highest Risk***.

Observou-se que $\boldsymbol{k = 18}$ proporcionava riscos de reidentificação relativamente baixos, sem prejudicar demasiado a utilidade dos dados na generalidade. Para além disso o valor de ***Suppressed Records*** é bastante mais reduzido do que para valores de k maiores.

3.1.2 Escolha de ℓ

Neste passo, foi necessário relembrar a percentagem de **valores únicos dos atributos sensíveis**, calculada em Python, que se na Sec.4.

- Os atributos `Balance` e `EstimatedSalary` têm uma **grande percentagem** de valores únicos, por isso admitem valores altos de ℓ;
- O atributo `CreditScore` tem **460** valores únicos;
- O atributo `NumOfProducts` tem apenas **4** valores únicos, portanto o seu valor de ℓ não poderá ser muito elevado.

Para diversos valores de k, percebeu-se que as transformações propostas e os riscos de reidentificação **não se alteravam**, considerando **valores baixos de ℓ**. Para além disso, verificou-se que o **fator com mais peso** nos resultados da anonimização era o **valor de ℓ** atribuído ao atributo sensível `NumOfProducts`, o que se explica pelos seus reduzidos valores únicos.

Como se consegue observar na Fig.3.1.2, experimentaram-se várias combinações de diferentes valores de ℓ. Considere-se o tuplo (k, ℓ-`EstimatedSalary`, ℓ-`Balance`, ℓ-`CreditScore`, ℓ-`NumOfProducts`). As combinações apresentadas são **(5, 4, 4, 3, 2 / 3)**, **(10, 9, 7, 6, 2 / 3)**, **(15, 14, 10, 8, 2 / 3)**, **(18, 17, 14, 10, 2 / 3)** e **(20, 19, 15, 12, 2 / 3)**, sendo que os valores de ℓ foram escolhidos tendo em conta as conclusões acerca do **número de valores únicos** de cada atributo.

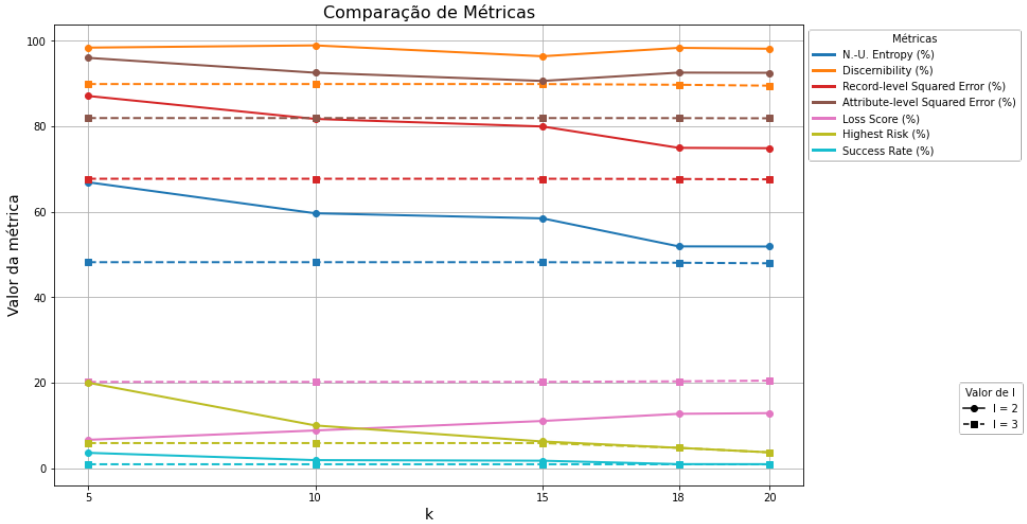


Figura 16: Análise dos valores de ℓ, considerando a transformação ótima.

É possível concluir-se que se **perde bastante utilidade** quando é considerado $\boldsymbol{\ell = 3}$, não sendo compensada pela diminuição dos riscos de reidentificação. Na verdade, a partir de $\boldsymbol{k = 15}$, os valores de ***Highest Risk*** são bastante semelhantes.

Apesar de em 3.1.1 se ter considerado $\boldsymbol{k = 18}$ a melhor solução, neste caso, atribuindo diferentes combinações de ℓ ao modelo, conseguiu-se que $\boldsymbol{k = 20}$ apresenta-se valores de ***Highest Risk*** e ***Success Rate*** ainda mais baixos, nomeadamente, **3.70%** e **0.94%** e, ainda, uma percentagem muito baixa de ***Suppressed Records*** de **0.38%**.

3.1.3 Resultados

Por fim, escolhendo $\boldsymbol{k = 20}$ e ℓ **= 19, 15, 12** e **2** para `EstimatedSalary`, `Balance`, `CreditScore` e `NumOfProducts`, respetivamente, é aplicada uma transformação que generaliza o atributo `Age` com **Nível 2** e o atributo `Tenure`, igualmente, com **Nível 2**, , sem generalizar os restantes.

Para além disso, conseguiram-se os seguintes valores de **métricas de utilidade** (Tab.2) e **riscos** (Fig.17):

Métrica	**(20, 19, 15, 12, 2)**
N.-U. entropy (%)	51.88
Discernibility (%)	98.17
Record-level squared error (%)	74.89
Attribute-level squared error (%)	92.54
Loss score (%)	12.89
Suppressed Records (%)	0.38

Tabela 2: Resultados obtidos para o tuplo (20, 19, 15, 12, 2).

Figura 17: Riscos de reidentificação do modelo.

Faixa Etária	**Média Salário (Anonimizado)**	**Média Salário (Original)**
[18, 30[	100,855.25	101,365.94
[30, 40[	98,616.32	99,355.95
[40, 50[	103,543.08	102,204.54
[50, 93[	96,803.87	97,988.03

Tabela 3: Comparação entre salários médios do *dataset* anonimizado e original por faixa etária.

Na Tab.3, são apresentados os resultados obtidos no recalculo da estatística inícial, isto é, a média de salários por faixa etária.

É possível observar que a anonimização **manteve a utilidade dos dados**, com **mínimas diferenças salariais** entre faixas etárias, garantindo **equilíbrio** entre privacidade e qualidade. Iremos avaliar o desempenho do próximo modelo.

3.2 *k-Anonimity* e *t-Closeness*

A segunda abordagem que se tomou combina os modelos ***k-Anonimity*** e ***t-Closeness***. Os principais parâmetros a ajustar são os valores de k e t.

3.2.1 Escolha de t

Inicialmente, procurou-se compreender o comportamento das métricas ao longo do intervalo de $\boldsymbol{k}$ entre **5** e **15**, para os valores de $\boldsymbol{t}$ considerados: **0.1, 0.15** e **0.2**. O objetivo desta análise preliminar foi identificar quais são as combinações de t e k que oferecem um melhor *trade-off* entre privacidade e utilidade, de forma a aprofundar a análise nestas combinações. Depois de encontrar o t mais relevante, vamos então testar mais valores de k, em 3.2.2.

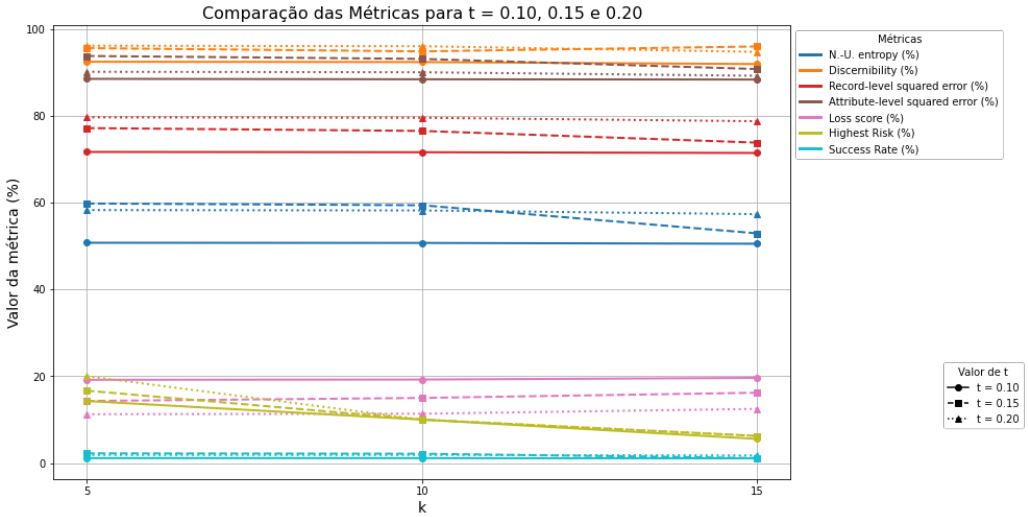


Figura 18: Valor das métricas para combinações de k e t.

Primeiramente, observa-se que o ***Highest Risk*** **diminui significativamente** entre $\boldsymbol{k = 5}$ e $\boldsymbol{k = 10}$, para todos os valores de t. Neste intervalo, os valores das **métricas de utilidade mantêm-se** relativamente estáveis, sem grandes variações, também, para todos os valores de t. Assim, a redução do risco de privacidade justifica-se, uma vez que é alcançada sem comprometer a utilidade dos dados.

No intervalo entre $\boldsymbol{k = 10}$ e $\boldsymbol{k = 15}$, verifica-se uma continuação da tendência de redução do ***Highest Risk*** para todos os valores de t. No entanto, neste caso, já se observam **alterações significativas** nas métricas de **utilidade** para cada t.

Assim, para $\boldsymbol{k = 10}$, observa-se que o valor do ***Highest Risk*** é de **10%** para todos os valores de t. No entanto, o caso com $\boldsymbol{t = 0.10}$ apresenta sistematicamente uma **pior utilidade** em todas as métricas, pelo que **não será considerado** uma opção viável.

Verifica-se, também, que é em $\boldsymbol{k = 15}$ que se atinge o **menor valor de risco**. Neste cenário, o valor $\boldsymbol{t = 0.20}$ destaca-se por manter **melhores resultados** na maioria das métricas de **utilidade**, com exceção da ***Discernibility*** e do ***Attribute-level Squared Error***, onde apresenta **ligeiras desvantagens** face ao $\boldsymbol{t = 0.15}$. Ainda assim, importa referir que o ***Loss***

Score associado ao $\boldsymbol{t = 0.20}$ é consideravelmente **inferior**. Deste modo, conclui-se que o valor de $\boldsymbol{t = 0.20}$ representa a melhor escolha.

Considerou-se que a configuração com $\boldsymbol{k = 15}$ e $\boldsymbol{t = 0.20}$ já representava uma boa opção, uma vez que apresentava valores reduzidos de risco: **6,25%** no ***Highest Risk*** e **1,76%** no ***Success Rate*** e **bons valores na utilidade**. No entanto, dado que $k = 15$ corresponde ao maior valor testado inicialmente, decidiu-se estender a análise para $\boldsymbol{k = 20}$, com o objetivo de **observar a evolução** das métricas para valores superiores.

Paralelamente, optou-se também por testar os valores **0.25**, **0.30** e **0.35** (os resultados de 0.30 são iguais aos de 0.35 para cada k, logo só vão ser representados os valores para 0.30) de t, de forma a comparar diretamente com o $\boldsymbol{t = 0.20}$.

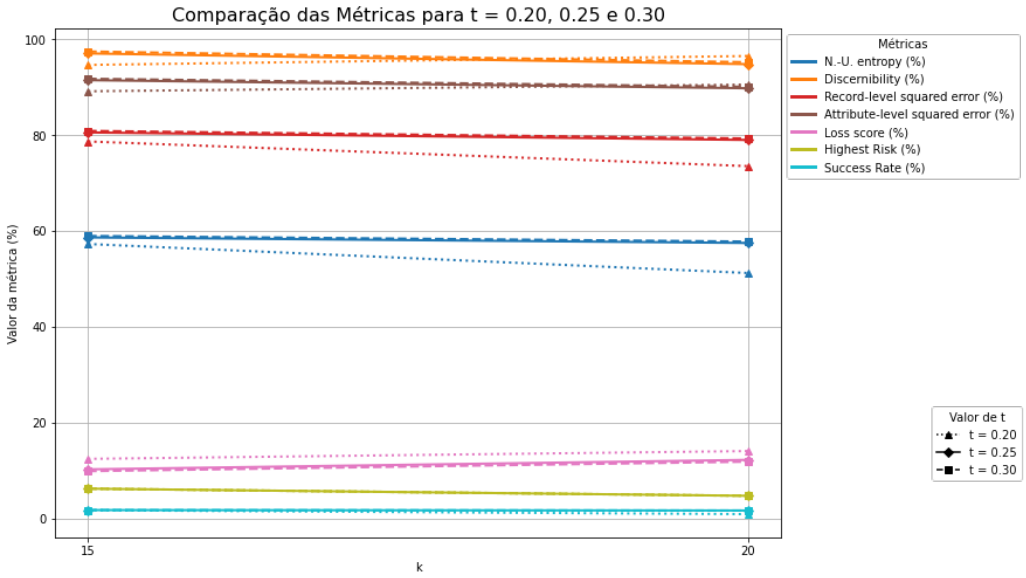


Figura 19: Valor das métricas para combinações de k e t

Observa-se que os **riscos** de privacidade se mantêm praticamente **inalterados** entre os diferentes valores de t, tanto no ***Highest Risk*** como no ***Success Rate***. No entanto, nas métricas de utilidade, o valor $\boldsymbol{t = 0.30}$ destaca-se por apresentar os melhores resultados na maioria das métricas, com exceção da ***Discernibility***, que sofre uma ligeira diminuição. Além disso, o ***Loss Score*** associado ao $\boldsymbol{t = 0.30}$ é **inferior** ao dos restantes, apesar da tendência de crescimento desta métrica entre $k = 15$ e $k = 20$. Concluiu-se, então, que o melhor valor de t será **0.30**.

Verifica-se ainda que, nesse mesmo intervalo de k, a ***Discernibility*** tende a **diminuir**, enquanto a quantidade de **supressão** aplicada **aumenta** consideravelmente — passa de **1,75%** em $\boldsymbol{k = 15}$ para **3,99%** em $\boldsymbol{k = 20}$ – valor acima do limite proposto nos requisitos de privacidade. Assim, podemos inferir que valores de k demasiado próximos de 20 não são vantajosos, dado o aumento significativo da supressão e o consequente impacto na utilidade dos dados. Um ponto de **equilíbrio** mais adequado deve situar-se entre $\boldsymbol{k} = 15$ e $\boldsymbol{k = 18}$.

3.2.2 Escolha do k

Será realizada a análise com valores de $\boldsymbol{k}$ entre **15** e **18**, mantendo $\boldsymbol{t = 0.30}$, com o objetivo de identificar com maior precisão qual o valor de k que proporciona o melhor **equilíbrio** entre utilidade e privacidade. Os valores de 16, 17 e 19 mostraram-se muito iguais aos de 15 e 18, portanto vão ser só representados esses, na Tab.4.

Atributo	$k = 15$	$k = 18$
N.-U. entropy (%)	59.00	58.46
Discernibility (%)	97.50	96.36
Record-level squared error (%)	80.93	80.20
Attribute-level squared error (%)	91.81	91.09
Loss score (%)	9.90	10.89
Highest Risk (%)	6.25	5.55
Success Rate (%)	1.79	1.74
N.-U. entropy Age (%)	34.61	34.27
Suppressed Records (%)	1.75	2.89

Tabela 4: Comparação dos atributos de privacidade e utilidade entre $k = 15$ e $k = 18$.

Após a análise comparativa dos resultados, optou-se pela escolha do $\boldsymbol{k = 18}$. Embora o valor $\boldsymbol{k = 15}$ apresente **melhores resultados** nas métricas de **utilidade**, a melhoria observada em termos de **privacidade** com $\boldsymbol{k = 18}$ foi considerada **mais relevante** para os objetivos da anonimização. Além disso, a **perda de utilidade** associada mostrou-se **pouco significativa**, permanecendo dentro de limites aceitáveis (por exemplo, um ***Loss Score*** de **10.89%** e **supressão** inferior a **3%**). Assim, a escolha de $\boldsymbol{k = 18}$ reflete um compromisso equilibrado entre utilidade e privacidade, privilegiando uma ligeira **melhoria na segurança sem comprometer a qualidade** dos dados.

3.2.3 Resultados

Para finalizar, o melhor **equilíbrio** entre entre privacidade e utilidade foi conseguido com $\boldsymbol{k = 18}$ e $\boldsymbol{t = 0.30}$, e a transformação aplicada generaliza o atributo `Age` com **Nível 2** e o atributo `Tenure`, com **Nível 1**, sem generalizar os restantes.

Apresenta-se, na Fig.20, os **riscos de reidentificação** finais.

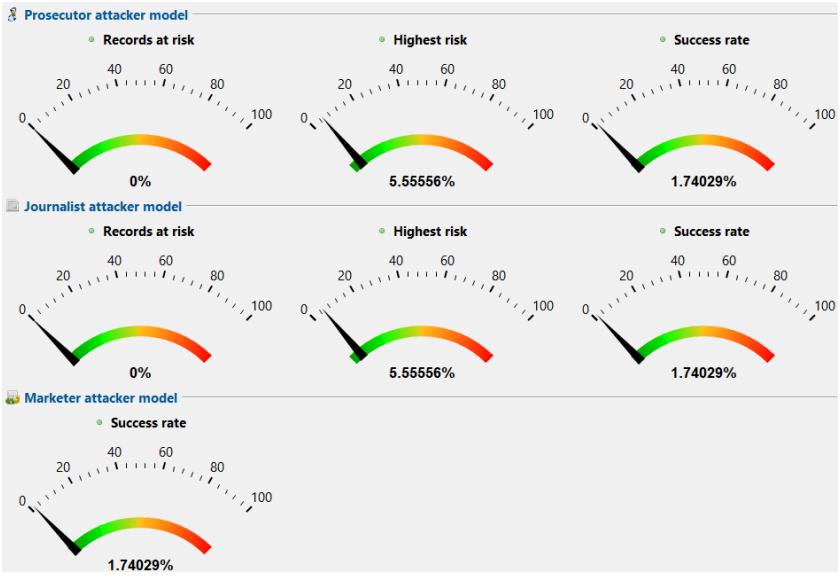


Figura 20: Riscos após da anonimização com 18-*anonimity* e 0.3-*closeness*.

Os resultados obtidos são, à primeira vista um pouco **contraintuitivos**, pois não se esperava que este modelo apresentasse riscos de privacidade maiores que o anterior. Contudo, este comportamento pode ser justificado pelo facto de o ***t-Closeness*** se focar na **similaridade**

das distribuições, neste caso, com **alta variabilidade de valores sensíveis**, sem garantir necessariamente uma **diversidade** de valores sensíveis dentro de cada classe de equivalência, ao contrário do ***ℓ-Diversity***.

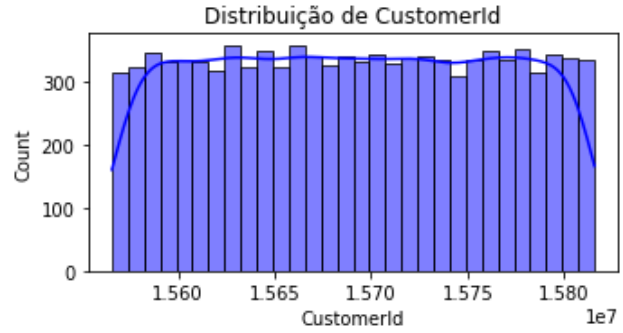
Faixa Etária	Média Salário (Anonimizado)	Média Salário (Original)
[18, 30[	101,338.05	101,365.94
[30, 40[	98,616.32	99,355.95
[40, 50[	103,526.27	102,204.54
[50, 93[	96,829.42	97,988.03

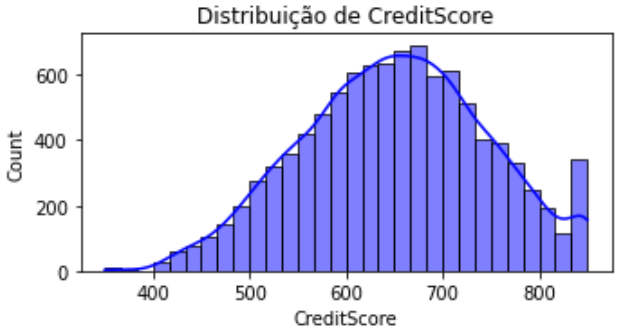
Tabela 5: Comparação entre salários médios do dataset anonimizado e original por faixa etária.

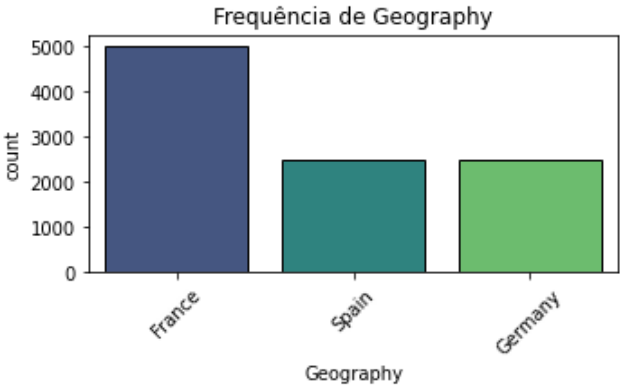
A comparação realizada entre os salários do *dataset* anonimizado e os valores reais demonstra que a anonimização **preservou** com eficácia a **utilidade dos dados**. As diferenças entre os valores são mínimas em todas as faixas etárias, o que indica que a aplicação das transformações de **anonimização não distorceu os dados**. Desta forma, confirma-se que é possível atingir um **equilíbrio** entre **proteção da privacidade** e **preservação da qualidade dos dados**.

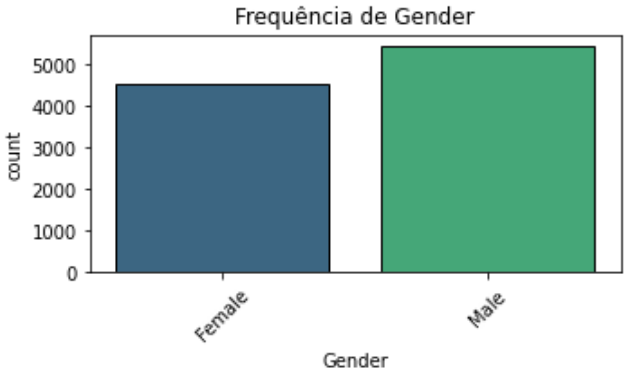
4 Anexos

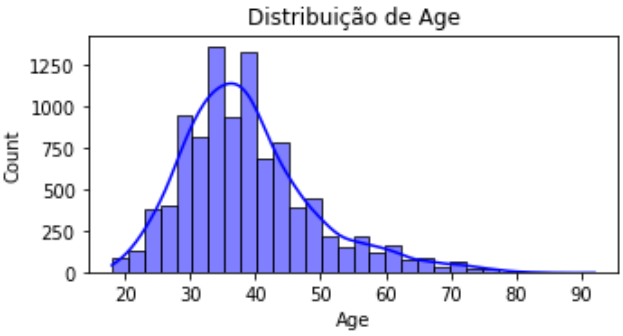
4.1 Distribuição do *Dataset*

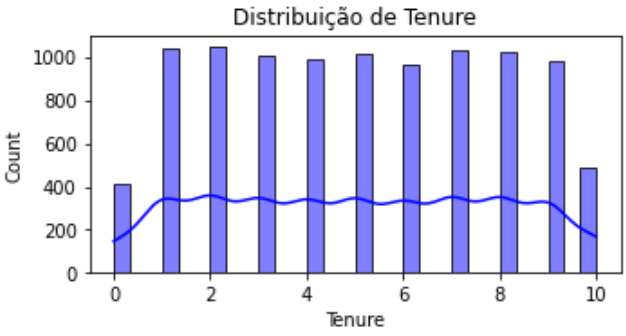


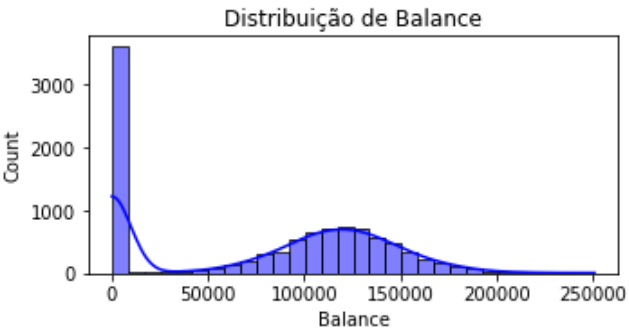


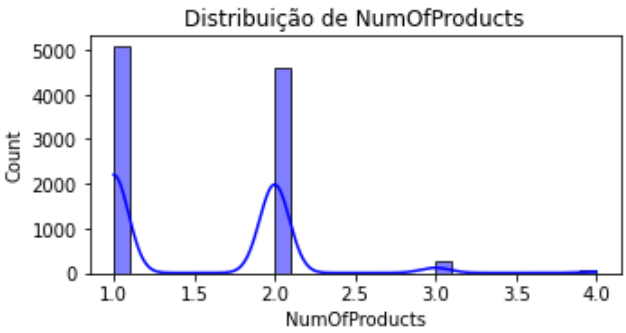


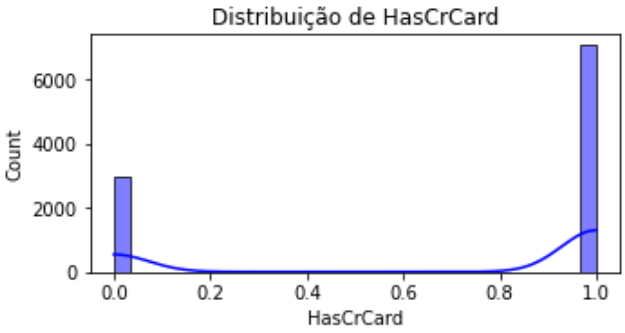


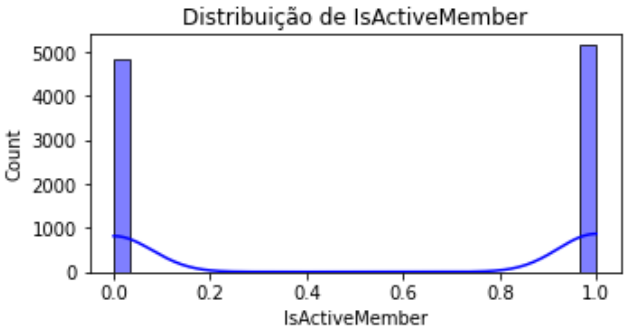


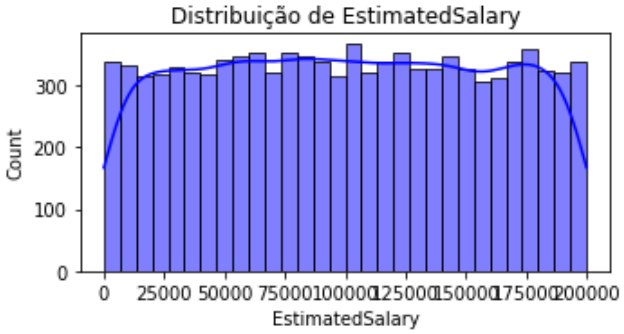


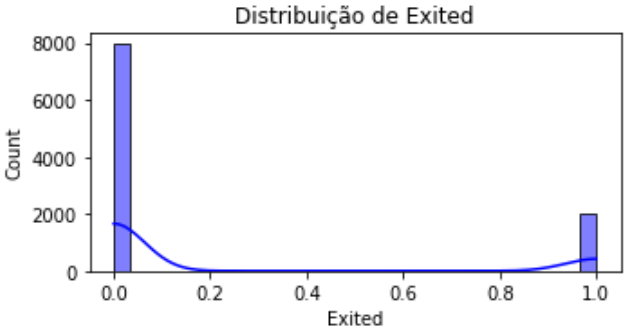





Distribuição de HasCrCard
Count
HasCrCard


Distribuição de IsActiveMember
Count
IsActiveMember


Distribuição de EstimatedSalary
Count
EstimatedSalary


Distribuição de Exited
Count
Exited

4.2 Qualidade e Conformidade dos Dados

```
#Contagem de Valores Nulos
data.isnull().sum()

data.isna().sum()
```

Coluna	Valores Nulos
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

Conforme é possível verificar, o *dataset* não contem valores nulos nem N/A, pelo que não existem valores em falta.

```
#Contagem de Valores Unicos
data.nunique()
```

Coluna	Valores Únicos
RowNumber	10000
CustomerId	10000
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2

```
#Valores Unicos de Geography
data['Geography'].unique()
```

```
array(['France', 'Spain', 'Germany'], dtype=object)
```

```
#Valores Unicos de Gender
data['Gender'].unique()
```

```
array(['Female', 'Male'], dtype=object)
```

```
#Valores Unicos de HasCrCard
data['HasCrCard'].unique()
```

```
array([1, 0])
```

```
#Valores Unicos de IsActiveMember
data['IsActiveMember'].unique()
```

```
array([1, 0])
```

```
#Valores Unicos de Exited
data['Exited'].unique()
```

```
array([1, 0])
```

```
#Valores Unicos de NumOfProducts
print(sorted(data['NumOfProducts'].unique()))
```

```
[1, 2, 3, 4]
```

```
#Valores Unicos de Tenure
print(sorted(data['Tenure'].unique()))
```

```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

```
#Contagem de Linhas Duplicadas
num_duplicados = data.duplicated().sum()
```

```
Número de registos duplicadas no dataset: 0.
```

Através da descrição dos dados anteriormente efetuada e da análise de valores únicos em cada coluna, também é possível constatar que:

- O intervalo de valores para `CreditScore` é [350, 850];
- Os valores dos atributos `Tenure` e `NumOfProducts`, ao variarem, respetivamente, de 0 a 10 e de 1 a 4, são razoáveis;
- As colunas categóricas não contêm erros de codificação, dado que `Geography` só toma os valores *France*, *Spain* e *Germany*, enquanto `Gender` só toma os valores *Male* ou *Female*;
- Os atributos binários assumem unicamente os valores 0 e 1, tal como esperado;
- Não existem registos duplicados, ou seja, cada cliente só aparece uma vez no *dataset*.

Tudo isto indica que não existem erros de medição nem inconsistências de valores/codificação nos dados constantes no *dataset* original.