FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Assignment 2: Privacy-Preserving Data Sharing

Segurança e Privacidade

Mestrado em Engenharia e Ciência de Dados

2022/2023

Duarte Emanuel Ramos Meneses – 2019216949 – duartemeneses@student.dei.uc.pt

Patrícia Beatriz Silva Costa – 2019213995 – patriciacosta@student.dei.uc.pt

Índice

Introdução

Cada vez mais a questão da privacidade dos dados é debatida em praça pública. Inclusive, muitas leis foram criadas com o intuito de combater esse problema. No entanto, que técnicas podem ser utilizadas para anonimizar os dados, aumentando a privacidade dos mesmos?

É com essa pergunta em mente que este trabalho prático procura analisar e avaliar a performance de três métodos distintos de fomentar a privacidade. Iremos experimentar modelos de privacidade na ferramenta ARX, *differential privacy* e ainda dados sintéticos.

Ao longo deste relatório explicaremos o processo de experimentação e analisaremos os resultados obtidos. Para cada *dataset* resultante, vamos realizar as mesmas análises que na meta 1 deste trabalho e ver o impacte de cada técnica utilizada. No fim, enunciaremos as vantagens e desvantagens de cada método.

Análises provenientes da meta 1

Ao longo deste relatório, iremos comparar as análises efetuadas na meta 1 com as realizadas com os *datasets* resultantes da anonimização.

- *<u>Análise 1:</u>*

Uma das análises que tínhamos efetuado na primeira meta envolvia a correlação. Sendo que um dos exercícios deste trabalho envolve *Differential Privacy*, esta métrica tornava-se pouco relevante do ponto de vista estatístico e muito exigente computacionalmente. Deste modo, optamos por substituir esta análise pela média dos valores por idade de cliente das colunas *past_avg_amount_annuity, past_avg_amt_application, past_avg_amt_credit e past_loans_total*. Os resultados foram os seguintes:

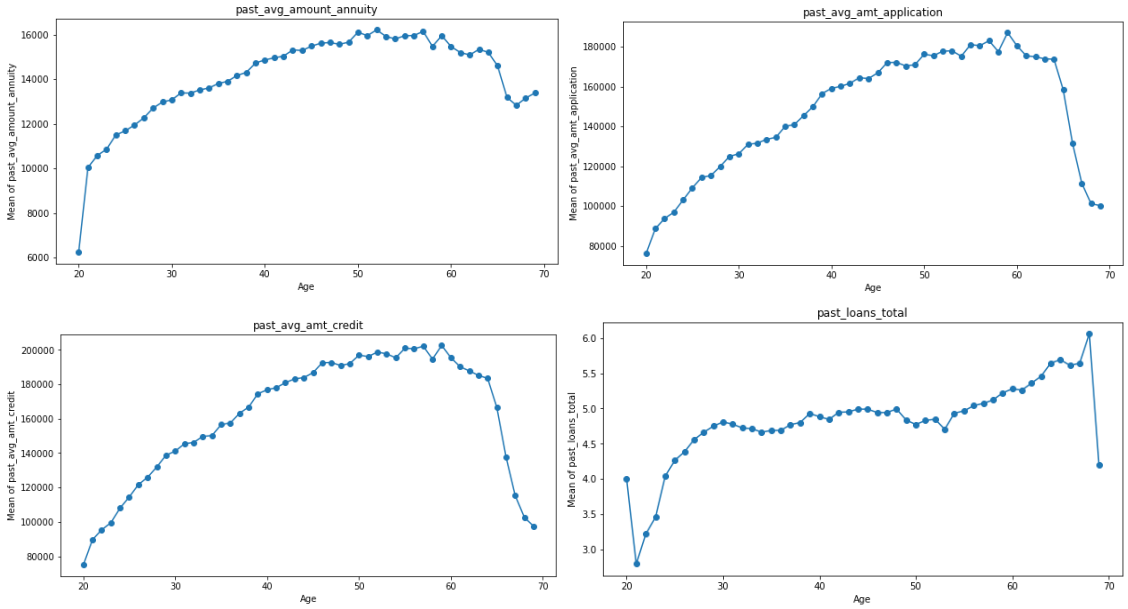


Figura 1 - Análise no dataset original da média dos past_event por idade dos clientes

Pelos gráficos acima, percebemos que existe um padrão semelhante em todos. Era de esperar que o montante aplicado e de crédito fosse aumentando com a idade uma vez que, quanto mais velha for uma pessoa, mais probabilidade tem de ter pedido empréstimos. Com isto, mais dinheiro pagou/recebeu. Esta tendência vê-se claramente no gráfico do total de empréstimos. À medida que a idade avança, o número de empréstimos aumenta.

No que diz respeito ao gráfico da anuidade, como o montante a pagar está de acordo com o rendimento do cidadão, percebemos claramente que existe um aumento no valor pago por ano na idade mais ativa.

Algo interessante a analisar é a curva descendente a partir dos 60 anos. Quanto à anuidade, explicamos essa redução uma vez que é normalmente a idade da reforma, pelo que os clientes começam a pagar menos. Já quanto aos outros gráficos, podemos deduzir que os empréstimos ganharam popularidade já depois destas pessoas terem necessitado para o que é comum (comprar casa, por exemplo). Deste modo, não têm tantos pedidos.

- *<u>Análise 2:</u>*

Decidimos manter a da primeira meta que consistia no impacte da idade dos clientes nos empréstimos em incumprimento. Para tal, analisamos a percentagem de pessoas com determinada idade com empréstimos em incumprimento. Os resultados foram os seguintes:

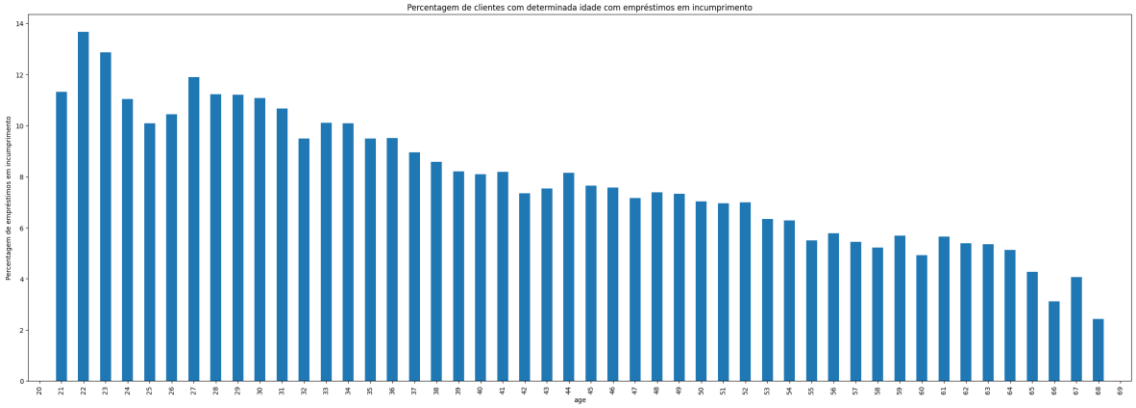


Figura 2 - Análise da percentagem de empréstimos em incumprimento por idade do dataset original

Tal como era expectável, quanto mais novos são os clientes, maior é a percentagem de empréstimos em incumprimento. Podemos explicar isto, uma vez que com a idade se vai ganhando estabilidade financeira. Posto isto, quanto mais velho for um cliente, teoricamente, mais capacidade tem de pagar um empréstimo.

1. Anonimização com modelos de privacidade

1.1. Caracterização do *dataset* através da classificação de atributos

Antes de caracterizar o *dataset* através da classificação de atributos, decidimos que devíamos remover algumas colunas que não continham informações relevantes para a análise dos empréstimos nem eram importantes para o estudo da privacidade. Deste modo, e com vista também a uma exigência computacional menor, removemos as seguintes colunas:

- 'car_age';
- 'days_employed';
- 'has_own_car';
- 'has_own_realty';
- 'housing_type';
- 'mobilephone_reachable';
- num_children';
- 'num_family_members';
- 'num_req_bureau_day';
- 'num_req_bureau_hour;
- 'num_req_bureau_month;
- 'num_req_bureau_qrt';
- 'num_req_bureau_week';
- 'num_req_bureau_year';
- 'occupation_type';
- 'provided_email';
- 'provided_homephone';
- 'provided_mobilephone';
- 'provided_workphone';
- 'region_rating';
- 'score_ext_1';
- 'score_ext_2';
- 'score_ext_3'.

O *dataset* continha o primeiro e último nome do cliente em duas colunas diferentes. Posto isto, criamos uma só intitulada de “nome” com a junção de ambos e eliminamos as outras duas.

Como algumas colunas apresentavam em certos casos valores nulos, decidimos que a melhor solução seria substituir esses valores por -1. Surgiu a ideia de eliminar as linhas que continham valores *None* mas, sendo que era só em alguns casos, iríamos perder imensa informação. Desta forma, conseguimos manter todos os dados, sendo possível identificar que valores eram inicialmente nulos.

Após este tratamento do *dataset*, classificamos os atributos:

Atributos	Classificação
loan_id	Identifying
infringed	Sensitive
gender	Quasi-identifying
annual_income	Quasi-identifying
income_type	Quasi-identifying
family_status	Quasi-identifying
age	Quasi-identifying
organization_type	Quasi-identifying
name	Identifying

Todas as outras variáveis foram classificadas como *Insensitive*.

1.2. Análise da distinção e separação dos QIDs

Quasi-identifier	Distinction	Separation
gender	0.00098%	44.98648%
family_status	0.00195%	55.38769%
income_type	0.0026%	64.17843%
age	0.01626%	97.69504%
organization_type	0.01886%	89.20774%
annual_income	0.82859%	94.41195%
gender, family_status	0.00455%	76.13336%
gender, income_type	0.00585%	80.58284%
income_type, family_status	0.01106%	83.70202%
gender, age	0.03349%	98.72032%
gender, organization_type	0.03902%	93.69274%
income_type, organization_type	0.06504%	92.78526%
income_type, age	0.0787%	98.90068%
family_status, age	0.08%	98.90373%
family_status, organization_type	0.09366%	95.31399%
age, organization_type	0.82989%	99.58376%
gender, annual_income	1.00484%	96.86392%
annual_income, income_type	1.2289%	97.84133%
annual_income, family_status	1.29459%	97.53041%
annual_income, organization_type	2.77746%	99.3786%
annual_income, age	3.44118%	99.86845%
gender, income_type, family_status	0.02081%	91.37435%
gender, income_type, organization_type	0.11967%	95.6132%
gender, income_type, age	0.14406%	99.37726%
gender, family_status, age	0.15772%	99.41035%
gender, family_status, organization_type	0.18146%	97.35096%
income_type, family_status, organization_type	0.26698%	96.92641%
income_type, family_status, age	0.32064%	99.4818%
gender, age, organization_type	1.54108%	99.73783%

gender, annual_income, income_type	1.54791%	98.80994%
gender, annual_income, family_status	1.62401%	98.6361%
income_type, age, organization_type	2.01098%	99.68405%
annual_income, income_type, family_status	2.09456%	99.02876%
family_status, age, organization_type	2.97745%	99.81697%
gender, annual_income, organization_type	3.79304%	99.63522%
annual_income, income_type, organization_type	4.2444%	99.58893%
gender, annual_income, age	4.72406%	99.92574%
annual_income, family_status, organization_type	5.53183%	99.73169%
annual_income, income_type, age	6.09799%	99.93424%
annual_income, family_status, age	6.90902%	99.93781%
annual_income, age, organization_type	18.16748%	99.97661%
gender, income_type, family_status, organization_type	0.47673%	98.20843%
gender, income_type, family_status, age	0.5912%	99.7156%
gender, annual_income, income_type, family_status	2.72966%	99.46923%
gender, income_type, age, organization_type	3.43728%	99.79197%
gender, family_status, age, organization_type	4.77186%	99.88944%
income_type, family_status, age, organization_type	5.8154%	99.8647%
gender, annual_income, income_type, organization_type	5.82711%	99.75014%
gender, annual_income, family_status, organization_type	7.45567%	99.8458%
annual_income, income_type, family_status, organization_type	8.28133%	99.82582%
gender, annual_income, income_type, age	8.57855%	99.96225%
gender, annual_income, family_status, age	9.51966%	99.96546%
annual_income, income_type, family_status, age	12.19924%	99.96921%
gender, annual_income, age, organization_type	23.64436%	99.9853%
annual_income, income_type, age, organization_type	25.22739%	99.98261%
annual_income, family_status, age, organization_type	29.6526%	99.98974%
gender, income_type, family_status, age, organization_type	8.61368%	99.91499%
gender, annual_income, income_type, family_status, organization_type	10.94693%	99.8973%
gender, annual_income, income_type, family_status, age	16.31259%	99.98261%
gender, annual_income, income_type, age, organization_type	31.45416%	99.9886%
gender, annual_income, family_status, age, organization_type	36.14082%	99.99376%
annual_income, income_type, family_status, age, organization_type	37.70889%	99.99259%
gender, annual_income, income_type, family_status, age, organization_type	44.51158%	99.99533%

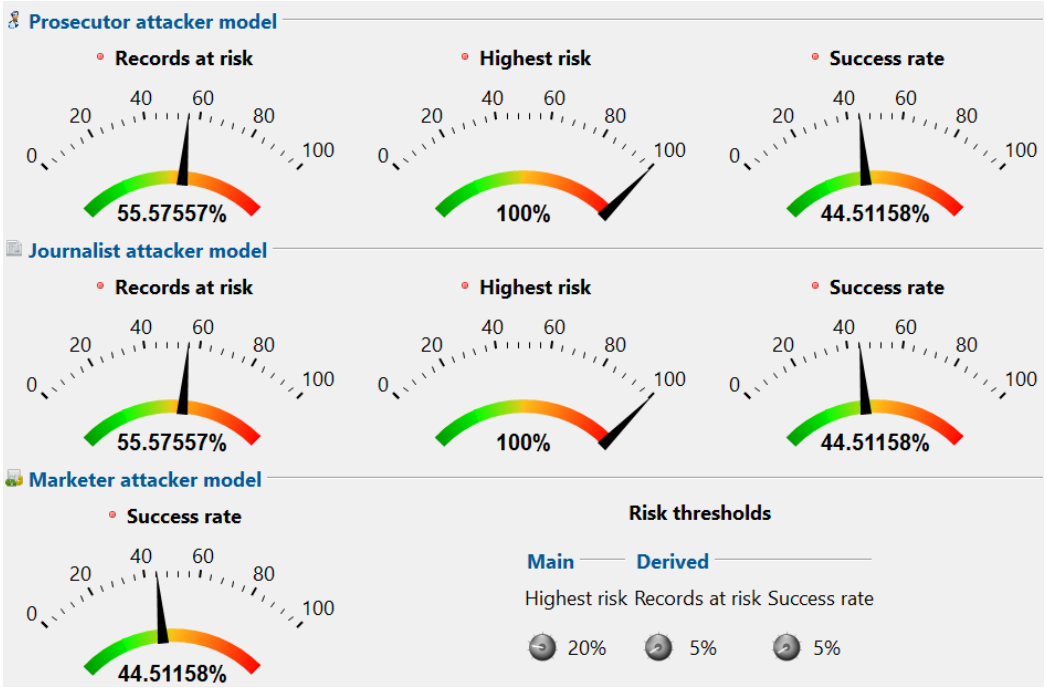
Figura 3 - Distinção e Separação dos QIDs do dataset original

Tal como era de esperar, o valor da distinção aumenta proporcionalmente ao número de QIDs juntos. No entanto, conseguimos ver que, com o mesmo número de QIDs, muitas vezes o valor da distinção baixa com o género. Isto deve-se ao facto de este atributo apenas possuir três valores, o que afeta drasticamente a distinção. Para esse valor ser elevado, os atributos combinados devem ter vários valores possíveis.

Quanto à separação, os valores apresentados no quadro acima são próximos dos 100% na maioria dos casos.

Uma vez que valores elevados de distinção e separação indicam possíveis QIDs, pensamos ter definido corretamente os quasi-identifiers do nosso dataset.

1.3. Medição dos riscos de privacidade do dataset na forma original



Measure	Value [%]
Lowest prosecutor risk	0.42918%
Records affected by lowest risk	0.07577%
Average prosecutor risk	44.51158%
Highest prosecutor risk	100%
Records affected by highest risk	31.43595%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	44.51158%
Sample uniques	31.43595%
Population uniques	2.99084%
Population model	PITMAN
Quasi-identifiers	age, annual_income, family_status, gender, income_type,...

Figura 4 - Medição dos riscos de privacidade do dataset na forma original

Pela imagem acima, fica evidente que os dados têm um elevado risco de privacidade, o que não pode acontecer. As secções seguintes irão apresentar técnicas de como reduzir este risco.

1.4. Hierarquia usada para os Quasi-Identifiers

Para incrementar o grau de privacidade dos nossos dados, decidimos criar hierarquias nos *QIDs*. Isto permite anonimizar os dados através de diversas formas.

1.4.1. Gender

Foi utilizado *ordering* de um nível pois consideramos que o género, sendo um atributo que contém poucos valores distintos, ou é apresentado ou não (anonimizando).

Level-0	Level-1
F	{F, M, XNA}
M	{F, M, XNA}
XNA	{F, M, XNA}

1.4.2. Annual Income

Por existirem valores iguais ao longo da coluna, apercebemo-nos que seria possível utilizar uma hierarquia de intervalos de 6 níveis. O primeiro nível contém um intervalo de 60 000 e os seguintes um tamanho de 2, 4, 8, 16, 32, respetivamente.

1.4.3. Income Type

Foi utilizado um *ordering* de três níveis. Podemos observar a junção de tamanho 2 em cada nível. Pensamos neste tipo de hierarquia uma vez que existem poucas letras em comum entre os valores possíveis e, como tal, esta é a forma mais eficaz de anonimizar os dados, agrupando-os.

Level-0	Level-1	Level-2	Level-3
Businessman	{Bu, Ca}	{Bu, Ca, Ml, Pe}	*
Commercial associate	{Bu, Ca}	{Bu, Ca, Ml, Pe}	*
Maternity leave	{Ml, Pe}	{Bu, Ca, Ml, Pe}	*
Pensioner	{Ml, Pe}	{Bu, Ca, Ml, Pe}	*
State servant	{Sta, St}	{Sta, St, Un, Wr}	*
Student	{Sta, St}	{Sta, St, Un, Wr}	*
Unemployed	{Un, Wr}	{Sta, St, Un, Wr}	*
Working	{Un, Wr	{Sta, St, Un, Wr}	*

1.4.4. Family Status

Foi utilizado um *ordering* de 2 níveis, onde podemos observar a junção de tamanho 3 no primeiro nível e de 2 no segundo. Tal como no *income_type* da secção anterior, optamos por esta hierarquia uma vez que existem poucas letras em comum entre os valores possíveis e, como tal, esta é a forma mais eficaz de anonimizar os dados, agrupando-os.

Level-0	Level-1	Level-2
Civil marriage	{Cm, Ma, Sp}	*
Married	{Cm, Ma, Sp}	*
Separated	{Cm, Ma, Sp}	*
Single / not married	{Snm, Un, Wi}	*
Unknown	{Snm, Un, Wi}	*
Widow	{Snm, Un, Wi}	*

1.4.5. Age

Para a idade, decidimos utilizar a hierarquia de intervalos. Os valores deste atributo variam entre 20 e 69, pelo que decidimos utilizar 3 níveis em que o primeiro corresponde a um intervalo de 10 valores (tamanho 1), o segundo de 20 (tamanho 2) e o terceiro de 80 (tamanho 4). Com isto, observamos que existe um agrupamento de todos os valores no último nível.

Observação: O nível 0 corresponde a valores sem qualquer agrupamento.

Level-0	Level-1	Level-2	Level-3
	[20,30[	[30,40[	[0,80[
	[30,40[	[30,40[	[0,80[
	[40,50[	[40,60[	[0,80[
	[50,60[	[40,60[	[0,80[
	[60,70[	[60,80[	[0,80[

1.4.6. Organization Type

Foi utilizado *masking*, de 22 níveis, sendo esse o tamanho máximo do tipo de organização. Esta hierarquia foi utilizada por haver letras em comum no início das diferentes organizações. Por uma questão de visualização, optamos por não apresentar o resultado, uma vez que a tabela contém imensos valores.

1.5. Requisitos de privacidade

Os pesos dos atributos foram mantidos a 0.5 e a *supression* foi limitada a 10%.

1.6. Modelos de privacidade

Como modelo de privacidade do dataset foi utilizado o L-diversity com L=2 para o atributo sensível "infringed" e K-Anonimity com K=12.

Escolhemos o L com valor 2 pois o atributo "infringed" contém dois valores distintos. Já no caso do K, testamos vários valores até que a medição de riscos apresentasse resultados aceitáveis.

a. Resultados

Distinção e separação dos QIDs

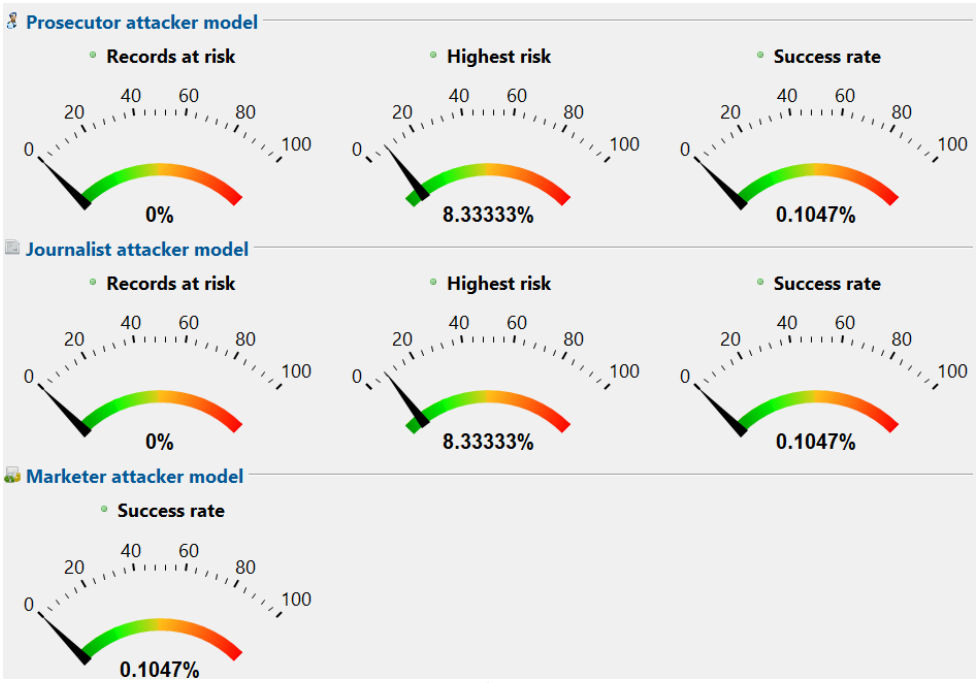
Quasi-identifier	Distinction	Separation
annual_income	0.00065%	1.10948%
organization_type	0.00065%	39.76064%
gender	0.00065%	44.95042%
income_type	0.00131%	64.0449%
family_status	0.00164%	55.1737%
age	0.00164%	78.24484%
annual_income, organization_type	0.00131%	40.33416%
gender, annual_income	0.00131%	45.45405%
gender, organization_type	0.00131%	65.72668%
income_type, organization_type	0.00229%	78.67908%
annual_income, income_type	0.00262%	64.39481%
gender, income_type	0.00262%	80.50354%
annual_income, family_status	0.00294%	55.81713%
annual_income, age	0.00294%	78.51633%
family_status, organization_type	0.00327%	72.99359%
gender, family_status	0.00327%	76.0097%
age, organization_type	0.00327%	86.84882%
gender, age	0.00327%	87.94468%
income_type, family_status	0.00654%	83.58871%
income_type, age	0.00654%	90.05272%
family_status, age	0.00818%	89.65527%
gender, annual_income, organization_type	0.00262%	65.99942%
annual_income, income_type, organization_type	0.00425%	78.86435%
gender, income_type, organization_type	0.00458%	88.03351%
annual_income, family_status, organization_type	0.00523%	73.32605%
gender, annual_income, family_status	0.00523%	76.30541%
gender, annual_income, income_type	0.00523%	80.66636%
annual_income, age, organization_type	0.00589%	86.9894%
gender, annual_income, age	0.00589%	88.07071%
gender, family_status, organization_type	0.00654%	85.1094%

gender, age, organization_type	0.00654%	92.44332%
annual_income, income_type, family_status	0.00916%	83.79541%
annual_income, income_type, age	0.01014%	90.15191%
annual_income, family_status, age	0.01112%	89.82313%
income_type, age, organization_type	0.01112%	93.82493%
income_type, family_status, organization_type	0.01145%	90.30549%
gender, income_type, family_status	0.01276%	91.31258%
gender, income_type, age	0.01309%	94.46507%
family_status, age, organization_type	0.01603%	93.7887%
gender, family_status, age	0.01603%	94.4477%
income_type, family_status, age	0.03108%	95.30692%
gender, annual_income, income_type, organization_type	0.00851%	88.12285%
gender, annual_income, family_status, organization_type	0.00982%	85.26961%
gender, annual_income, age, organization_type	0.01145%	92.51178%
annual_income, income_type, family_status, organization_type	0.01505%	90.41444%
gender, annual_income, income_type, family_status	0.01734%	91.4099%
annual_income, income_type, age, organization_type	0.01734%	93.87831%
gender, annual_income, income_type, age	0.01963%	94.51142%
annual_income, family_status, age, organization_type	0.02094%	93.87589%
gender, annual_income, family_status, age	0.02094%	94.52613%
gender, income_type, family_status, organization_type	0.02159%	94.71324%
gender, income_type, age, organization_type	0.02225%	96.41109%
gender, family_status, age, organization_type	0.03108%	96.55647%
annual_income, income_type, family_status, age	0.03697%	95.36839%
income_type, family_status, age, organization_type	0.04973%	97.12535%
gender, income_type, family_status, age	0.05758%	97.46502%
gender, annual_income, income_type, family_status, organization_type	0.02814%	94.76646%
gender, annual_income, income_type, age, organization_type	0.03239%	96.437%
gender, annual_income, family_status, age, organization_type	0.03861%	96.59917%
annual_income, income_type, family_status, age, organization_type	0.05857%	97.15838%
gender, annual_income, income_type, family_status, age	0.0674%	97.49408%
gender, income_type, family_status, age, organization_type	0.09161%	98.38783%
gender, annual_income, income_type, family_status, age, organization_type	0.1047%	98.40404%

Figura 5 - Distinção e separação do dataset anonimizado

Analisando os valores da distinção e separação da imagem acima com os apresentados na secção 2.2. deste relatório, percebemos que estes reduziram quando foram aplicadas técnicas de anonimização. Isto é positivo uma vez que, quanto menores forem os valores da distinção e da separação, menor é também o risco de existirem QIDs, o que garante mais privacidade no dataset.

Medição dos riscos de privacidade do dataset anonimizado



Measure	Value [%]
Lowest prosecutor risk	0.00687%
Records affected by lowest risk	4.76281%
Average prosecutor risk	0.1047%
Highest prosecutor risk	8.33333%
Records affected by highest risk	0.00393%
Estimated prosecutor risk	8.33333%
Estimated journalist risk	8.33333%
Estimated marketer risk	0.1047%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	age, annual_income, family_status, gender, income_type,...

Figura 6 - Medição dos riscos de privacidade do dataset anonimizado

Tal como se esperava, o risco de privacidade diminuiu. Isto comprova que as técnicas utilizadas funcionam e anonimizam de facto os dados. Tínhamos inicialmente mais de metade dos dados em risco e, neste momento, esse risco é nulo. O risco máximo era de 100% e agora é pouco mais de 8%.

Concluímos assim que os modelos de privacidade que implementamos resultaram.

Utilidade

Ao anonimizarmos os dados, perdemos alguma informação, uma vez que alguns são agrupados, outros são reduzidos aos seus primeiros caracteres,… Isto aumenta a privacidade mas, por outro lado, leva à perda de informação. Esta é uma questão que afeta a utilidade do dataset. Deste modo, concluímos que, apesar de termos incrementado a privacidade, a utilidade dos dados diminuiu.

1.8. Repetição das análises feitas na Meta 1

- ***Análise 1:***

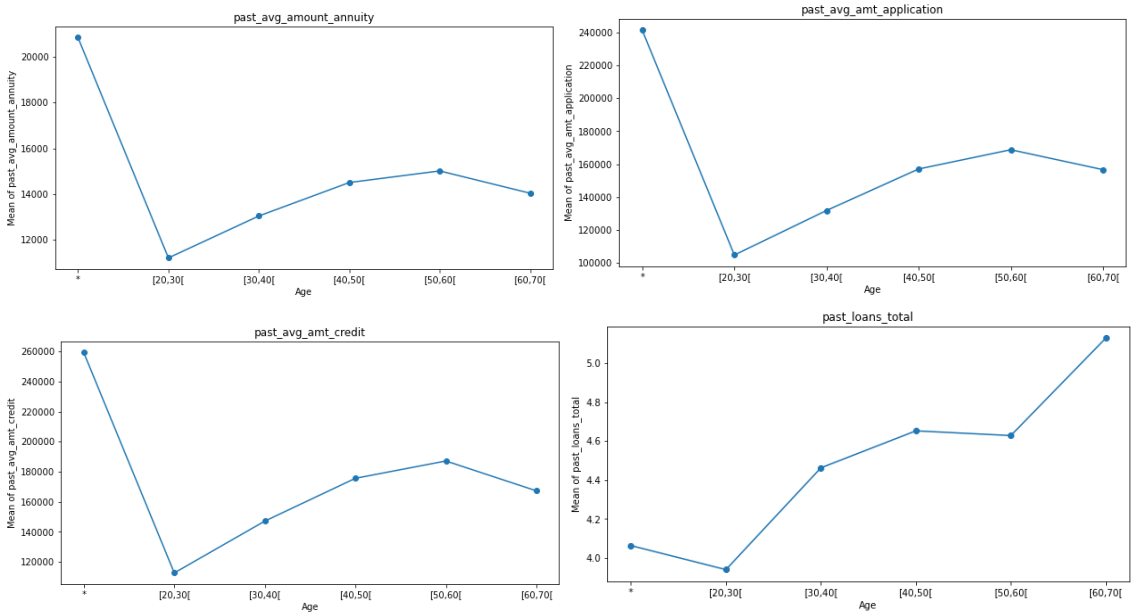


Figura 7 - Análise da média dos past_event por idade dos clientes no dataset com modelos de privacidade

Sendo que utilizamos hierarquia de intervalos para a anonimização da idade, era previsível que agora a análise aparecesse agrupada por intervalos. No entanto, a tendência das linhas em todos os gráficos mantém-se exatamente como na meta 1. Isto permite ter uma noção das médias de cada coluna por intervalo de idade, tendo os dados mais anonimizados do que nas análises presentes na secção 1 deste relatório. No entanto, a análise não é tão detalhada.

- ***Análise 2:***

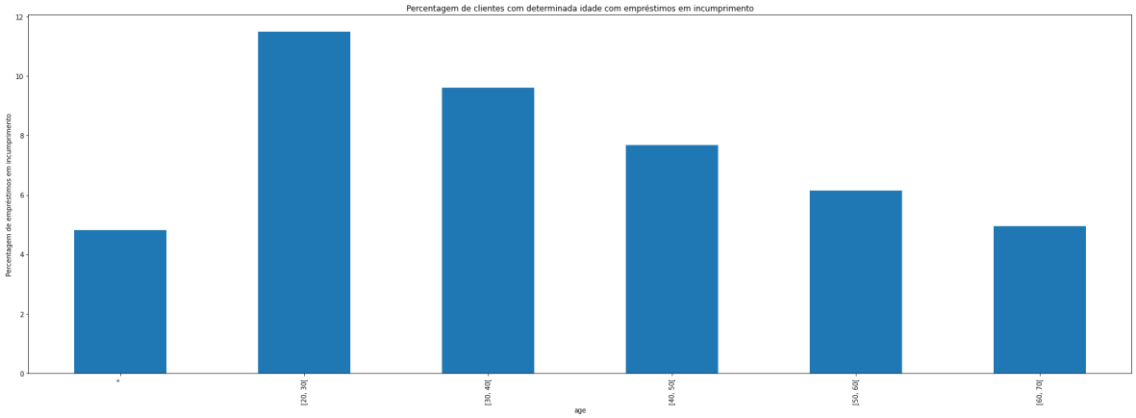


Figura 8 - Análise dos empréstimos infringidos por idade do dataset anonimizado

Tal como referimos anteriormente, era previsível que a análise aparecesse agrupada por intervalos de idade. Mais uma vez, a tendência de decréscimo das colunas à medida que a idade avança mantém-se. Dá para analisar consoante os intervalos de idade, pelo que não é tão detalhado.

Vantagens

- Anonimização dos dados, o que incrementa a privacidade.

Desvantagens

- Perda de informação;

- Caso se pretenda analisar detalhadamente, por exemplo, por idade, a anonimização pode dificultar esse processo já que por intervalos não se analisa tão bem;
- Redução da utilidade do *dataset*.

2. Differential Privacy

2.2. Sensibilidade nas duas análises pré *Differential Privacy*

Para calcular a sensibilidade dos dados, optamos por determinar a média de todos os valores da coluna, Após isto, calculamos a média dos valores dessa mesma coluna, tirando sempre uma linha. A sensibilidade é dada pela maior diferença entre a média com todas as linhas e as médias com menos uma linha.

No exercício 1 optamos por substituir os dados nulos por -1. No entanto, isso iria afetar a média. Com isto, optamos por, neste ponto, substituir os valores *None* por 0.

Os resultados para as nossas análises foram as seguintes:

- *Análise 1:*

	past_avg_amount_annuity	past_avg_amt_application	past_avg_amt_credit	past_loans_total
Original Sensitivity	0.932297	12.696286	12.646841	0.000222

Com isto, fica evidente que nas colunas *past_avg_amount_annuity* e *past_loans_total*, a sensibilidade é reduzida. Isto significa que não existe nenhum valor que se destaque, o que permite que os dados estejam mais anonimizados. O mesmo já não se verifica nas outras colunas.

- *Análise 2:*

	Age
Original Sensitivity	0.000083

Tal como na análise 1, a coluna da idade apresenta um valor de sensibilidade reduzido. Isto significa que não existe nenhum valor que se destaque, o que ajuda à anonimização.

2.3. Implementação da Differential Privacy

Neste ponto, para cada linha das colunas em questão, adicionamos ruído calculado através de *np.random.laplace*, com pico de distribuição em 0 e escala igual à sensibilidade apresentada na secção anterior a dividir por épsilon. Decidimos utilizar vários valores de épsilon (os mais comuns) e comparar os resultados.

Os valores da sensibilidade após a implementação da *Differential Privacy* são os seguintes:

- *Análise 1:*

	past_avg_amount_annuity	past_avg_amt_application	past_avg_amt_credit	past_loans_total
Original Sensitivity	0.932297	12.696286	12.646841	0.000222
Epsilon = 0.01	0.931556	12.703533	12.650964	0.000222
Epsilon = 0.2	0.932257	12.695973	12.646719	0.000222
Epsilon = ln(2)	0.932294	12.696307	12.646955	0.000222
Epsilon = ln(3)	0.9323	12.6962	12.646747	0.000222

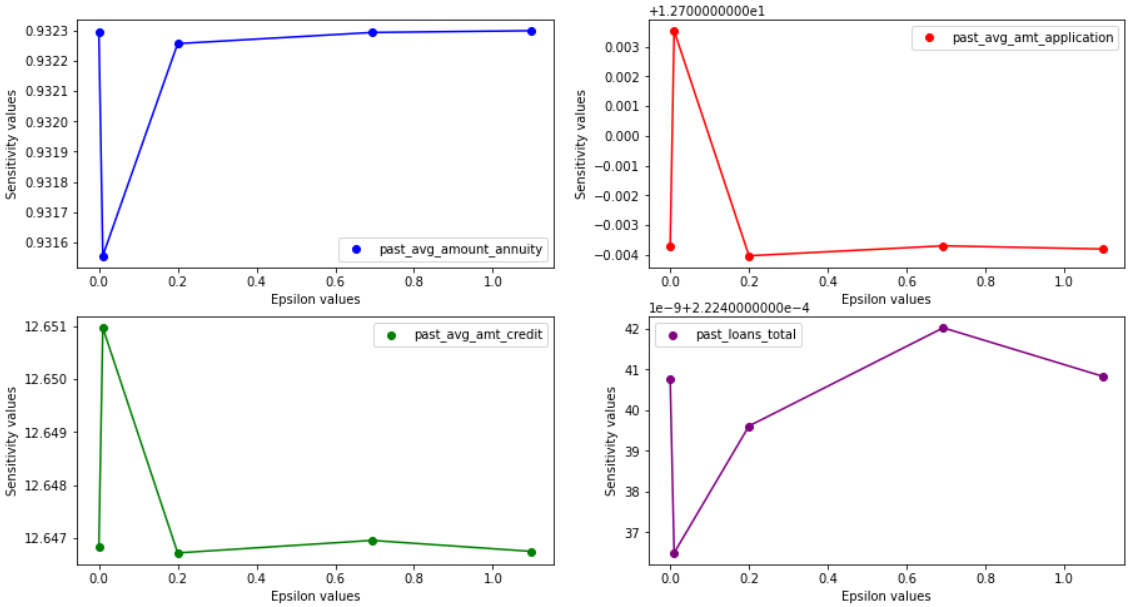


Figura 9 - Análise da sensitivity com os diferentes epsilons em cada past_event

Pelos gráficos acima, vemos que os valores da sensibilidade não se alteraram muito. No entanto, é com o épsilon igual a 0.01 que se nota a maior diferença.

- *Análise 2:*

	Age
Original Sensitivity	0.000083
Epsilon = 0.01	0.000085
Epsilon = 0.2	0.000085
Epsilon = ln(2)	0.000085
Epsilon = ln(3)	0.000085

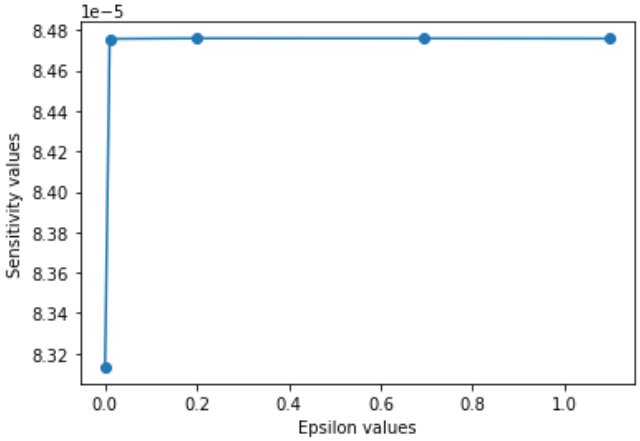


Figura 10 - Análise da sensitivity com os diferentes épsilones

Tal como na análise 1, a sensibilidade não se alterou muito. O valor é igual para todos os épsilones, tendo apenas aumentado muito residualmente quando comparado com o dataset sem ruído.

2.4. Repetição das análises feitas na Meta 1

- ***Análise 1:***

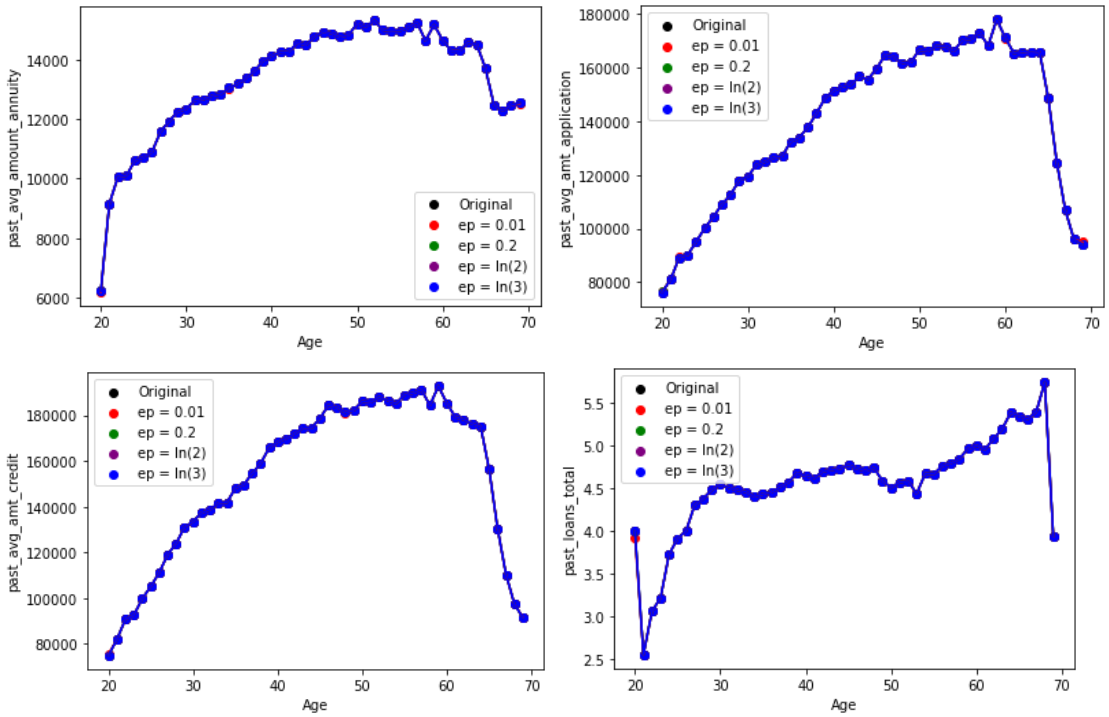


Figura 11 - Análise da média dos past_event por idade dos clientes após a aplicação da Differential Privacy para cada valor de epsilon

Comparando os gráficos acima com os da meta 1, vemos que a tendência das curvas é quase igual. Aliás, nestes gráficos as linhas sobrepõem-se, pelo que concluímos que os dados alterados não diferem muito dos originais. Isto deve-se a estarmos a analisar médias. A adição de ruído nos valores não afeta muito este tipo de métrica, o que comprovamos pelas imagens.

- ***Análise 2:***

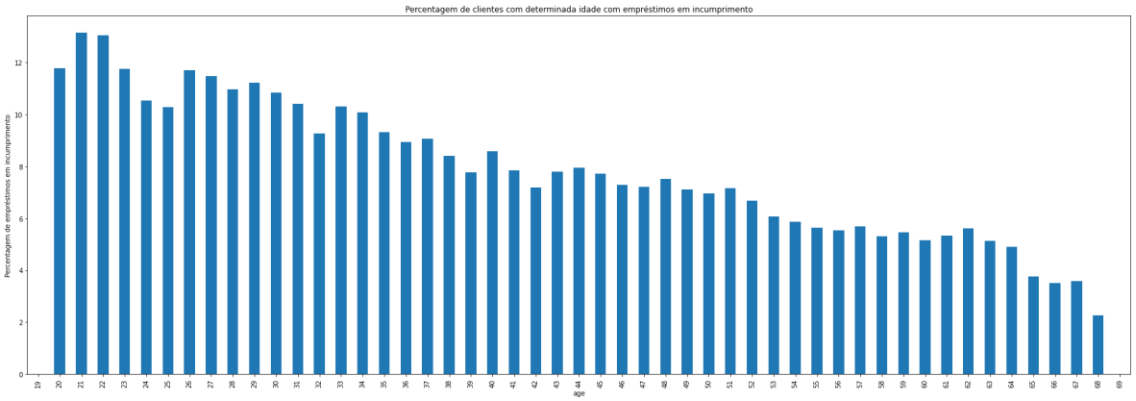


Figura 12 - Análise dos empréstimos infringidos por idade com Differential Privacy de epsilon = 0.01

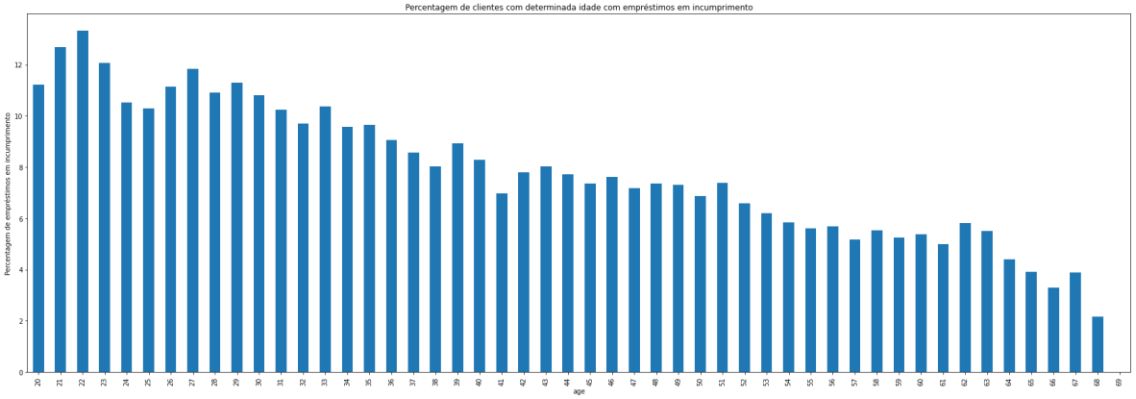


Figura 13 - Análise dos empréstimos infringidos por idade com Differential Privacy de epsilon = 0.2

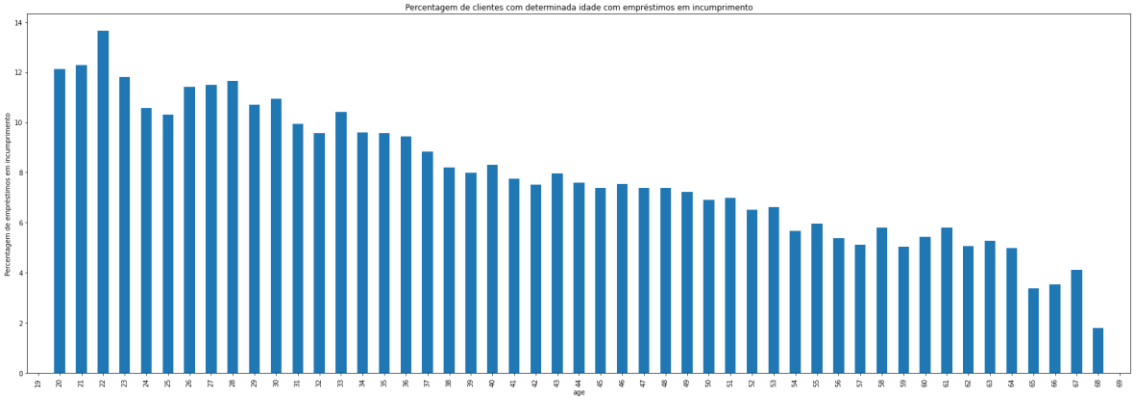


Figura 14 - Análise dos empréstimos infringidos por idade com Differential Privacy de epsilon = ln(2)

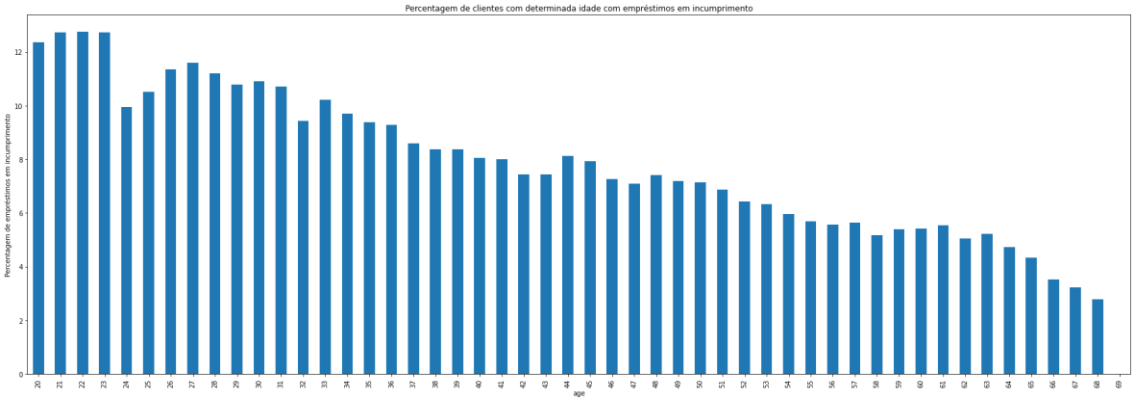


Figura 15 - Análise dos empréstimos infringidos por idade com Differential Privacy de epsilon = ln(3)

Ao contrário da análise 1, nesta já se veem diferenças. Por exemplo, para épsilon igual a 0.01 e ln(2), existem clientes com 19 anos, o que não é verdade. Vemos também que, de épsilon para épsilon, embora a tendência das barras se mantenha aproximadamente a mesma, existem algumas mudanças. Isto deve-se ao facto de o ruído ser adicionado à idade dos clientes. Isto leva, por exemplo, que um cliente classificado com 20 anos passe a ter outra idade, o que irá afetar logicamente a análise, pois passa para outra barra.

2.5. Conclusões

Vantagens

- Dados anonimizados, visto que o ruído os modifica;
- Análise possível, uma vez que, tal como vimos, os resultados são muito semelhantes (sobrepõem-se).

Desvantagens

- Computacionalmente muito exigente, sobretudo no cálculo da sensibilidade;
- Análise não tão precisa, já que os dados estão modificados.

3. Synthetic Data

3.1. Planeamento da implementação

Após pesquisar por diversos sites e analisar o material disponibilizado nas aulas da cadeira, optamos por utilizar o modelo *GaussianCopula* da *package sdv* para gerar os dados sintéticos. Este é um modelo que nos permite obter os resultados pretendidos e não é muito exigente computacionalmente quando comparado com outros (CTGAN).

Com o *GaussianCopula* conseguimos definir que atributos devem ter valores únicos e ainda gerar dados com informações completamente diferentes do *dataset* de treino, o que ajuda na anonimização. É ainda possível avaliar os dados gerados, pelo que consideramos ser uma boa opção para gerar o pedido.

3.3. Passo a passo da geração dos dados sintéticos

Para gerar dados sintéticos, começamos por criar o modelo pretendido (*GaussianCopula*). Após isto, utilizámo-lo para dar *fit* com o nosso *dataset* e gerar dados com o tamanho requerido (307511 linhas, como originalmente).

Para melhorar a privacidade dos dados originais, optamos por, no momento da criação do modelo, especificar que campos não deveriam conter informações iguais às de treino. Chegamos à conclusão que o primeiro e último nome, bem como o ID do empréstimo deveriam ser completamente diferentes dos do *dataset* modelo (ID e alguns QIDs).

Como o ID do empréstimo deve ser único, especificamos isso também na criação do modelo.

3.4. Métricas de avaliação

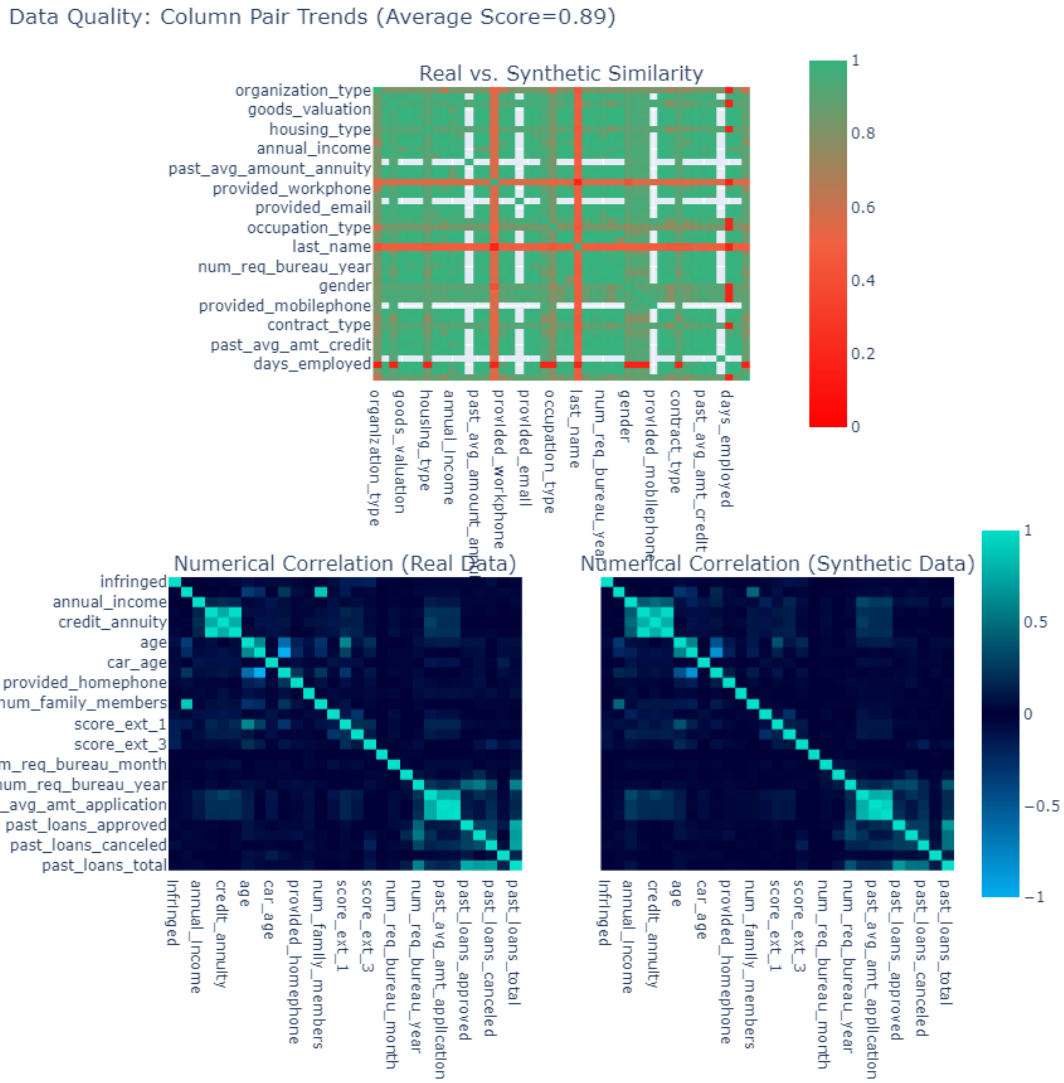


Figura 16 - Qualidade dos dados

Analisando o primeiro gráfico, percebemos que muitos dados sintéticos são semelhantes aos originais, tendo um valor médio de semelhança (qualidade) de 0.89. Como são muitos dados, nem todas as colunas estão representadas. No entanto, isso pode ser visualizado no *notebook*, passando o rato por cima de cada célula.

Já quanto aos gráficos da correlação, vemos que nos dados sintéticos diminui um pouco comparando com os originais, mas os atributos mais correlacionados mantêm-se.

	Column	Metric	Quality Score
0	infringed	KSComplement	0.980619
1	num_children	KSComplement	0.751684
2	annual_income	KSComplement	0.677494
3	credit_amount	KSComplement	0.858203
4	credit_annuity	KSComplement	0.937471
5	goods_valuation	KSComplement	0.864631
6	age	KSComplement	0.977071
7	days_employed	KSComplement	0.261158
8	car_age	KSComplement	0.914369
9	provided_mobilephone	KSComplement	0.999997
10	provided_workphone	KSComplement	0.899691
11	provided_homephone	KSComplement	0.895906
12	mobilephone_reachable	KSComplement	0.998133
13	provided_email	KSComplement	0.989542
14	num_family_members	KSComplement	0.954774
15	region_rating	KSComplement	0.930474
16	score_ext_1	KSComplement	0.857435
17	score_ext_2	KSComplement	0.923881
18	score_ext_3	KSComplement	0.934617
19	num_req_bureau_hour	KSComplement	0.993887
20	num_req_bureau_day	KSComplement	0.994402
21	num_req_bureau_week	KSComplement	0.971044
22	num_req_bureau_month	KSComplement	0.561122
23	num_req_bureau_qrt	KSComplement	0.489148
24	num_req_bureau_year	KSComplement	0.896877
25	past_avg_amount_annuity	KSComplement	0.898138
26	past_avg_amt_application	KSComplement	0.857264
27	past_avg_amt_credit	KSComplement	0.863424
28	past_loans_approved	KSComplement	0.918376
29	past_loans_refused	KSComplement	0.598919
30	past_loans_canceled	KSComplement	0.551822
31	past_loans_unused	KSComplement	0.924753
32	past_loans_total	KSComplement	0.874287
33	contract_type	TVComplement	0.921860
34	gender	TVComplement	0.911987
35	has_own_car	TVComplement	0.958245
36	has_own_realty	TVComplement	0.962183
37	income_type	TVComplement	0.907727
38	education	TVComplement	0.942015
39	family_status	TVComplement	0.955790
40	housing_type	TVComplement	0.920380
41	occupation_type	TVComplement	0.850279
42	organization_type	TVComplement	0.829148
43	first_name	TVComplement	0.571749
44	last_name	TVComplement	0.477765

Figura 17 - Comparação da qualidade de cada coluna

De um modo geral, esta é a qualidade (semelhança) dos dados gerados em cada coluna. Como era de esperar, os dados numéricos, sendo aqueles que menos variam, são os que apresentam melhores métricas. Já os textuais não são tão semelhantes, sobretudo se tiverem vários valores possíveis (como o primeiro e último nome). Por exemplo, a education_type, tendo poucos valores diferentes, apresenta uma métrica de 0.94, o que é bastante positivo.

Decidimos comparar a frequência da idade e dos infringed entre o dataset original e o gerado sinteticamente. Os resultados encontram-se abaixo:

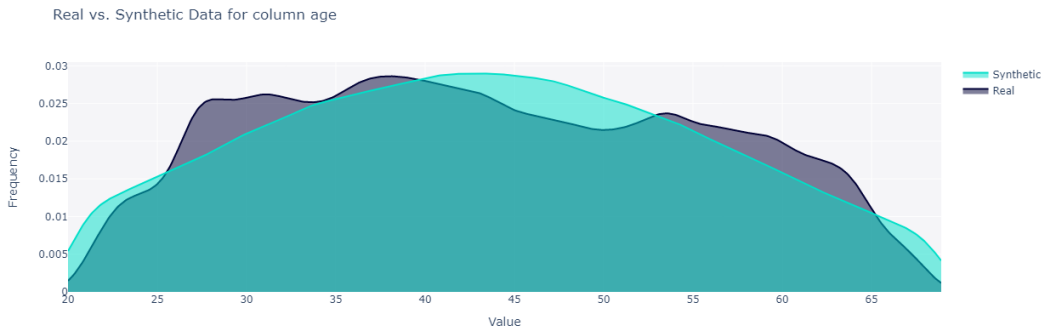


Figura 18 - Comparação da distribuição da idade

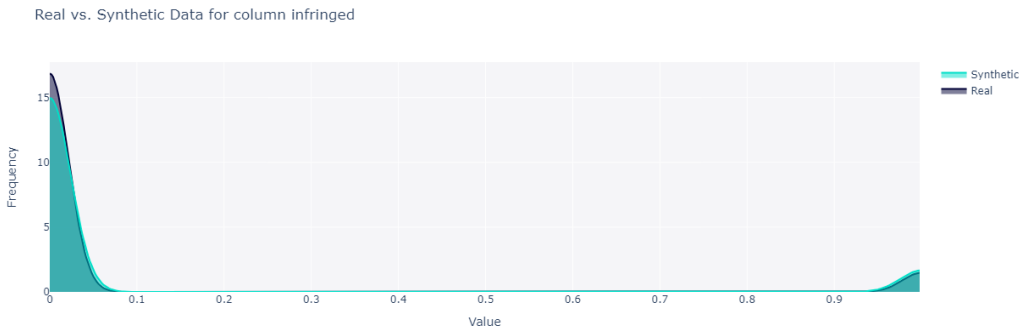


Figura 19 - Comparação da distribuição do atributo infringed

Embora se note que a frequência nos dados sintéticos segue a tendência dos originais, percebemos que a evolução nos sintéticos é muito mais suave, sobretudo na idade. Isto era de esperar visto que o modelo analisa a distribuição do inicial e tenta replicar mediante essa distribuição. Ora, posto isto, podemos deduzir que estamos perante um caso de *under-fitting*, sobretudo nas colunas em que existem mais valores possíveis. Na coluna dos *infringed*, por exemplo, já não se nota uma diferença tão notória uma vez que só há dois valores possíveis.

3.5. Repetição das análises feitas na Meta 1

- ***Análise 1:***

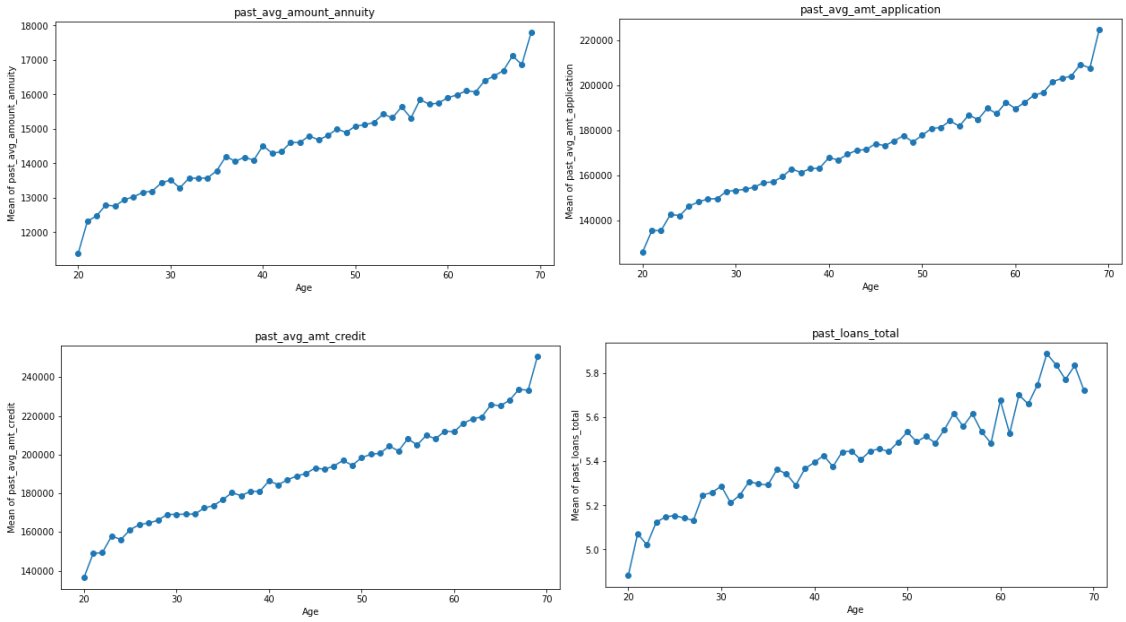


Figura 20 - Análise dos past_events por idade do dataset sintético

Analisando os gráficos acima, percebemos que estamos perante os resultados mais diferentes dos originais que obtivemos ao longo deste relatório. Isso era expectável uma vez que os dados utilizados são todos gerados sinteticamente.

Como essa geração é feita através de um modelo que analisa a distribuição dos dados, os resultantes advêm da tendência detetada pelo modelo. Posto isto, percebemos que, tal como já tínhamos referido acima, estamos na presença de *under-fitting*, uma vez que se percebe que os gráficos acima seguem a tendência dos originais mas que não têm em conta alguns *outliers* que inviabilizariam esta evolução.

Enquanto nos gráficos originais observamos uma diminuição da média a partir dos 60 anos, aqui esses valores continuam a subir, o que poderá significar que apenas seguem a tendência.

- *Análise 2:*

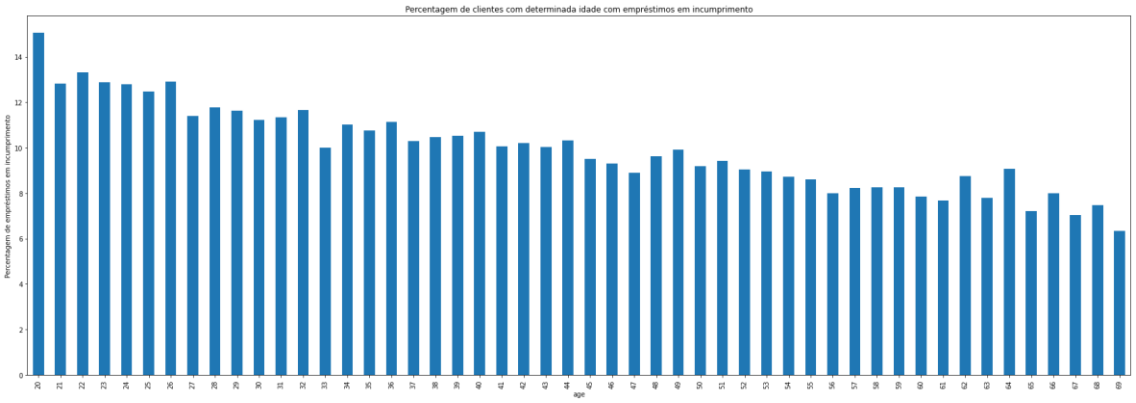


Figura 21 - Análise dos empéstimos infringidos por idade do dataset sintético

Neste caso, a presença de *under-fitting* já não é tão notória. A tendência do gráfico originalmente já é de diminuição das barras à medida que a idade vai avançando. No entanto, essa diminuição é bem mais acentuada, o que indicia que, apesar da tendência ter sido bem detetada pelo modelo, este não conseguiu detetar o decréscimo com a idade mais avançada.

Vantagens

- Anonimização completa dos dados.

Desvantagens

- Perda de informação;
- Análise não tão detalhada, uma vez que os dados sintéticos são gerados através da distribuição dos originais.

Conclusão

Ao longo deste trabalho analisamos o desempenho de diferentes técnicas de anonimização dos dados. Embora algumas anonimizem bem, ou seja, incrementam a privacidade, por vezes fazem com que se perca informação relevante para as análises.

Com isto, percebemos que temos de encontrar um bom equilíbrio entre anonimização e utilidade do *dataset*. Do ponto de vista da privacidade, quanto mais anonimizados os dados estiverem, melhor é. No entanto, quanto mais isso acontecer, menos úteis os dados se tornam.

Deste modo, pensamos que a escolha da técnica a utilizar está dependente do tipo de análise que se pretende fazer. Como vimos, a que encontrou melhor equilíbrio entre anonimização e utilidade dos dados foi a *Differential Privacy*. No entanto, enquanto a análise 1 não sofreu grandes alterações, a análise 2 foi afetada pelo ruído adicionado. Isto indica-nos que para analisar médias a *Differential Privacy* é adequada mas para idades nem tanto.

Em suma, este trabalho ajudou a colocar em prática a matéria lecionada nas aulas teóricas e práticas da cadeira, o que permitiu que consolidássemos e entendêssemos melhor os conceitos abordados em Segurança e Privacidade.

Referências

- Quasi-Identifier Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Reidentification. (s.d.). Publishing Open Access research journals & papers | Hindawi. https://www.hindawi.com/journals/wcmc/2021/7154705/, acedido em 03 de dezembro de 2022
- Configuration | ARX - Data Anonymization Tool. (s.d.). ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing. https://arx.deidentifier.org/anonymization-tool/configuration/, acedido em 03 de dezembro de 2022
- Tabular Models — SDV 0.17.2 documentation. (s.d.). The Synthetic Data Vault. Put synthetic data to work! https://sdv.dev/SDV/developer_guides/sdv/tabular.html, acedido em 15 de dezembro de 2022
- numpy.random.laplace — NumPy v1.24 Manual. (s.d.). NumPy. https://numpy.org/doc/stable/reference/random/generated/numpy.random.laplace.html, acedido em 15 de dezembro de 2022
- GaussianCopula Model — SDV 0.17.2 documentation. (s.d.). The Synthetic Data Vault. Put synthetic data to work!, acedido em 15 de dezembro de 2022https://sdv.dev/SDV/user_guides/single_table/gaussian_copula.html#can-i-evaluate-the-synthetic-data
- Standard Providers — Faker 15.3.4 documentation. (s.d.). Welcome to Fakerâ??s documentation! — Faker 15.3.4 documentation. https://faker.readthedocs.io/en/master/providers.html, acedido em 17 de dezembro de 2022
- Antunes N., (2022). Slides Teóricos, MECD 2022/23, acedido em 18 de dezembro de 2022
- Cardoso N., (2022). Material Prático, MECD 2022/23, acedido em 18 de dezembro de 2022