

Segurança e Privacidade

Mestrado em Engenharia e Ciência de Dados

2022/2023

Assignment 2:

Privacy-Preserving

Data Sharing

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

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

Índice

[Introdução 3](#_Toc122277504)

[1. Anonimização com modelos de privacidade 4](#_Toc122277505)

[1.1. Caracterização do dataset através da classificação de atributos 4](#_Toc122277506)

[1.2. Análise da distinção e separação dos QIDs 4](#_Toc122277507)

[1.3. Medição dos riscos de privacidade do dataset na forma original 5](#_Toc122277508)

[1.4. Hierarquia usada para os Quasi-Identifying 5](#_Toc122277509)

[1.4.1. Gender 5](#_Toc122277510)

[1.4.2. Annual Income 5](#_Toc122277511)

[1.4.3. Income Type 6](#_Toc122277512)

[1.4.4. Family Status 6](#_Toc122277513)

[1.4.5. Age 6](#_Toc122277514)

[1.4.6. Organization Type 6](#_Toc122277515)

[2. Differential Privacy 7](#_Toc122277516)

[3. Synthetic Data 7](#_Toc122277517)

[Conclusão 7](#_Toc122277518)

[Referências 8](#_Toc122277519)

# Introdução

Cada vez mais a questão da privacidade dos dados é debatida em praça pública. Inclusive, foram criadas muitas leis com o intuito de

Nos dias que correm, o mundo é controlado por informação (dados). Como tal, para que não haja nenhuma desgraça na sociedade, é necessária a presença de segurança nas trocas de informação.

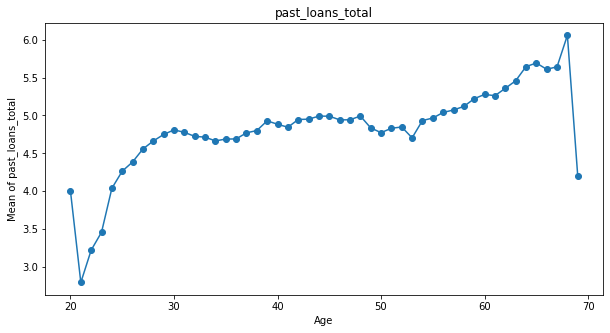
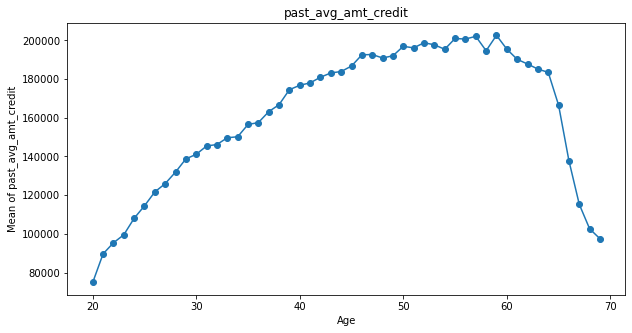
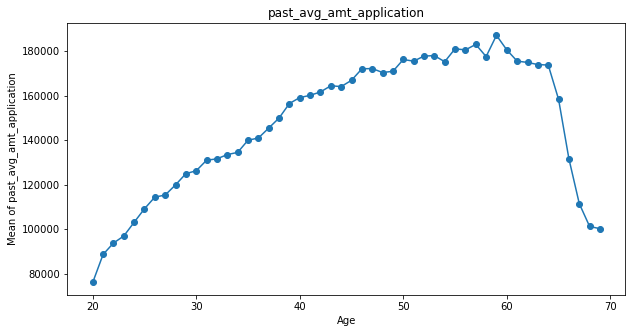
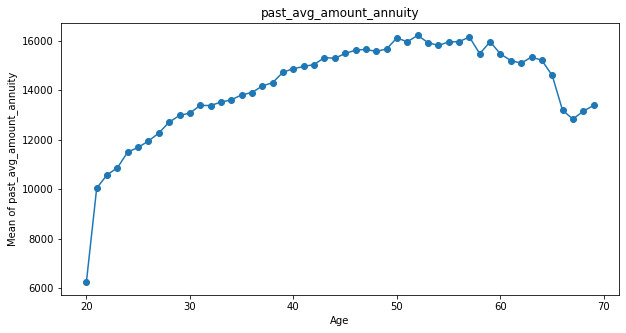
É com isso em vista que este trabalho prático procura simular a troca de informações confidenciais entre duas empresas. Para tal, são utilizados mecanismos de troca de chaves e de encriptação/desencriptação, de modo a garantir confidencialidade, integridade e autenticidade. Sem estes mecanismos, a comunicação entre as duas entidades estaria comprometida e exposta a possíveis atacantes.

Ao longo deste relatório vamos abordar o modelo de ameaça ao sistema e o esquema de comunicação. Vamos ainda analisar o impacte de alguns dados em empréstimos em incumprimento, bem como comparar o tempo de execução do nosso programa com e sem mecanismos de validação da autenticidade e da integridade.

# Análises provenientes da meta 1

Ao longo deste relatório, iremos comparar as análises efetuadas na meta 1 com as efetuadas utilizando o dataset anonimizado.

No entanto, uma das análises que tínhamos realizado 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 das colunas past\_avg\_amount\_annuity, past\_avg\_amt\_application, past\_avg\_amt\_credit e past\_loans\_total, por idade de cliente. Os resultados foram os seguintes:

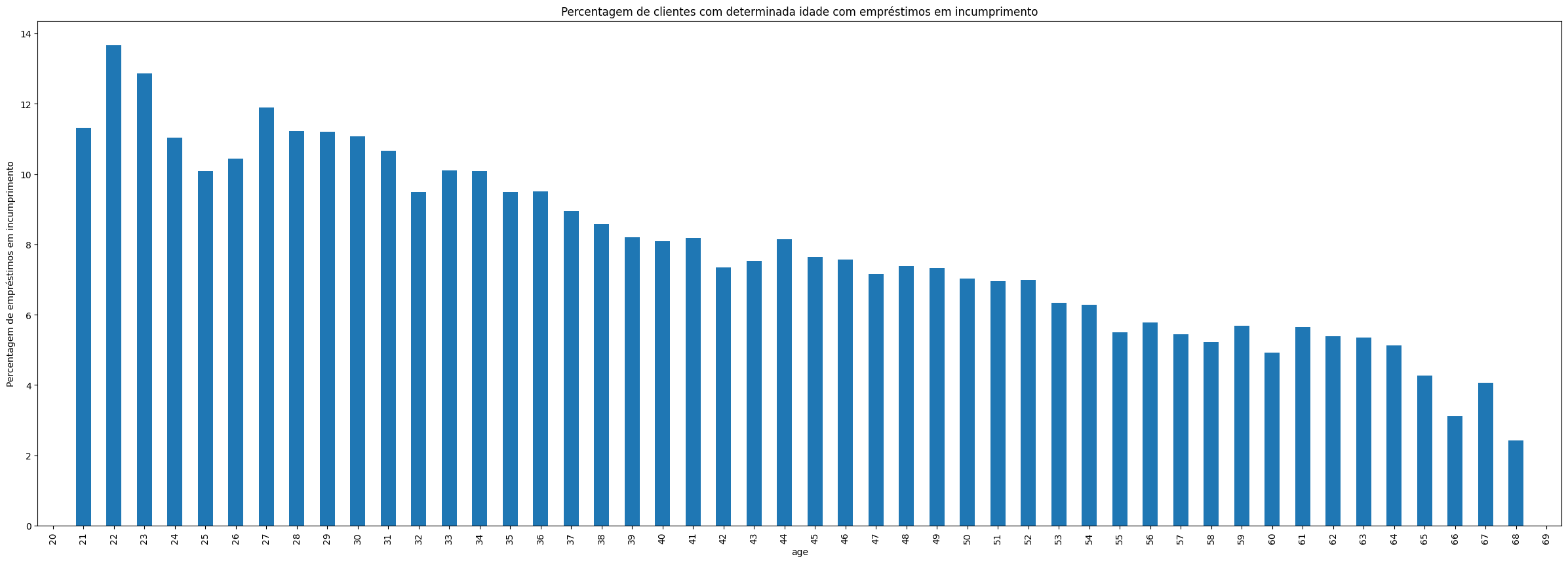


Pelos gráficos acima, percebemos que existe um padrão semelhante em todos. Era de esperar que o montante aplicado e de crédito fosse aumentando com a idade uma vez que, quanto mais velha for uma pessoa, mais provável é de ter pedido empréstimos, logo mais dinheiro pagou/recebeu. Esta tendência vê-se 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 um aumento no valor pago por ano na idade mais ativa. Explicamos isto uma vez que, quanto mais recebe, mais paga pelo empréstimo.

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 começam a pagar menos. Já quanto aos outros gráficos, podemos deduzir que os empréstimos ganharam popularidade já depois destas pessoas terem necessitado para comprar casa, logo, não têm tantos pedidos nem dinheiro gasto/recebido.

Quanto à segunda análise, 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:



Tal como era espectável, quanto mais novos são os clientes, a percentagem de  
empréstimos em incumprimento é superior. 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.

# 2. 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 também o primeiro e último nome do cliente em duas colunas diferentes. Posto isto, criamos uma só coluna “nome” com a junção de ambos e eliminamos as outras duas.

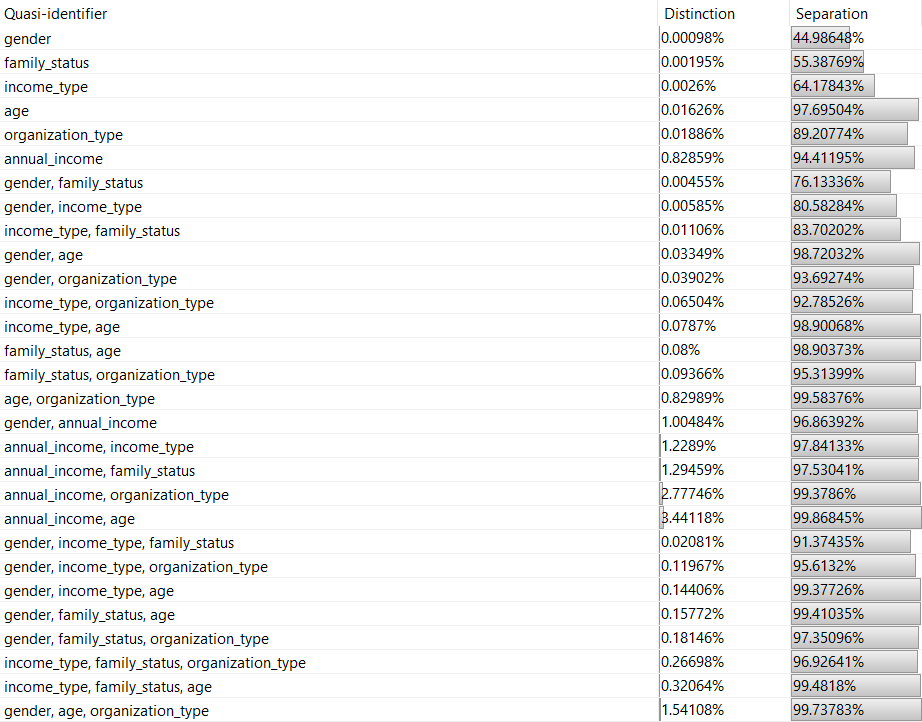
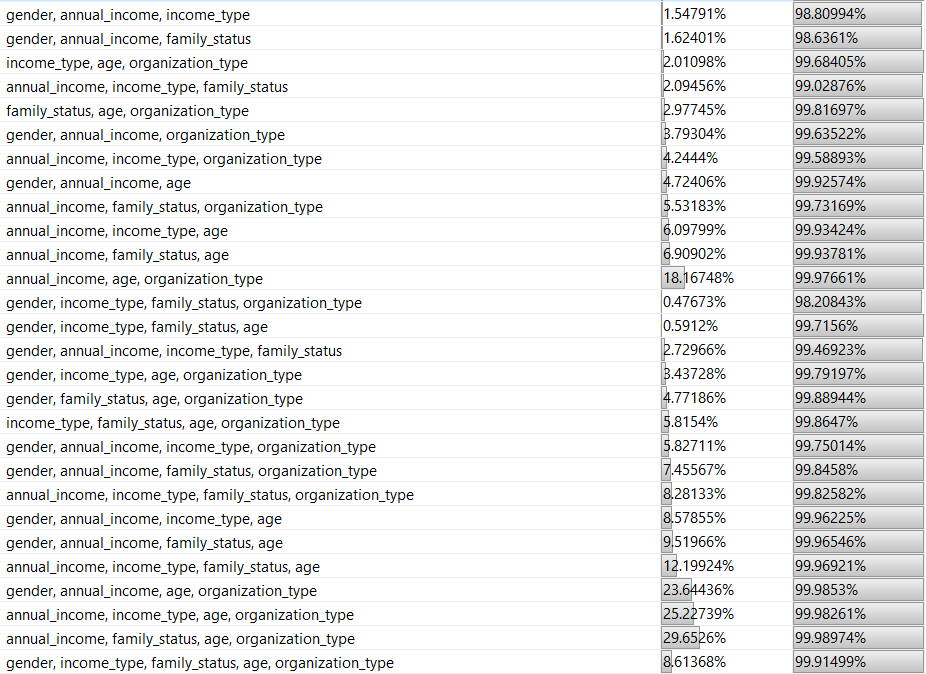
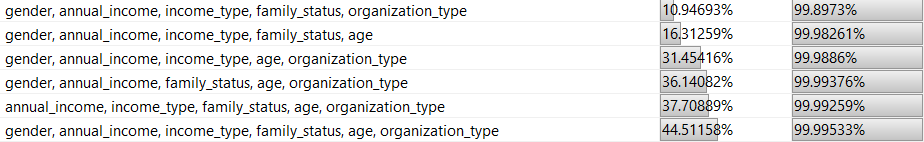
Como algumas colunas apresentavam em certos casos valores nulos, decidimos que a melhor solução seria substituir esses valores por -1. Surgiu a ideia de eliminar as linhas que continham valores None mas, sendo que era só em algumas colunas, estaríamos a perder imensa informação com essa abordagem. Desta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

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

|  |  |
| --- | --- |
| **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 do dataset foram classificadas como Insensitive.

## 1.2. Análise da distinção e separação dos QIDs

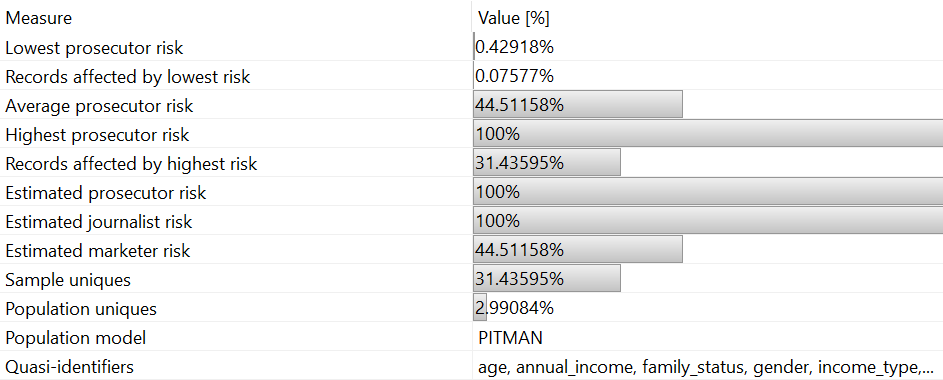
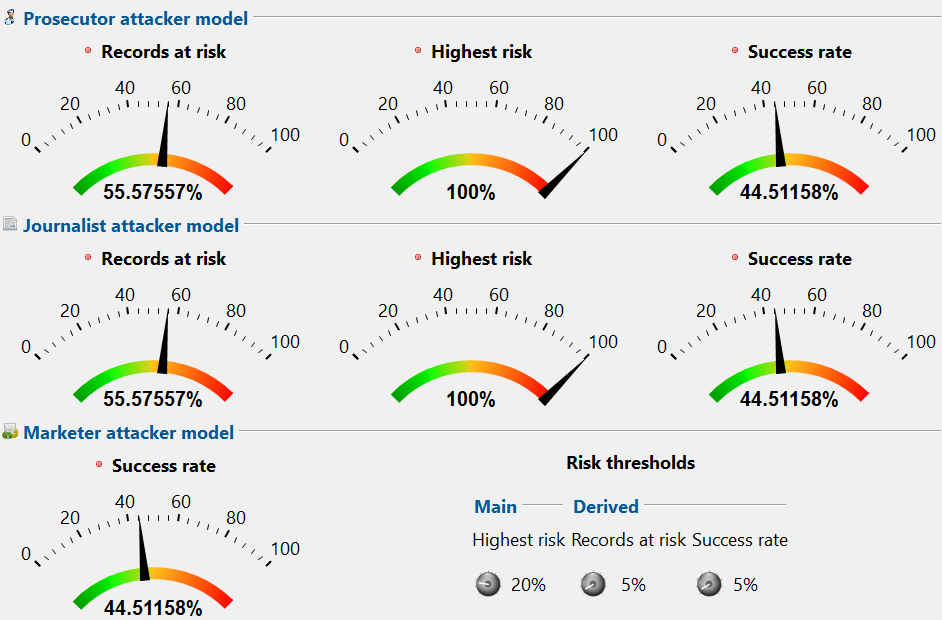
  

Tal como era de esperar, o valor da distinção aumenta proporcionalmente ao número de QIDs juntos. No entanto, conseguimos ver que, com o mesmo número de QIDs, muitas vezes o valor da distinção baixa com o género. Isto deve-se ao facto de este atributo apenas possuir três valores, o que afeta drasticamente 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 muito elevados, muito 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



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 técnicas.

### 1.4.1. Gender

Foi utilizado *ordering* de um nível pois consideramos que o género, sendo um atributo que contém poucos valores distintos, ou deve ser 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 utilizamos 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 tem imensos valores.

## 1.5. Requisitos de privacidade

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

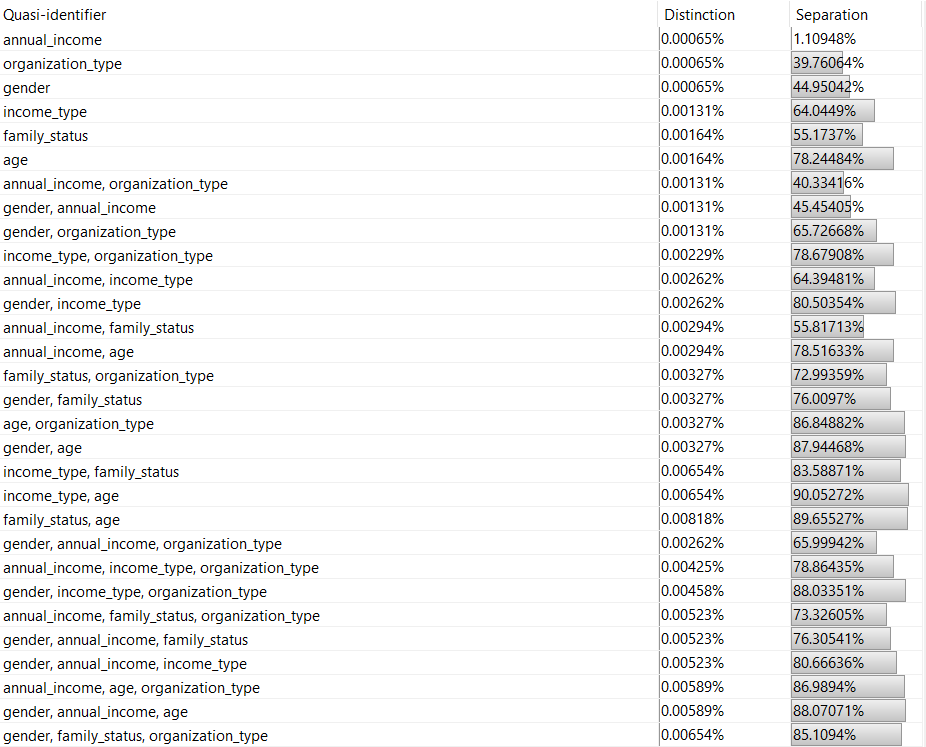
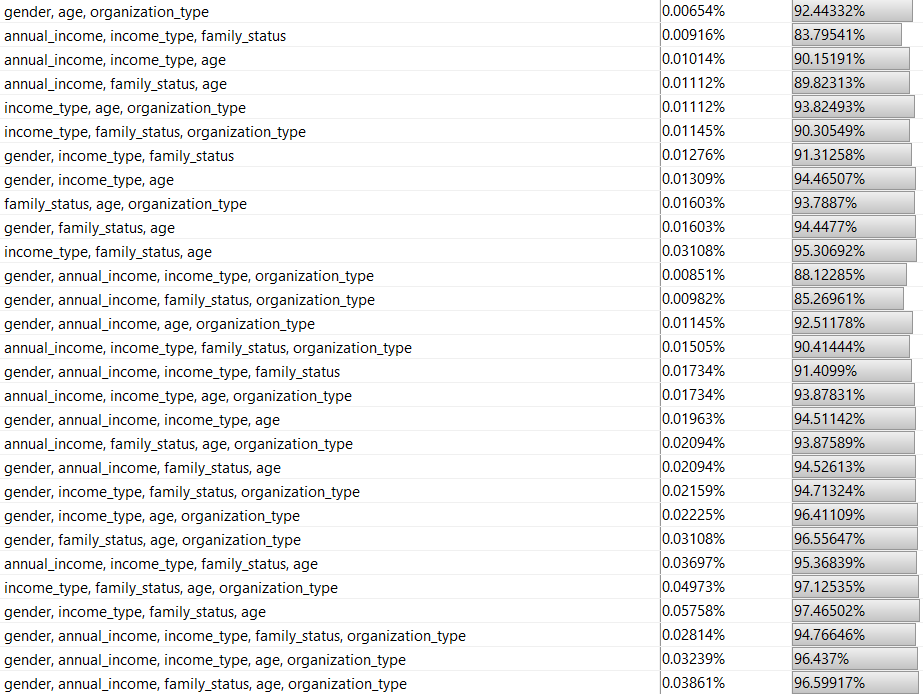
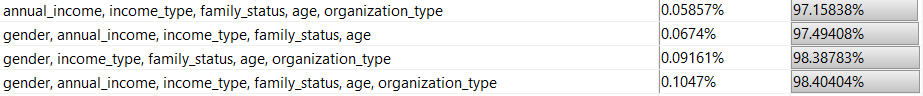
## 1.6. Modelos de privacidade

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. No caso do K, vários valores forem testados até que a medição de riscos demostrasse resultados aceitáveis.

### a. Resultados

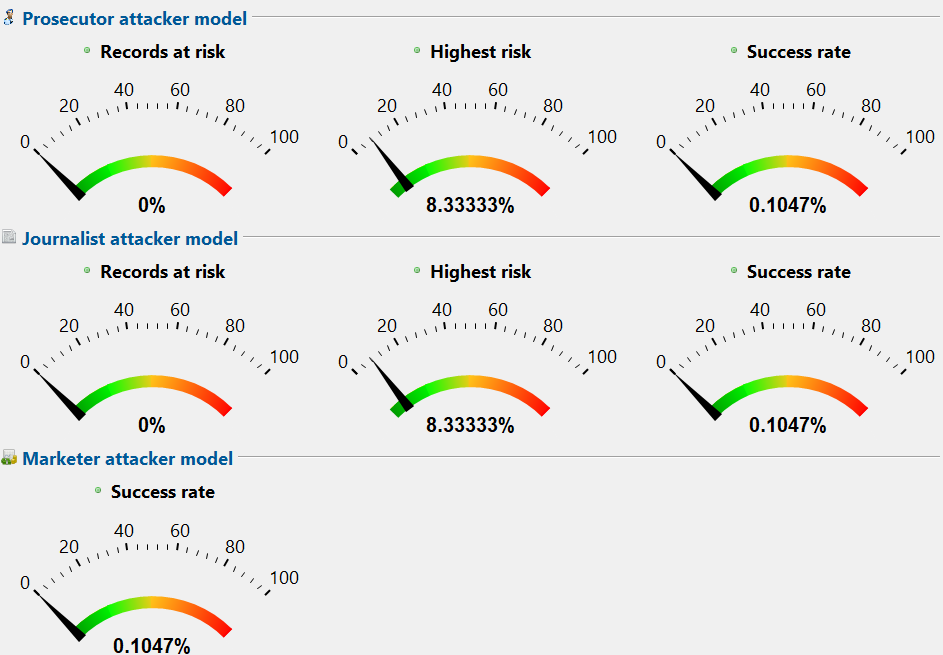
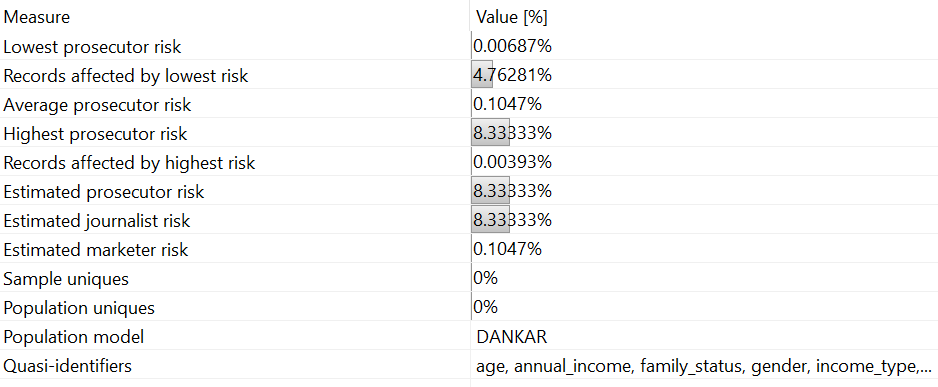
#### 1. Distinção e separação dos QIDs

Analisando os valores da distinção e separação da imagem acima com os apresentados na secção 1.2. deste relatório, percebemos que 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 po risco de existirem QIDs, o que nos garante mais privacidade no dataset.

Os valores da separação reduziram mas não significativamente. Já no que diz respeito aos da distinção, diminuiram bastante, quase até valores nulos.

#### 2. Medição dos riscos de privacidade do dataset anonimizado

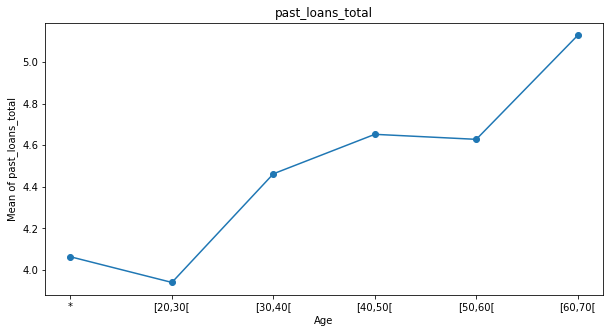
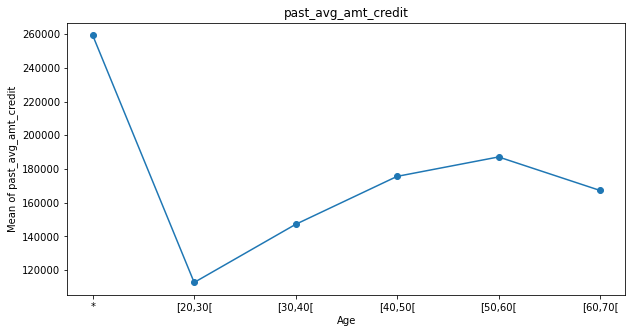
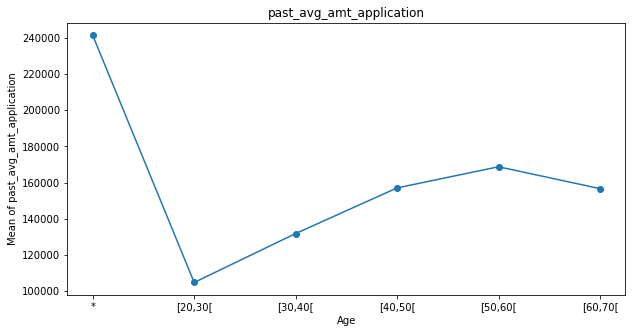
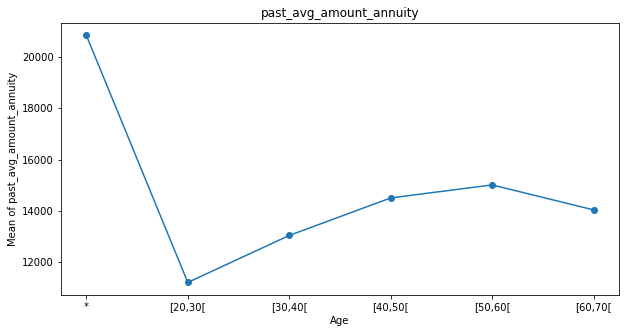
Tal como se esperava, o risco de privacidade diminuiu. Isto comprova que as técnicas utilizadas funcionam e anonimizam de facto os dados. Tínhamos inicialmente mais de metade dos dados em risco e, neste momento, esse risco é nulo. O risco máximo era de 100% e agora é pouco mais de 8%.

Concluímos assim que os modelos de privacidade que implementamos resultaram.

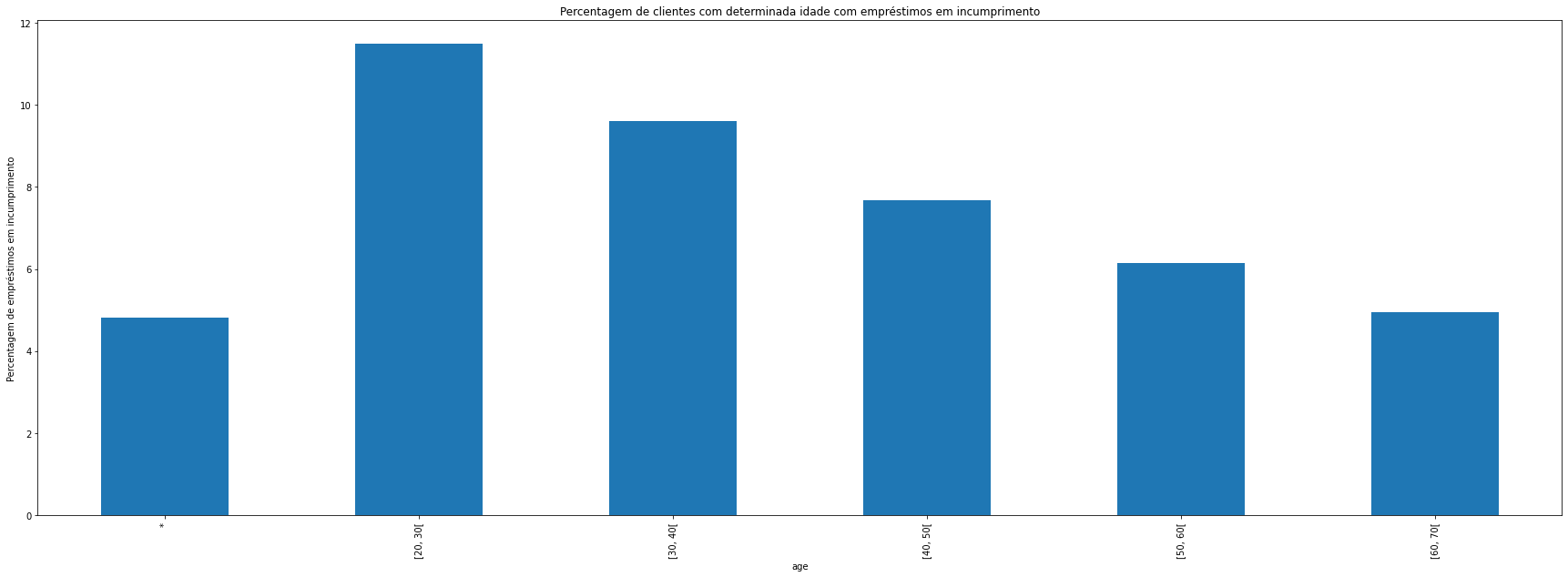
#### 3. Utilidade

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



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



Tal como referimos nas análises anteriores, era previsível que agora 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.

## 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.

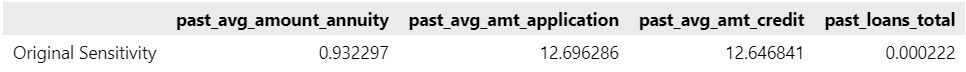
# Differential Privacy

## 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 maior diferença entre a média com todas as linhas e as médias com menos uma linha é considerada a sensibilidade.

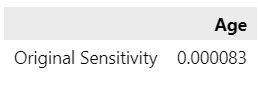
Os resultados para as nossas análises foram as seguintes:

* Análise 1



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



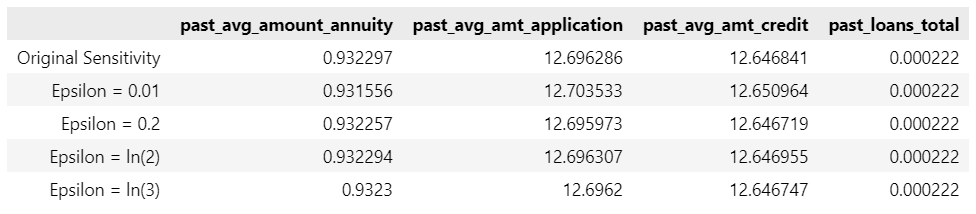
Tal como na análise 1, a coluna da idade apresenta um valor de sensibilidade reduzido. Isto significa que não existe nenhum valor que se destaque, o que ajuda à anonimização.

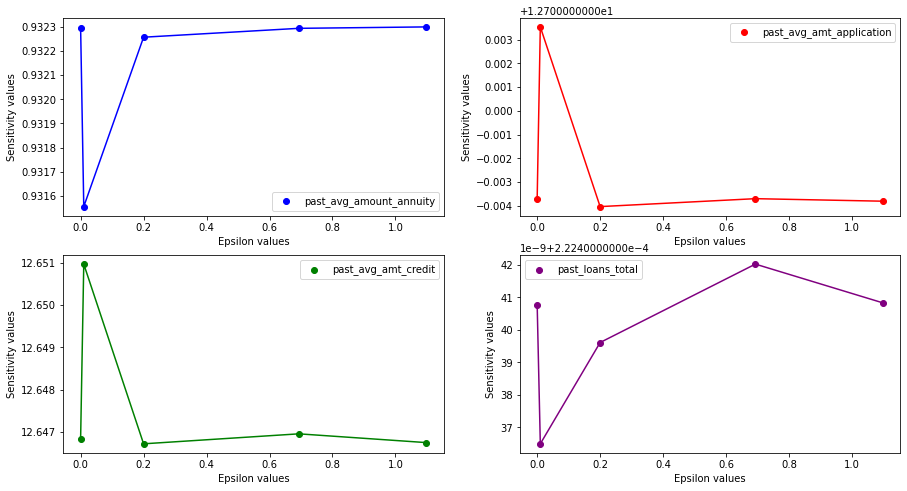
## 2.3. Implementação da Differential Privacy

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 pelo 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:

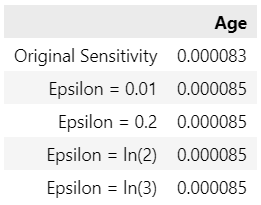
* Análise 1:

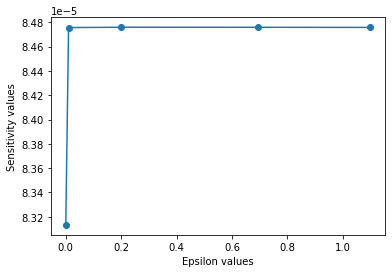




Pelos gráficos acima, vemos que os valores da sensibilidade não se alteraram muito. No entanto, é com o épsilon igual a 0.01 que se nota a maior diferença.

* Análise 2:

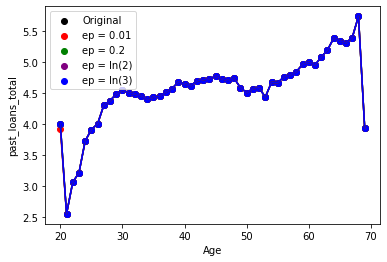
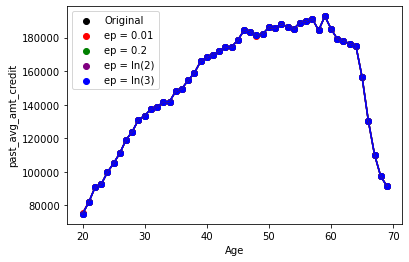
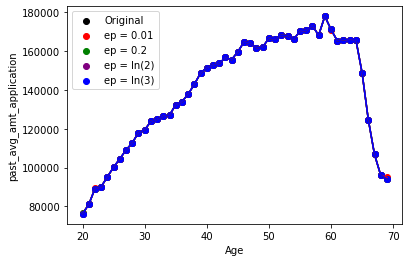
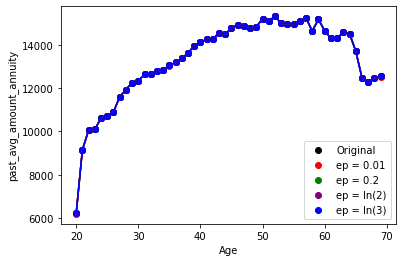




Tal como na análise 1, a sensibilidade não se alterou muito. O valor é igual para todos os épsilones, tendo apenas aumentado muito residualmente quando comparado com o dataset sem ruído.

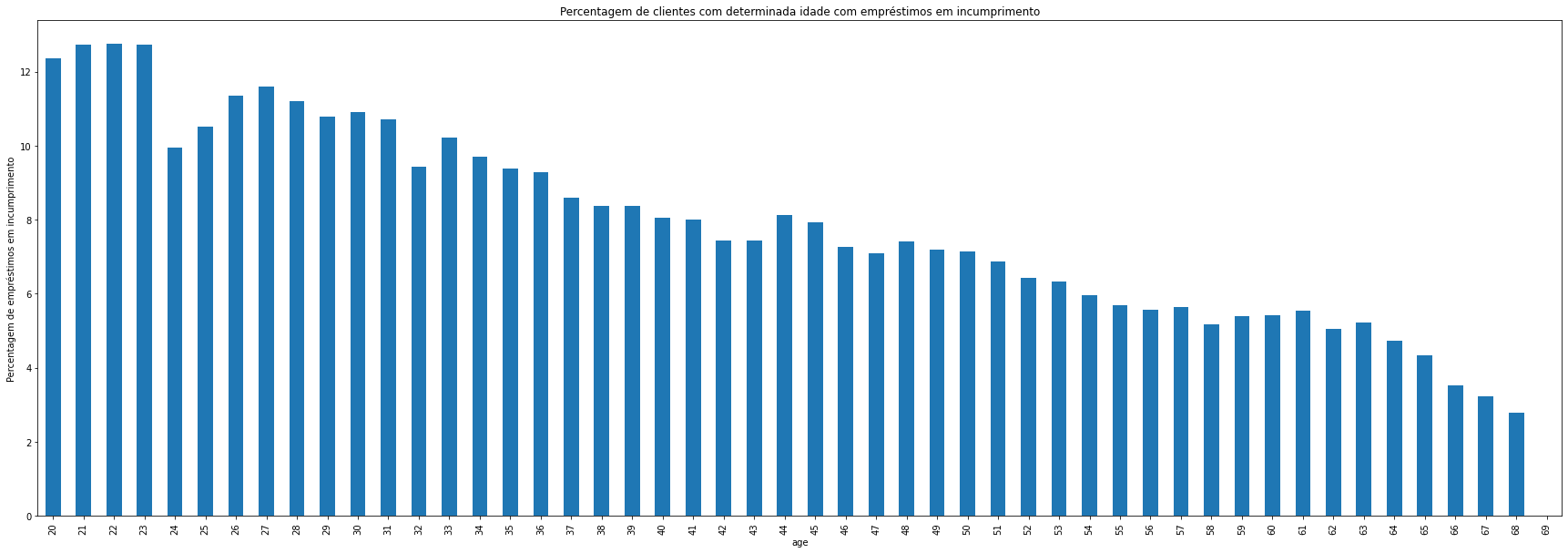
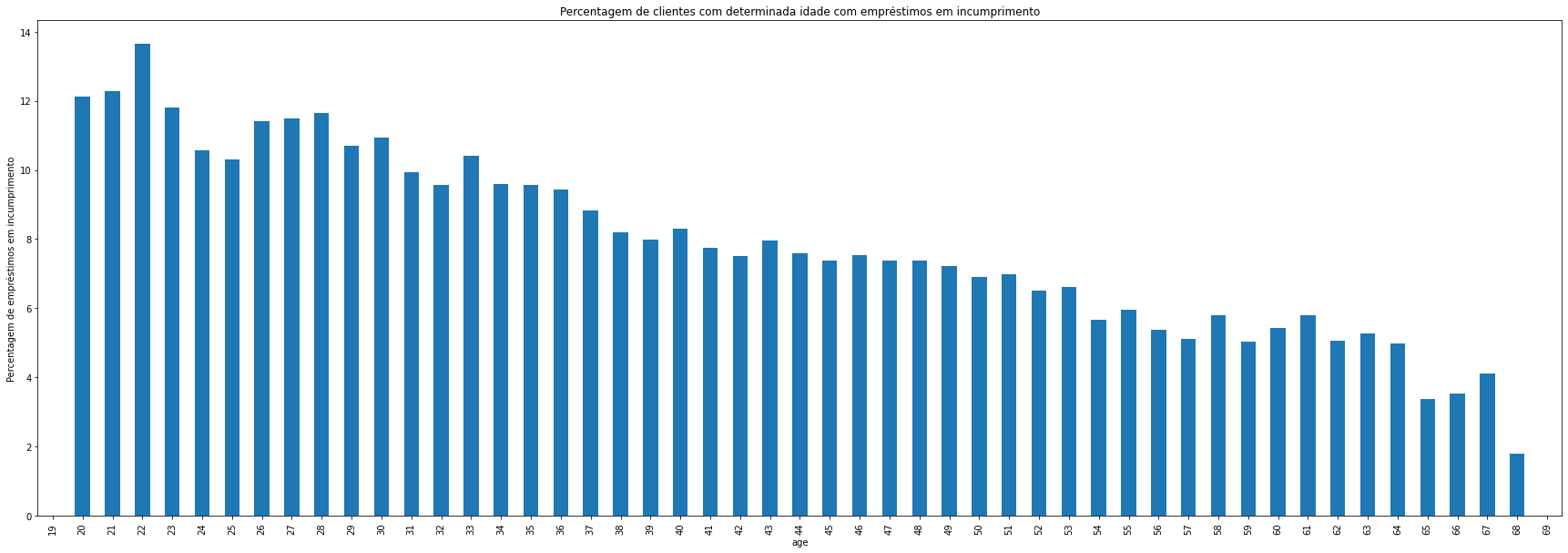
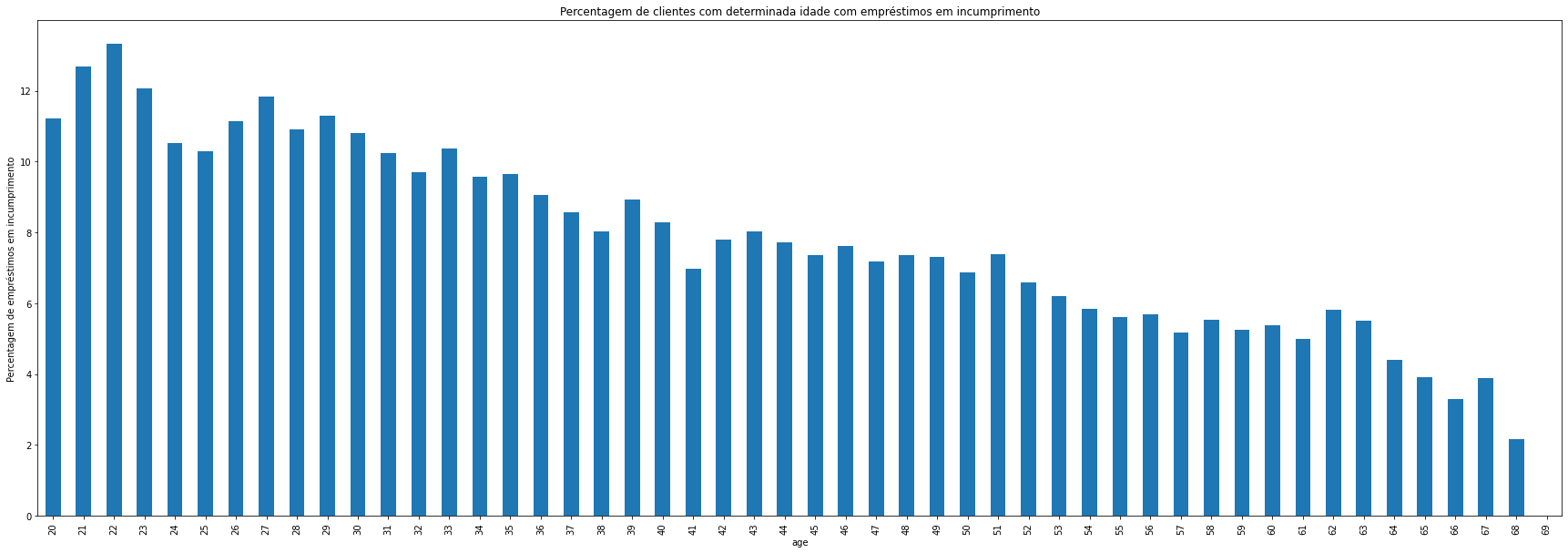
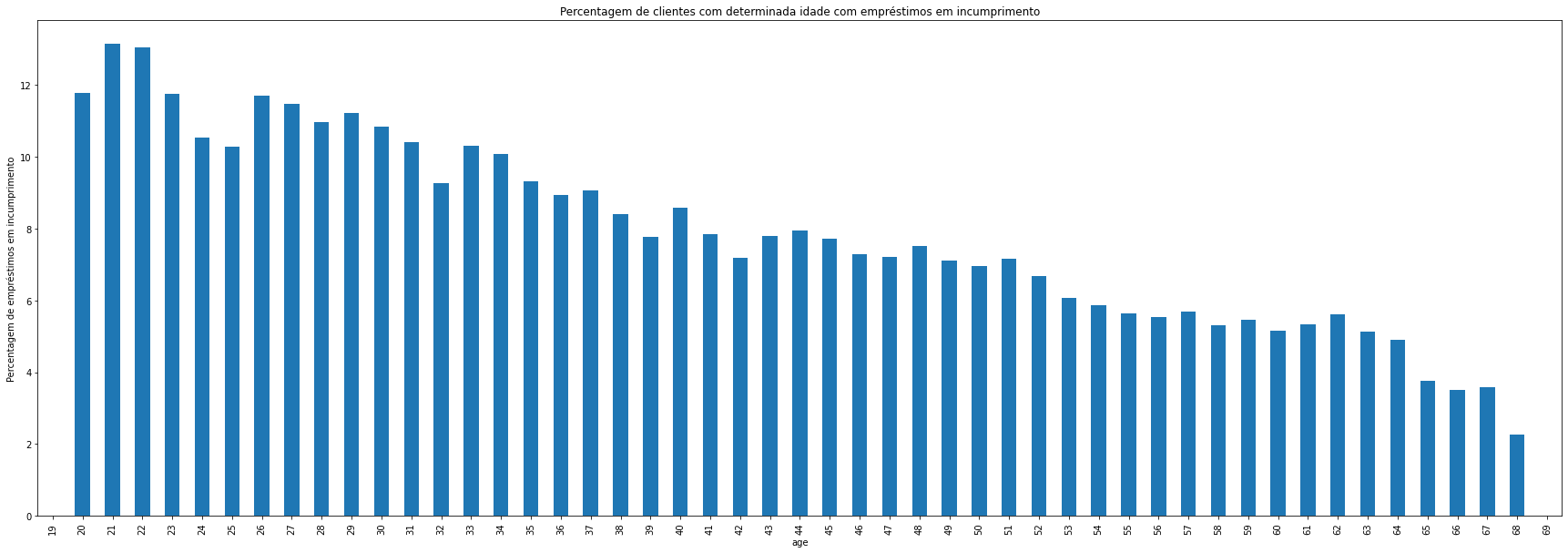
## 2.4. Análises da meta 1

* Análise 1:



Comparando os gráficos acima com os da meta 1, vemos que a tendência da curva da média de cada atributo com a evolução da idade é quase igual. Aliás, nos nosso gráficos, as linhas sobrepõem-se, pelo que concluímos que os dados alterados não diferem muito dos originais. Isto deve-se por estarmos a analisar médias. A adição de ruído nos valores não afeta muito as médias pelo que vemos.

* Análise 2:



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 esta análise, pois passa a contar 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 de como implementar

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 pretendido (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. Logicamente, optamos por analisar que QIDs / IDs não deveriam ser replicados. Chegamos à conclusão que o primeiro e último nome, bem como o ID do empréstimo deveriam ser completamente diferentes dos do dataset modelo.

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

# Conclusão

Ao longo deste trabalho ficou claro que devemos sempre encriptar ficheiros com algoritmos que tenham mecanismos de validação da integridade e autenticidade. Isto leva a que haja uma maior segurança já que, mesmo que um invasor intersete o ficheiro, não consegue alterá-lo uma vez que não tem permissões para tal.

Percebemos também que, analisando o *dataset* disponibilizado, existem colunas referentes aos empréstimos passados que estão correlacionadas com os empréstimos em incumprimento. Olhando também para a idade de quem infringe, reparamos que existe uma maior percentagem de infratores nas gerações mais novas, o que pode significar uma menor estabilidade financeira nestas idades.

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

* Lutes, J. (2020, 2 de dezembro). *Correlation Is Simple With Seaborn And Pandas*. Medium. https://towardsdatascience.com/correlation-is-simple-with-seaborn-and-pandas-28c28e92701e - acedido em 8 de outubro 2022
* *Pandas - Data Correlations*. (s.d.). W3Schools Online Web Tutorials. https://www.w3schools.com/python/pandas/pandas\_correlations.asp - acedido em 8 de outubro 2022
* *Pandas Correlation of Columns*. (s.d.). Spark by {Examples}. https://sparkbyexamples.com/pandas/pandas-correlation-of-columns/ - acedido em 8 de outubro 2022
* Gurav, S. (2022, 10 de agosto). *5 Pandas Group By Tricks You Should Know in Python*. Medium. https://towardsdatascience.com/5-pandas-group-by-tricks-you-should-know-in-python-f53246c92c94 - acedido em 8 de outubro 2022
* Lee, A. (2020, 10 de maio). *Making Plots with the Pandas groupby*. Medium. https://python.plainenglish.io/making-plots-with-the-pandas-groupby-ac492941af28 - acedido em 8 de outubro 2022
* *Plot the Size of each Group in a Groupby object in Pandas - GeeksforGeeks*. (s.d.). GeeksforGeeks. https://www.geeksforgeeks.org/plot-the-size-of-each-group-in-a-groupby-object-in-pandas/ - acedido em 8 de outubro 2022
* *Pandas GroupBy - Count occurrences in column - GeeksforGeeks*. (s.d.). GeeksforGeeks. https://www.geeksforgeeks.org/pandas-groupby-count-occurrences-in-column/ - acedido em 8 de outubro 2022
* *Run a basic correlation between two columns of a dataframe*. (s.d.). Stack Overflow. https://stackoverflow.com/questions/35095249/run-a-basic-correlation-between-two-columns-of-a-dataframe - acedido em 8 de outubro 2022
* *Welcome to pyca/cryptography — Cryptography 39.0.0.dev1 documentation*. (s.d.). Welcome to pyca/cryptography — Cryptography 39.0.0.dev1 documentation. https://cryptography.io/en/latest/ - acedido em 7 de outubro 2022
* Antunes N., (2022). Slides Teóricos, MECD 2022/23 - acedido em 14 de outubro 2022
* Cardoso N., (2022). Material Prático, MECD 2022/23 - acedido em 14 de outubro 2022