Segurança e Privacidade

Assignment 2

Privacy-Preserving Data Sharing

Relatório

André Abreu Couto | 2019215056 Edgar Rafael Cristo Antunes | 2016251340

Introdução

No âmbito da cadeira de Segurança e Privacidade foi realizado o *Assignment 2- Privacy-Preserving Data Sharing* que tinha como principais objetivos:

- Perceber como os modelos de privacidade podem ser utilizados em cenários de processamento de dados;
- Aplicar as técnicas de privacidade para solucionar problemas da vida real;
- Comparar as vantagens e desvantagens das técnicas utilizadas.

Primeiramente, utilizamos a ferramenta ARX para anonimizar os dados antes de os partilhar, utilizando os modelos de privacidade aprendidos, com o objetivo de manter os níveis adequados de utilidade e privacidade dos dados.

Em segundo lugar, utilizamos *differential privacy* como segunda solução, adicionando ruído aos registros, garantindo a sua privacidade.

Finalmente a terceira solução foi gerar dados sintéticos a partir dos do *dataset* original.

Exercício 1: Anonymization with Privacy Models

1.

A classificação dos atributos e a razão da sua escolha segue em anexo num ficheiro Excel denominado "*Att Classification*" . Resumidamente, os atributos que não utilizamos na análise do primeiro Assignement foram removidos para aumentar a eficácia do nosso trabalho, reduzindo o tempo gasto e priorizando aqueles que achamos mais importantes. Os atributos classificados como *QID's* são aqueles que sozinhos não conseguiriam identificar um indivíduo, mas combinados sim. Os atributos classificados como *Sensitive Attributes* são aqueles aos quais os indivíduos não querem ser associados, mas que nós precisamos para fazer a análise. Os atributos classificados como *Identifying Attributes* são aqueles que estão diretamente ligados ao indivíduos e como tal vão ser eliminados.

2.

Quando analisamos os valores de *distinction* e *separation* obtemos resultados como:

Quasi-identifier	Distinction	Separation
infringed	0.00065%	14.84238%
contract_type	0.00065%	17.22953%
gender	0.00098%	44.98648%
age	0.01626%	97.69504%
annual_income	0.82859%	94.41195%
first_name	1.62498%	99.71358%
credit_amount	1.82205%	99.31273%
credit_annuity	4.44635%	99.85528%
last_name	9.16488%	99.91296%
loan_id	100%	100%

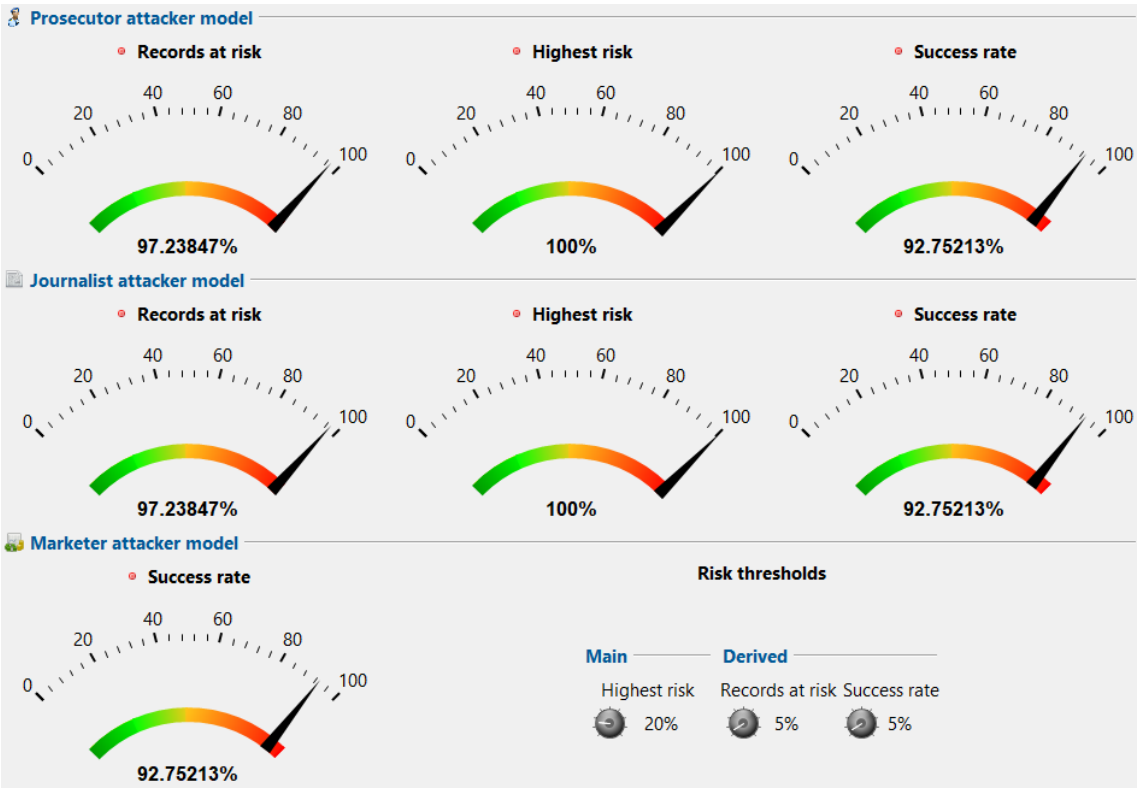
Assim, claramente o *loan_id* é um *identifying attribute* uma vez que atinge valores de *distinction* e *separation* de 100%.

first_name, last_name	86.60048%	99.99975%

O *first_name e last_name* quando combinados também atingem valores muito perto dos 100% e portanto podem ser classificados como *identifying.* Quanto aos outros atributos que classificamos como *QIDs* também atingem valores de *separation* elevados.

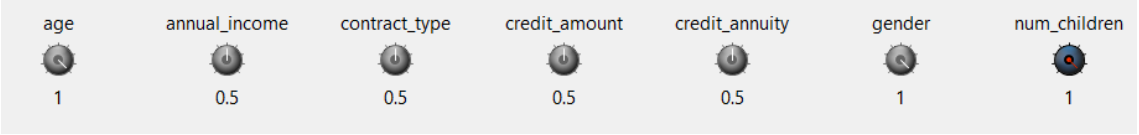
3.

Na análise de risco do *dataset* original, em qualquer modelo do *attacker* (*Prosecutor scenario, Journalist Scenario, Marketer Scenario),* é possível verificar que a maior parte dos dados estão em risco, principalmente no *Prosecutor scenario* (que ataca um indivíduo) e no *Journalist Scenario* (que ataca qualquer indivíduo) sendo que a percentagem de sucesso é também bastante elevada.



4 & 5.

Quanto aos QID's atribuímos um maior peso à idade, ao género e ao número de filhos uma vez que são os atributos que utilizamos na segunda análise.



Aplicamos hierarquias a todos os QID's como é possível verificar no ficheiro ARX que segue também em anexo.

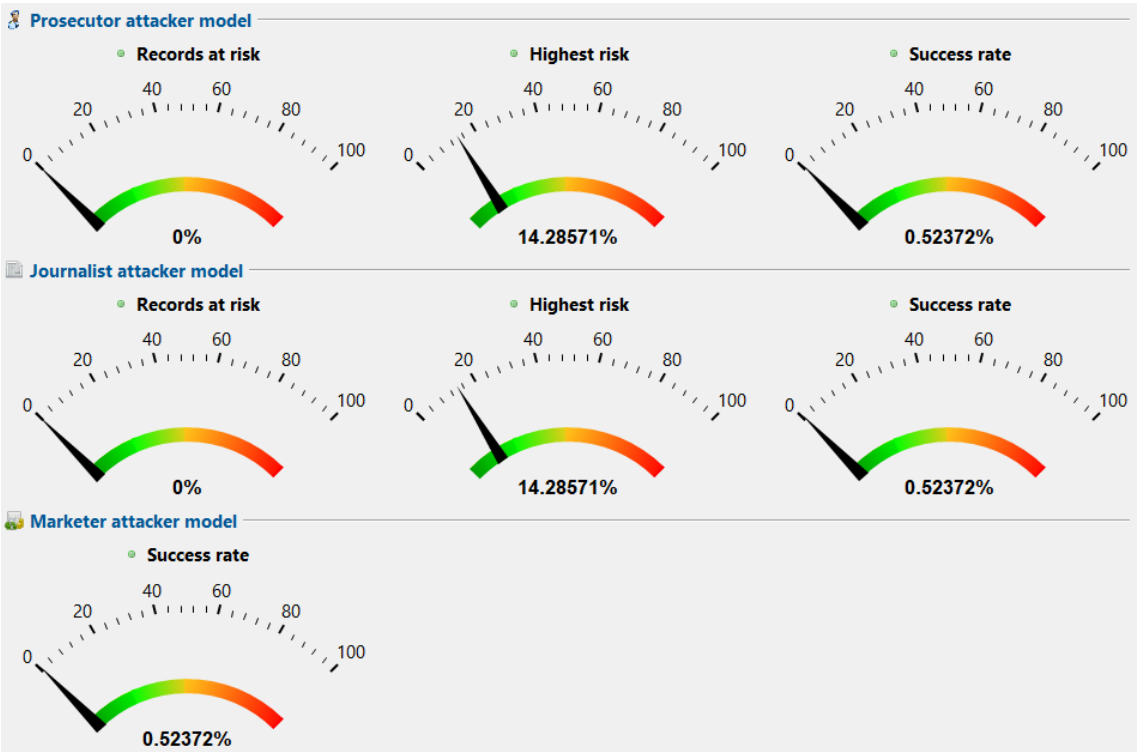
Aplicamos o modelo de privacidade *k-anonymity,* com k=3 uma vez que ao aumentar o k, apesar de aumentarmos a privacidade, diminuímos a utilidade. Para um *dataset* tão grande este k pode ser considerado pequeno, mas a verdade é que após anonimizarmos os dados a classe mínima tinha 7. E como tal decidimos manter o valor uma vez que conseguimos fazer a análise pretendida.

Para os atributos sensíveis, e uma vez que o *k-anonymity* não se preocupa com esses atributos, aplicamos o modelo *l-diversity*, com l=2 para o atributo *infringed*, l=3 para *past_avg_amount_annuity, past_avg_amt_application* e *past_avg_amt_credit,* e l=4 para os *past_loans* restantes.

Consideramos um *suppression limit* de 40% admitindo que podem ser removidos alguns atributos ou todos os dados de um atributo. Definimos ainda que os valores de generalização e supressão que o ARX deve preferir são semelhantes.

6 & 7.

Aplicamos então a anonimização e os resultados foram bastante positivos.



Quanto ao risco, de valores perto de 100% passaram para valores perto de 0%, o que poderia indicar uma destruição de dados. Mas, a nível de privacidade conseguimos alcançar o objetivo. Para perceber se realmente aconteceu essa perda vamos analisar a utilidade.

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
age	String	30.64281%	13.87144%	34.65514%	12.6806%	51.06162%
annual_income	String	30.64281%	11.55953%	6.5997%	0.90443%	69.37602%
contract_type	String	30.64281%	69.35719%	69.35719%	84.40096%	69.35719%
credit_amount	String	30.64281%	9.90817%	8.31556%	0.47845%	63.52301%
credit_annuity	String	30.64281%	26.00895%	69.35719%	50.7656%	15.35292%
gender	String	30.64281%	69.35719%	69.35719%	67.12395%	72.78424%
num_children	String	30.64281%	17.3393%	54.49494%	0.25181%	67.81847%

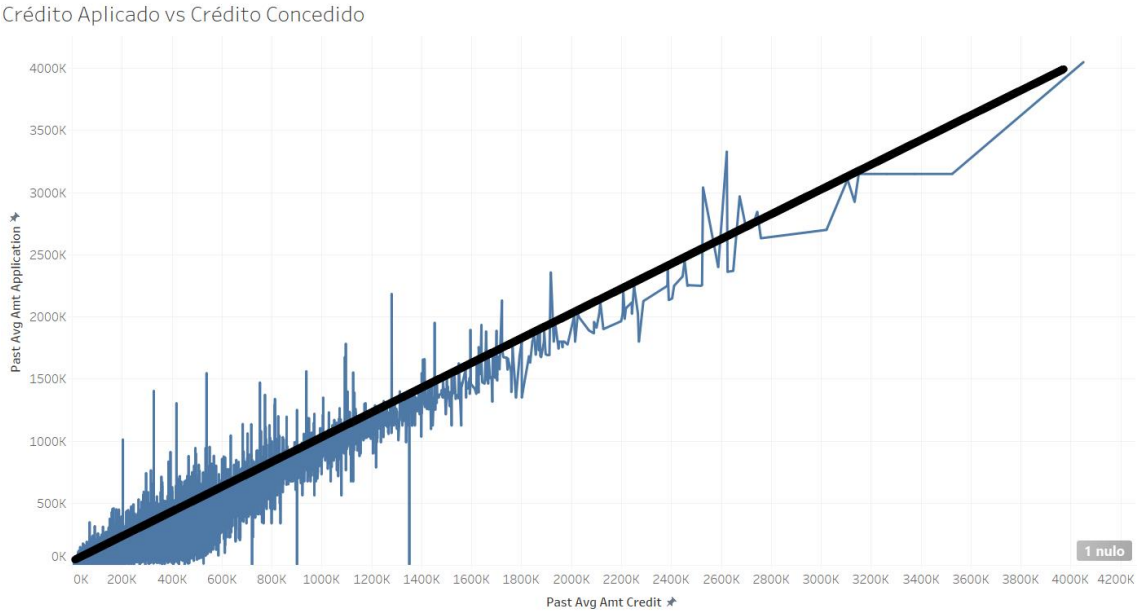
Model	Quality
Gen. intensity	31.0574%
Granularity	41.12344%
N.-U. entropy	22.80188%
Discernibility	69.27313%
Average class size	99.93826%
Record-level squared error	68.47751%
Attribute-level squared error	69.37084%
Aggregation-specific squared error	0.11764%

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	190.94091 (0.06209%)	190.94091 (0.08953%)
Maximal class size	2887 (0.93883%)	2887 (1.35361%)
Minimal class size	7 (0.00228%)	7 (0.00328%)
Suppressed records	94230 (30.64281%)	0
Number of classes	1117	1117
Number of records	307511	213281

Conseguimos perceber que cerca de 30% dos dados foram perdidos, mas como se trata de um projeto em que o objetivo é analisar contratos de crédito consideramos que esta perda de valores está no limite. É também possível verificar que a nível *discernibility* atingiu um valor de cerca de 70%, tal como o *squared error,* o que é ótimo uma vez que quanto mais perto dos 100% melhor. Claro que em métricas que utilizam os tamanhos da classe máxima e mínima não vamos obter valores tão positivos, mas a verdade é que, como vamos poder ver em seguida, fizemos a análise que pretendíamos. E como tal não fizemos mais modificações no modelo que escolhemos utilizar.

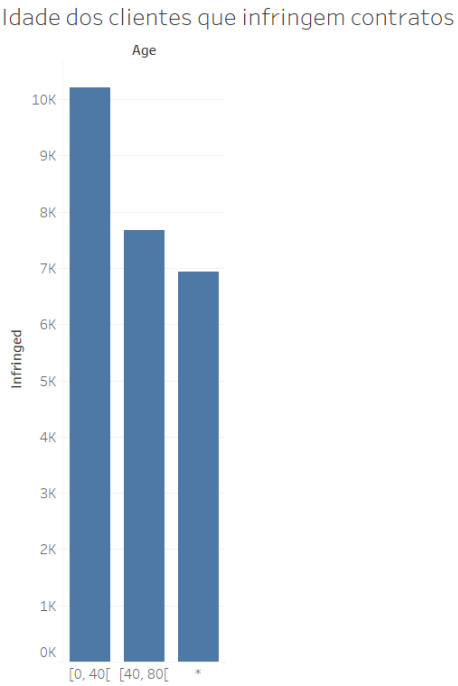
8.

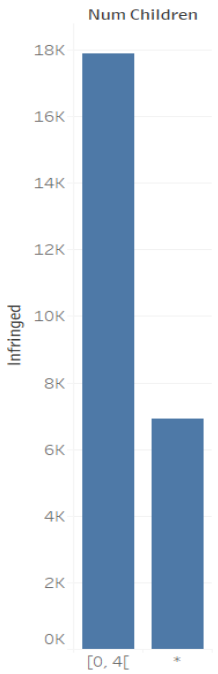
Para refazer as análises do Assignment 1 voltamos a recorrer ao *Tableau.* Quanto à primeira parte percebemos que, apesar dos dados estarem anonimizados, conseguimos fazer a mesma análise e o resultado foi o seguinte.



O gráfico reflete a relação que existe entre a média de crédito concedido a cada cliente e a média de crédito aplicado pelo cliente, nos clientes que infringiram os contratos de empréstimo. A linha preta representa aquele que deveria ser o registo ideal, ou seja, os valores de crédito concedido coincidiam com os valores de crédito aplicado. Mais uma vez podemos verificar que na maior parte dos clientes os valores estão bastante abaixo do pretendido.

Quanto à segunda análise:





Podemos verificar que apenas é possível fazer uma análise mais genérica daquela que tínhamos feito no primeiro Assignment uma vez que antes percebíamos melhor a faixa etária dos indivíduos que infringiam de alguma forma o contrato, e agora apenas temos um intervalo de quarenta anos onde eles incidem. A mesma situação acontece com o número de filhos uma vez que apenas conseguimos perceber que os indivíduos têm entre 0 e 4 filhos.

Assim, os modelos implementados foram eficazes uma vez que nos permitiram garantir a privacidade dos indivíduos. Para além disso, fizemos a primeira análise da mesma forma, só que com os dados anonimizados, e o resultado foi semelhante. A segunda análise mostra-nos as desvantagens da anonimização dos dados visto que estes estão mais generalizados perdendo um pouco a sua utilidade.

Exercício 2: Differencial Privacy

A segunda estratégia proposta pelos engenheiros da *controlER* para resolver o problema foi utilizar *diferencial privacy*.

O *diferencial privacy* tem como principal objetivo divulgar o *dataset* com o mínimo de informações possíveis acerca do indivíduo para que estes se sintam seguros em revelar os seus dados. Como? Perturbando os dados, ou seja, adicionando ruído a partir de uma distribuição de Laplace.

No ficheiro *differencialPrivacy.ipynb* está a implementação do modelo, bem como das funções *count*, *sum* e *mean* que utilizamos para comparar os dados originais com os dados perturbados.

Apesar de não ser permitido utilizar as bibliotecas de terceiros, nós utilizamos apenas para comparar os valores que obtivemos através das funções que criamos.

Escolhemos *Epsilon=0,01* visto que quanto mais pequeno é o seu valor, maior é o ruído adicionado.

Utilizamos também *data = data.fillna(0)* para que os espaços vazios fossem completados com zeros uma vez que estávamos a ter problemas com os mesmos.

Aplicamos o modelo a Aplicamos as funções nos atributos que utilizamos nas análises anteriores, ou seja: *past_avg_amt_credit*, *past_avg_amt_application, age* e *num_children.*

Serve como exemplo os resultados obtidos para o *past_avg_amt_credit:*

```
COUNT:

4685 4607.799728107734 1.6478179699523126
PYDP COUNT

4685 4763 -1.6648879402347918

Sum:

731913621.7887328 910031977.8655331 -24.335980472872563

PYDP SUM

731913621.7887328 594372826.9490967 18.791943577098404


Mean:


156224.89259097818 187419.26720361089 -19.96760829550037

PYDP Mean

156224.89259097818 136893.79685095575 12.373889601981892
```

O primeiro valor que aparece para cada função é com o atributo original, o valor seguinte é quando o ruído é adicionado e o terceiro valor é a percentagem de erro. As funções com *PYDP* são utilizadas com a biblioteca apenas para comparação.

Assim, é possível perceber que nas funções *sum* e *mean* é onde se verificam as maiores discrepâncias nos valores, mas o mesmo acontece quando se utiliza a biblioteca. Para além disso os nossos valores não diferem muito daqueles conseguidos utilizando a biblioteca. E como tal acreditamos que o modelo está a ser implementado com sucesso.

Em todas as nossas funções, a percentagem média de erro é de apenas 8.5%, o que consideramos ótimo uma vez que quanto mais próximos do 0 melhor.

Assim, consideramos este método eficiente na medida em que os dados estão perturbados e, portanto, os indivíduos estão protegidos. Quanto à sua utilidade, uma vez que a percentagem de erro em média é de 8,5%, é possível fazer uma análise eficaz.

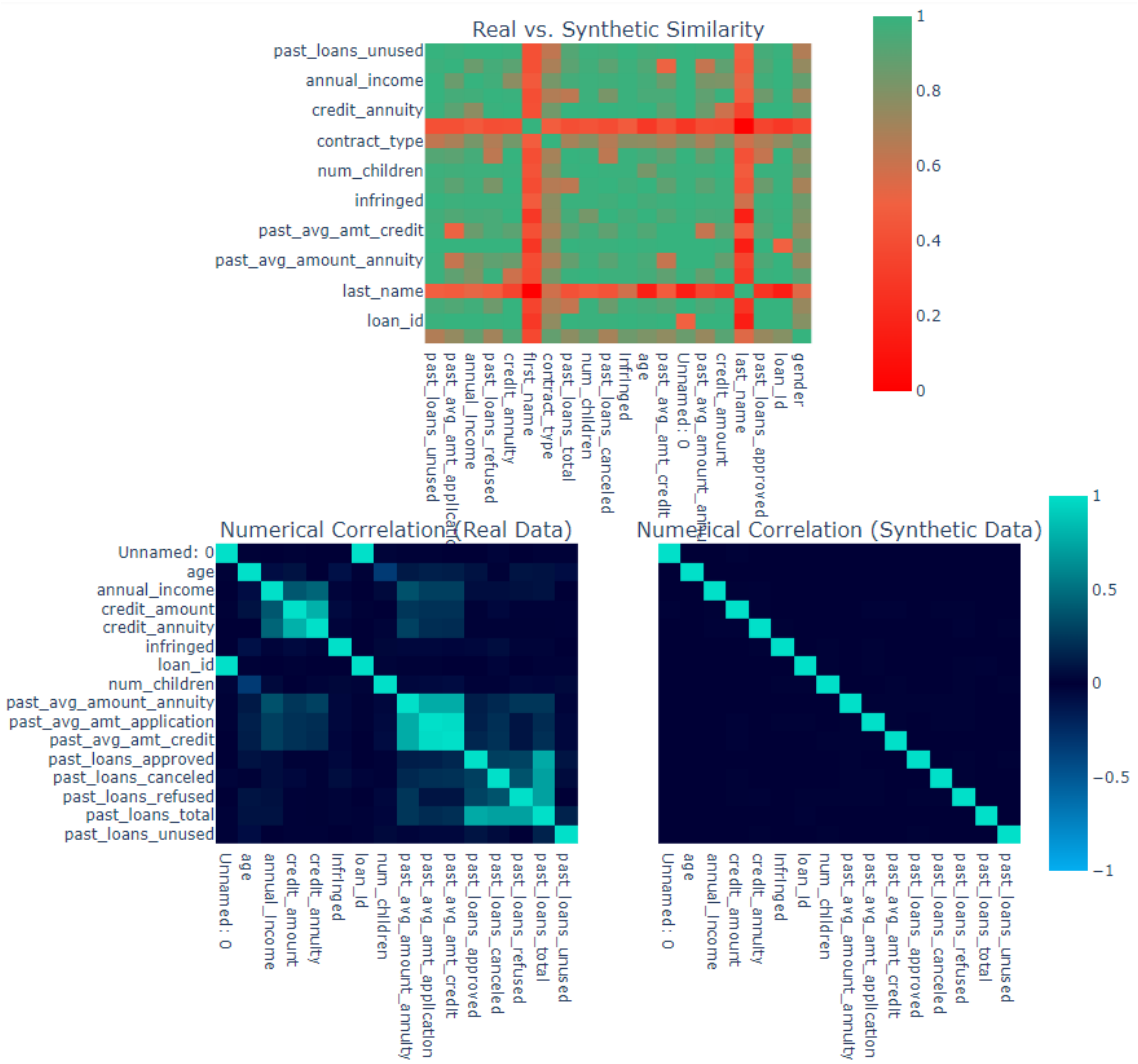
Exercício 3: Synthetic Data

A última solução proposta foi gerar dados sintéticos a partir do *dataset* original, utilizando *SDV.*

Como estávamos a ter problemas em relação ao tamanho do *dataset* resolvemos utilizar 125000 linhas, o máximo que conseguimos.

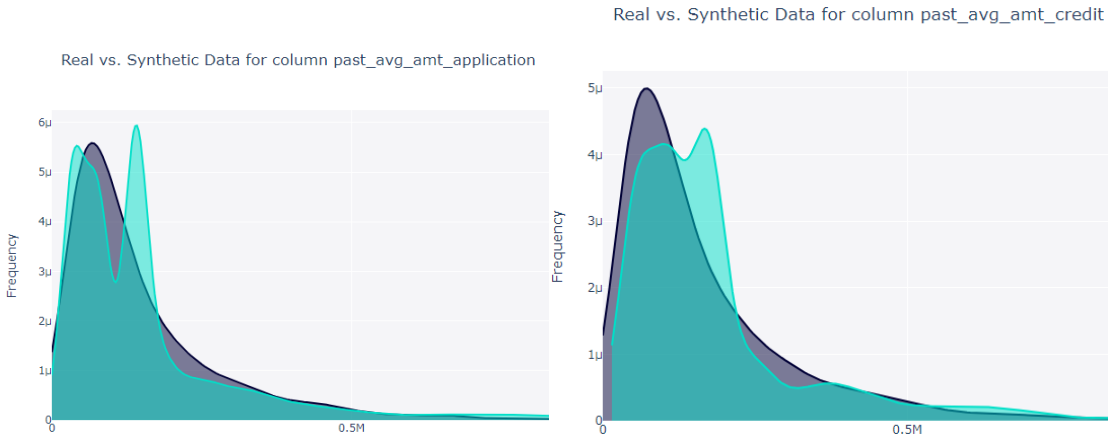
A solução e respetivas avaliações estão implementadas no ficheiro *syntheticData.ipynb.*

Atingimos um *overall_score* de cerca de 88%, sendo que o de qualidade foi de 81,39%. Portanto, no geral os resultados dos dados gerados foram bastante positivos. Principalmente nas colunas *credit_annuity, gender* nas quais os scores foram superiores a 93%. O *first_name* e *last_name* foram as colunas com piores resultados, contudo, acima de 50%.

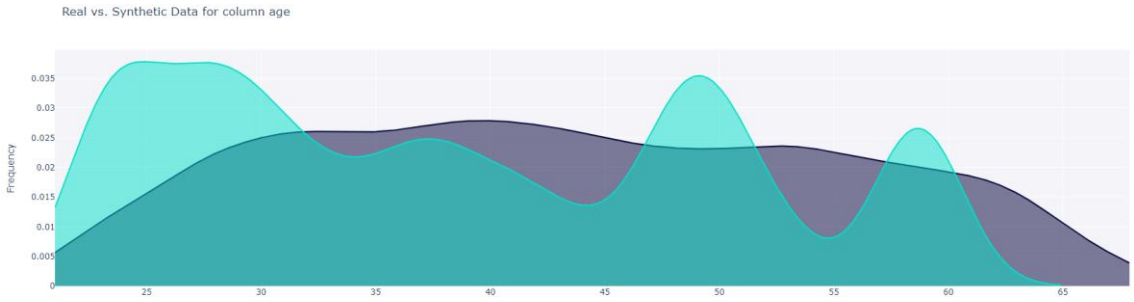


Quanto ao coeficiente de correlação de Pearson podemos verificar que os valores na diagonal são todos 1 o que indica forte correlação dos dados dados. No primeiro gráfico podemos ver melhor a avaliação dos dados e concluir aquilo que dissemos anteriormente sobre as colunas com melhores e piores resultados.

Em relação às categorias que utilizámos nas análises anteriores podemos verificar na maior parte delas os gráficos para analisar a *KS statistic* os dados sintéticos estão sobrepostos com os reais:



Apenas verificamos uma maior discrepância na idade:



Este método é então eficaz na medida em que os dados dos indivíduos estão protegidos. Os dados são sintéticos, ou seja as informações que nós temos não são na verdade dos indivíduos em questão. No entanto, e para o propósito do nosso projeto, os dados são bons o suficiente para fazer a análise pretendida.

Conclusão

Em forma de conclusão resta apenas referir que o projeto relativo a este relatório permitiu-nos perceber a importância da privacidade no tratamento de dados. Os diferentes modelos que implementámos mostraram-nos que não é fácil trabalhar com problemas reais. Existe uma linha muito ténue entre anonimizar os dados e destruí-los, por isso garantir a sua utilidade foi um dos principais obstáculos que conseguimos ultrapassar. Escolher o processo mais adequado para cada situação exige perceber muito bem as vantagens e desvantagens que cada um destes modelos pode trazer. Consideramos, portanto, que todos os objetivos foram cumpridos com sucesso.

Bibliografia

- Apontamentos teóricos e práticos da cadeira de Segurança e Privacidade.