Classificação de Clientes no Mercado Cambial utilizando K-means, DBSCAN e Clusterização.

CLASSIFICATION OF FOREIGN EXCHANGE CUSTOMERS USING K-MEANS, DBSCAN AND CLUSTERS.

Carvalho Pinto, João V. de., Piva, Murilo da Cunha.
jv-carvalhoo@hotmail.com, murilocpiva@gmail.com

Centro Universitário Facens - Sorocaba, SP, Brasil

Submetido em: 15 jan. de 2022. Aceito em: --

RESUMO

Entender o perfil de investimento de clientes é uma forma indispensável para que a gerência da empresa possa elaborar estratégias de maneira mais assertiva de acordo com cada tipo de público. Para auxiliar na tarefa de classificação de perfis, são utilizados algoritmos de classificação que identificam padrões ou semelhanças entre os dados, podendo assim, encontrar grupos que possuem características mais marcantes que outros. O objetivo deste trabalho é a análise dos resultados de dois algoritmos de classificação não-supervisionado, K-Means e DBSCAN, aplicados sobre dados de clientes de uma empresa de operação de câmbio. O desempenho de cada algoritmo é comparado por meio das métricas de avaliação Soma dos Erros Quadráticos e Coeficiente de Silhueta, que, baseados em cálculos estatísticos, demonstram o quanto aquele modelo de classificação é bom ou não. Apesar de os resultados entre os algoritmos serem próximos, o K-Means foi capaz de classificar e gerar clusters de forma mais conclusiva, com uma melhor distribuição que o DBSCAN que concentrou a maior parte dos dados no mesmo cluster, mesmo este possuindo melhores métricas.

Palavras-chave: Classificação de clientes. Algoritmos de Classificação. Mercado financeiro.

ABSTRACT

Understanding the customer's profile investment is an essential way for companies to be able to create more assertively strategies according to their public. To assist classifying profiles, classification and clustering algorithms are used to identify patterns or similarities between data, thus being able to find groups that show stronger characteristics than others. The objective of this work is to analyze the results of two unsupervised classification algorithms, K-Means and DBSCAN, applied to customer data of a foreign exchange broker. The performance of each algorithm is analyzed through the Sum of Squared Errors and Silhouette Coefficient, which are based on statistical calculations that demonstrate how accurate that classification model is. Although the results between the algorithms are close, K-Means was able to classify and create clusters more conclusively than DBSCAN, with a better distribution, which concentrated most data on the same cluster, even though having better metrics.

1 INTRODUÇÃO

Segundo Rossi (2010), em uma definição simplificada, o mercado de câmbio internacional é o lócus de negociação e troca entre moedas. Por definição, as transações nesse mercado estabelecem as taxas de câmbio ‘spot’ (liquidez imediata) e ‘futura’ (liquidez futura) entre as diversas moedas do sistema internacional.

Assim como em qualquer mercado financeiro, o mercado cambial é imprevisível, complexo e altamente movimentado. Uma operação de câmbio pode ser simplesmente a troca de uma moeda por outra, em papel moeda, por um viajante em um aeroporto de uma metrópole qualquer, como o pagamento de uma importação feita por uma empresa multinacional renomada.

No Brasil, estão autorizados a prestar serviços de câmbio, com aval do Banco Central do Brasil, Bancos e Corretoras de Câmbio, sendo um mercado competitivo entre as instituições pela melhor taxa da moeda oferecida e pela captação e preservação de clientes.

Porém, como é um mercado muito diversificado e abrangente, não existe um nicho específico de cliente que buscam realizar operações de câmbio, diante disto, este trabalho foi desenvolvido, em parceria com a instituição Patacão DTVM, que atua há mais de 50 anos no mercado, para buscar entender e classificar quais os perfis de clientes que buscam este tipo de serviço, qual perfil é mais propenso a realizar uma operação de grande escala, qual perfil tem mais chances de operar diversas vezes, entre outras análises.

Com o avanço da tecnologia e da Internet, está cada vez mais fácil realizar ou solicitar uma operação de câmbio, assim como tentativas de operações maliciosas, influenciando no perfil do cliente ao longo dos anos.

O objetivo deste trabalho é entender qual o perfil de cliente mais valioso para uma instituição do ramo captar, cultivar melhores relações e obter um maior retorno financeiro, visando um maior sucesso do setor comercial da empresa, ao utilizar ferramentas de aprendizado de máquina e análises de dados.

Na seção seguinte, é descrito como outras pesquisas e trabalhos de outros autores e pesquisadores influenciaram nas buscas e resultados deste artigo, assim como qual foi o diferencial para o desenvolvimento da pesquisa proposta neste artigo. Na seção 3 é descrita de forma detalhada todas ferramentas, algoritmos, métodos e bases utilizadas no desenvolvimento deste trabalho, assim como métricas utilizadas na avaliação de cada método utilizado e como interpretá-las. Na seção 4, é discutido todos resultados encontrados, utilizando as ferramentas e métodos descritos na seção 3, comparando-os e classificando quais foram os melhores resultados encontrados e qual seria o ideal para alcançar o objetivo proposto por este trabalho.

2 REFERENCIAL TEÓRICO

A classificação e análise de perfis de clientes é um tema amplamente utilizado e pesquisado nas mais diversas áreas profissionais, para tentar prever o padrão de consumo de um determinado grupo de clientes, entender qual

perfil alvo para ações de marketing e decisões corporativas estratégicas baseadas em análise de dados.

Essa tarefa de segregação e classificação de um grupo de clientes é um assunto explorado mundialmente, incluindo no mercado financeiro internacional, seja por bancos ou outras instituições de câmbio.

Algoritmos de classificação são algoritmos de aprendizado de máquina, supervisionados, semi-supervisionados ou não-supervisionados, utilizados em praticamente todas áreas de atuação, para agrupar dados de uma base com características semelhantes ou algo em comum, possibilitando novos insights e previsões por parte do analista de dados.

Um algoritmo de aprendizado de máquina supervisionado necessita de dados de entradas com suas respectivas saídas, para serem apresentadas ao algoritmo de aprendizagem utilizado durante o processo de treinamento (DE PÁDUA BRAGA, 2007).

Já em um aprendizado não-supervisionado, não há inputs de entradas e saídas, assim, a rede utiliza padrões, regularidades e correlações para agrupar os conjuntos de dados em classes. As propriedades que a rede vai “aprender” sobre os dados podem variar em função do tipo de arquitetura utilizada e da lei de aprendizagem (DE PÁDUA BRAGA, 2007).

Algoritmos de aprendizado de máquina utilizados para fins de categorização e clusterização, são quase majoritariamente não-supervisionados (METZ, 2006) O resultado obtido por meio dos algoritmos de clustering é um conjunto de agrupamentos de dados, no qual cada agrupamento é denominado cluster. Pode-se caracterizar um cluster como sendo um agrupamento composto de um número não fixo de objetos (exemplos) similares, de acordo com uma medida de similaridade.

Wang e Petrounias (2017) desenvolveram uma pesquisa para estudar o comportamento e classificar demograficamente os clientes que utilizam o serviço mobile de um banco chinês, utilizando árvores de decisão e análises cruzadas, para entender seus costumes e direcionar produtos e serviços específicos baseados na análise realizada.

Já Gahlaut e Singh (2017) utilizaram modelos neurais e de regressão para classificar clientes com base em suas características pessoais e costumes para prever seu nível de risco para o banco ao oferecer um serviço como empréstimo ou atribuição de limite de crédito.

Em outro exemplo, similar ao trabalho desenvolvido neste artigo, Aryuni et al. (2018), classificaram clientes de um banco utilizando ferramentas de machine learning semelhantes às deste artigo, K-Means e K-Medoids, comparando o resultado de ambos, porém, seu objetivo foi para prevenção de fraude e clusterização de usuários de mobile banking.

Mesmo que um trabalho semelhante a este tenha sido desenvolvido, os resultados serão divergentes, pois um detalhe como localização demográfica da instituição financeira pode ter um grande peso no perfil de seus clientes.

Neste artigo, tem-se o objetivo de entender e classificar o perfil dos clientes do mercado cambial, um mercado financeiro pouco explorado neste quesito, de uma instituição localizada no coração da maior cidade do hemisfério sul do mundo, para que a equipe comercial e de marketing possam captar mais clientes e tomar ações para alavancar o cadastro e operações com o perfil predominante. Utilizando dois tipos de redes neurais com aprendizado não-supervisionado, buscamos segregar os clientes com base na sua

quantidade de operações, montante operado, taxa média do dólar assim como sua idade.

3 MATERIAIS E MÉTODOS

Para realizar o trabalho de segregação e classificação dos clientes, tanto para clientes pessoas físicas ou jurídicas, realizando operação para fim comercial, foi utilizado majoritariamente o algoritmo de classificação e clusterização não supervisionado K-Means e DBSCAN, com os resultados sendo avaliados pelas métricas SSE e Coeficiente de Silhueta, comentados e demonstrados nos itens a seguir.

3.1 Base de dados

A base de dados utilizada foi provida pela instituição financeira, atuante no mercado de câmbio nacional, Patacão DTVM, empresa atuante no mercado cambial há 50 anos, disponibilizada para fins educativos com a condição da exclusão de qualquer dado sensível que pudesse vulnerabilizar qualquer de seus clientes, parceiros ou colaboradores. As bases de dados foram segregadas em duas variantes:

- Clientes Pessoa Física do mercado Comercial (comércio exterior);
- Clientes Pessoa Jurídica do mercado Comercial (comércio exterior).

As duas bases utilizadas totalizam aproximadamente 1.000 dados de clientes que realizaram alguma operação de câmbio entre os anos de 2017 e 2021. Foi realizado um tratamento dos dados de todas bases, removendo cadastros incompletos, com dados inconclusivos ou irrelevantes para as análises pretendidas, por meio de análises descritivas utilizando ferramentas como boxplot e histogramas, para auxiliar na identificação e investigação de outliers e dados discrepantes.

Foram utilizados os seguintes atributos cadastrais para a análise e categorização dos clientes: CEP, idade (caso aplicável), total em R$ operado no período, média em R$ operado, quantidade de operações realizadas, taxa média em USD.

3.2 Algoritmos de Classificação

Foram utilizados dois tipos de algoritmos de classificação diferentes, ambos não-supervisionados, para posterior comparação de desempenho e resultado. O primeiro, chamado de K-Means, Steinley (2006):É projetado para particionar dados bidirecionais e bimodais (ou seja, N objetos cada um tendo medições em variáveis P) em classes K (C1, C2, ..., CK), onde Ck é o conjunto de Nk objetos no cluster k, e K é dado pelo próprio usuário.

Já o segundo, chamado de “Density-based spatial clustering of applications with noise”, ou DBSCAN, segundo Schubert et al. (2017), traduzido deliberadamente:

O modelo introduzido pelo DBSCAN usa uma estimativa simples do nível de densidade mínimo, com base em um limite para o número de vizinhos,

minPts, dentro do raio ***ε***, com uma distância de medida arbitrária. Objetos com mais de *minPts* vizinhos dentro deste raio ***ε*** incluindo o ponto de consulta são considerados um ponto central.

Resumidamente, são algoritmos utilizados para avaliar, categorizar e agrupar (“clusterizar”) os dados tendo como base características em comum. Para isso, como a base de dados utilizada possui um grande volume, foi realizado um processo de normalização, onde é realizado o balanceamento e escalonagem de todos dados e colunas numéricas analisadas tendo como alvo a padronização dos mesmos, para apresentarem uma dimensão compatível entre si.

3.2.1 K-Means

Como mencionado anteriormente, o K-Means é um algoritmo com aprendizado não supervisionado, proposto por MacQueen em 1967, e utilizado até os dias de hoje como uma ferramenta poderosa em processos de classificação e clusterização, já que ele se adapta bem a um grande número de amostras e tem sido usado em uma grande variedade de áreas de aplicação em muitos campos diferentes. Utilizando o parâmetro de entrada *K*, como o nome diz, que é utilizado para determinar o número de clusters desejado, o algoritmo tentará agrupar uma população *n* da base de dados provida por meio de partições com quantidade *K*.

O “Means” do nome da ferramenta diz respeito ao valor médio das características avaliadas pelo algoritmo, adotadas como centroide dos clusters, o ponto médio central do agrupamento calculado pelo algoritmo.

O valor apropriado de K de uma base de dados é encontrado por meio de uma série de tentativas e erros, ainda mais por ser um valor subjetivo, já que muitas vezes não se possui a noção de quantos clusters é o ideal para repartir uma base de dados. Assim, ao invés de selecionar somente um valor de K, o recomendado é realizar o processamento do algoritmo com uma variedade substancial do parâmetro K, para posteriormente avaliar, através de métricas discutidas posteriormente neste artigo, qual o número de clusters mais apropriado para atender o objetivo pretendido.

O algoritmo K-Means trabalha basicamente da seguinte maneira, seguindo 4 passos:

i. o algoritmo escolhe aleatoriamente os centroides iniciais, utilizando K amostras do conjunto de dados X. Após a inicialização, o K-Means consiste em fazer um *loop* entre as duas próximas etapas.
ii. o algoritmo atribui cada amostra da base de dados ao seu centroide mais próximo.
iii. são criados novos centroides tomando o valor médio de todas as amostras atribuídas anteriormente a cada centroide da segunda etapa.
iv. a diferença entre o antigo e o novo centroide é calculada e o algoritmo repete essas duas últimas etapas até que esse valor seja menor que um limite. Em outras palavras, ele se repete até que os centroides não se movam significativamente.

Porém o K-Means utiliza como métrica de autoavaliação uma variável chamada ‘inércia’. A inércia pode ser reconhecida como uma medida de quão coerentes os clusters são internamente. Podendo influenciar negativamente na avaliação de clusters em bases de dados que possuem uma distribuição mais ‘alongada’ ou irregular.

3.2.2 DBSCAN

Outro algoritmo utilizado, conforme mencionado anteriormente, foi o DBSCAN, um algoritmo que possui a percepção de clusters como áreas de alta densidade separadas por áreas de baixa densidade. Devido a essa visão um tanto genérica, os clusters encontrados pelo DBSCAN podem ter qualquer formato, ao contrário de K-Means, que assume que os clusters têm um formato mais convexo e regular. O componente central do DBSCAN é o conceito de amostras de núcleo, que são amostras que estão em áreas de alta densidade.

O DBSCAN utiliza parâmetros implementados pelo usuário, chamados de ***minPts*** e ***ε,*** cujos, respectivamente possuem os seguintes atributos:

- ***ε*****:** É o tamanho do ‘raio’ a ser considerado como a área ‘vizinha’ em volta do ponto analisado, ou seja, a distância analisada entre um ponto e a presença, ou não, de outro ponto.
- **minPts:** É o número mínimo de pontos necessários para se configurar uma área de densidade relevante, dentro do raio ***ε*** determinado anteriormente.

Pontos que não atingirem os requisitos mínimos imputados pelo usuário são classificados como outliers ou ruído, diferentemente do K-Means, que considera estes pontos como parte de algum cluster, podendo influenciar negativamente em seu resultado.

i. O algoritmo do DBSCAN funciona seguindo os seguintes passos, descritos por Schubert et al. (2017): o banco de dados é verificado linearmente em busca de objetos que ainda não foram processados. Calculando os vizinhos de cada ponto utilizando ***ε*** e **minPts**, identificando os pontos principais.
ii. os pontos não essenciais são atribuídos ao ruído e, quando um ponto central é descoberto, seus vizinhos são expandidos iterativamente e adicionados ao cluster.
iii. os objetos que foram atribuídos a um cluster serão então ignorados quando encontrados mais tarde pela varredura linear.

3.3 Métricas de Avaliação

Para avaliar o desempenho dos algoritmos anteriormente mencionados, são utilizadas métricas de avaliação, ou seja, valores calculados por uma função ou manualmente que refletem a acurácia do valor calculado e previsto pelos algoritmos comparados com o valor real. O valor destas métricas reflete diretamente a qualidade de um modelo, portanto se forem mal escolhidas, será

impossível avaliar se o modelo de fato está atendendo os requisitos desejados.

As métricas utilizadas para avaliar o desempenho de uma clusterização por K-Means são a Soma dos Erros Quadráticos (SSE) e o Coeficiente de Silhueta, detalhados nos itens a seguir:

3.3.1 Soma dos Erros Quadráticos (SSE)

Soma dos Erros Quadráticos (SSE) é um método estatístico usado para medir a diferença entre os dados obtidos pelo modelo de previsão que foi feito anteriormente. SSE é frequentemente usado como uma referência de pesquisa Na determinação de clusters ideais. Segundo Aryuni et al. (2018), é calculado conforme apresentado na Equação 1.

$$SSE = \sum_{i=1}^{n} (y_i - \overline{y_i})^2 \qquad (1)$$

onde y_i é o valor de cada ponto de dados no cluster e $\overline{y_i}$ é o centroide do cluster.

O valor de SSE desejado sempre é o menor possível. Essa métrica é utilizada para determinar qual o melhor número de clusters para a devida base de dados, usando um método conhecido como *"método do cotovelo"* que se baseia na análise da visualização gráfica dos valores de SSE obtidos a cada loop dos algoritmos utilizados, geralmente K-Means ou DBSCAN. Conforme o número de clusters aumenta, o valor de SSE consequentemente diminui, dado que quanto maior a quantidade de clusters menor a distância entre centros. Ao observar uma estabilidade após uma queda constante entre os valores, formando um "cotovelo", assume-se que é a melhor quantidade de clusters para aquela situação. Vide exemplo a seguir:

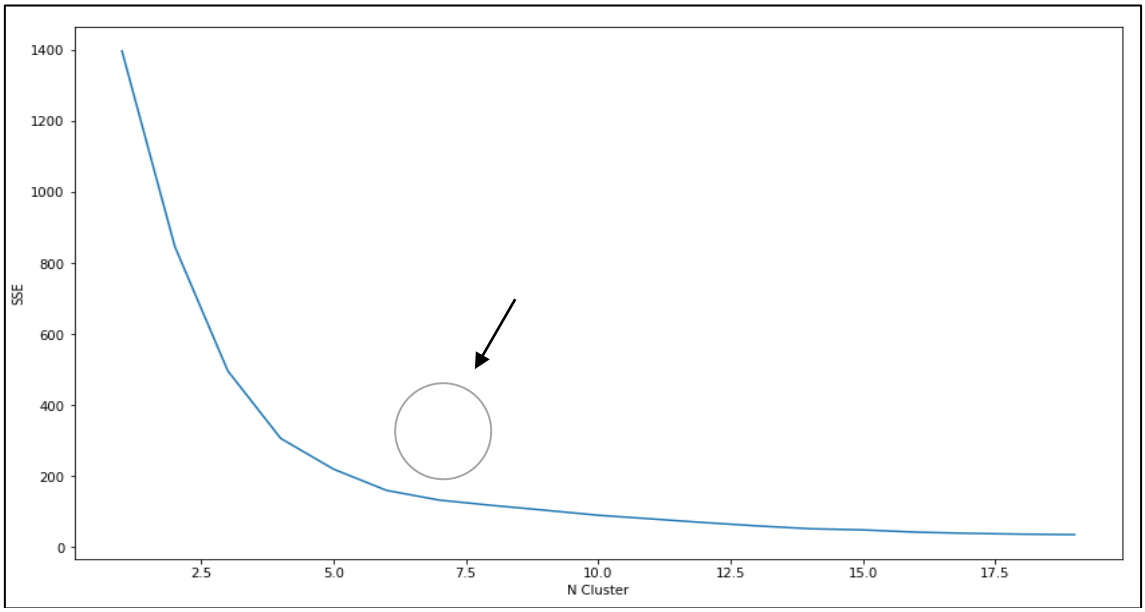

Figura 1 - Exemplo de método do "cotovelo" para definição do número de clusters.

No caso demonstrado acima, o número de cluster ideal, utilizando o método do cotovelo, seria entre 4 e 6.

3.3.2 Coeficiente de Silhueta

O coeficiente de silhueta foi proposto primeiramente por Rousseeuw (1987). A ideia é, através de uma série de cálculos matemáticos, auxiliar o pesquisador a escolher o número ótimo de clusters e, ao mesmo tempo, permitir que se construa uma representação gráfica do agrupamento encontrado.

Para cada objeto i do cluster K, obtém-se um valor denominado s, que reflete a qualidade da alocação dos objetos em cada cluster, calculado pela equação abaixo na Equação 2:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \qquad (2)$$

onde $a(i)$ é a dissimilaridade média do objeto i em relação a todos os objetos do mesmo grupo K1, e $b(i)$ é a dissimilaridade média entre o objeto i em relação a todos os objetos do grupo vizinho mais próximo a ele, grupo K2. s é um valor adimensional e pode variar de -1 a 1, quanto mais próximo de 1, mais bem classificado ele está, e quanto mais próximo a -1, quer dizer que ele foi mal classificado. Caso s esteja próximo de 0, significa que o dado classificado está em um ponto intermediário entre K1 e K2 por exemplo.

Após o cálculo de cada s, é realizado o cálculo médio de todo agrupamento com o devido número de clusters analisados pelo loop, no caso do algoritmo K-Means. Este cálculo da média de todos agrupamentos foi posteriormente denominado por Kaufman e Rousseeuw (1990) como Coeficiente de Silhueta (CS), variando seu valor de 0 a 1, sendo 1 um conjunto de agrupamento ótimo e quanto mais próximo de 0 um agrupamento quase inexistente, conforme demonstrado pela Tabela 1.

Coef. De Silhueta (CS)	**Interpretação**
0,71 a 1,00	Grupos descobertos possuem uma estrutura muito robusta
0,51 a 0,70	Grupos possuem uma estrutura razoável
0,26 a 0,50	Os grupos encontrados possuem uma estrutura fraca e pode ser artificial. É aconselhável tentar outros métodos sobre o conjunto de dados
Menor do que 0,25	Nenhuma estrutura foi descoberta

Tabela 1 - Interpretação Subjetiva do Coef. de Silhueta.

3.4 Comparação dos Algoritmos

Para a execução do estudo, ao utilizar o algoritmo K-Means, foi realizado uma série de execuções do algoritmo, armazenando o resultado de cada iteração, utilizando os parâmetros apresentados na Tabela 2.

K-MEANS		
	Versão do Algoritmo	K-Means ++
	Tamanho dos Clusters analisados	De 2 a 20, incrementando 1 a cada loop.
	Número Máximo de Iterações por Cluster analisado	300

Tabela 2 – Parâmetros utilizados com K-Means.

Assim, foram analisados um total de 19 resultados diferentes, ou seja, 19 clusters de tamanhos diferentes. Os resultados e métricas foram armazenados individualmente para posterior análise e comparação, descritos no Item a seguir.

Ao utilizar o algoritmo DBSCAN, foram realizadas uma série de ajustes e testes, para obter resultados plausíveis e adequados para as bases de dados utilizadas, levando em consideração os atributos escalonados, diferentemente do K-Means, o número de clusters não é sugerido pelo usuário, e sim encontrado pelo algoritmo com base nos parâmetros configurados. Concluiu-se que os parâmetros apresentados na Tabela 3 eram os mais adequados:

DBSCAN		
	Versão do Algoritmo	DBSCAN – SciKit Learn
	Valor de ε	De 0,25 a 4, incrementando 0,125 a cada loop.
	Valor de MinPts	De 2 a 10, incrementando 1 a cada loop.

Tabela 3 – Parâmetros utilizados com DBSCAN.

Deste modo, foram analisados um total de 240 combinações de parâmetros diferentes, resultando em um número variado de clusters. Os resultados e métricas foram armazenados individualmente para posterior análise e comparação, descritos no Item a seguir.

4 RESULTADOS E DISCUSSÃO

Os resultados obtidos utilizando os algoritmos K-Means e DBSCAN, em ambas bases de dados, de clientes de câmbio comercial para pessoas físicas e pessoas jurídicas, serão demonstrados a seguir:

4.1 CLIENTES CÂMBIO COMERCIAL PESSOA FÍSICA

Os resultados a seguir foram obtidos utilizando os algoritmos e parâmetros descritos nos itens anteriores, as Figuras 2 e 3 agregam todas iterações utilizando o algoritmo K-Means, demonstrando a evolução das métricas SSE e Coeficiente de Silhueta utilizando a base de dados de clientes de câmbio comercial pessoas físicas.

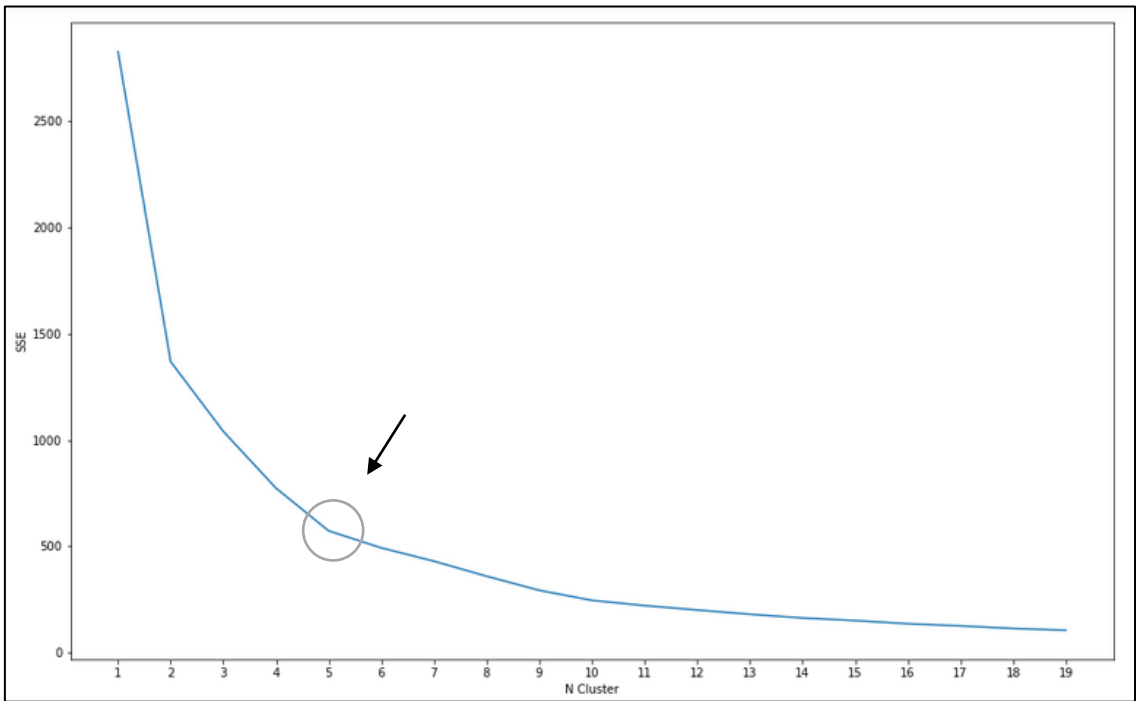

Figura 2 - SSE Câmbio Comercial PF - K-Means

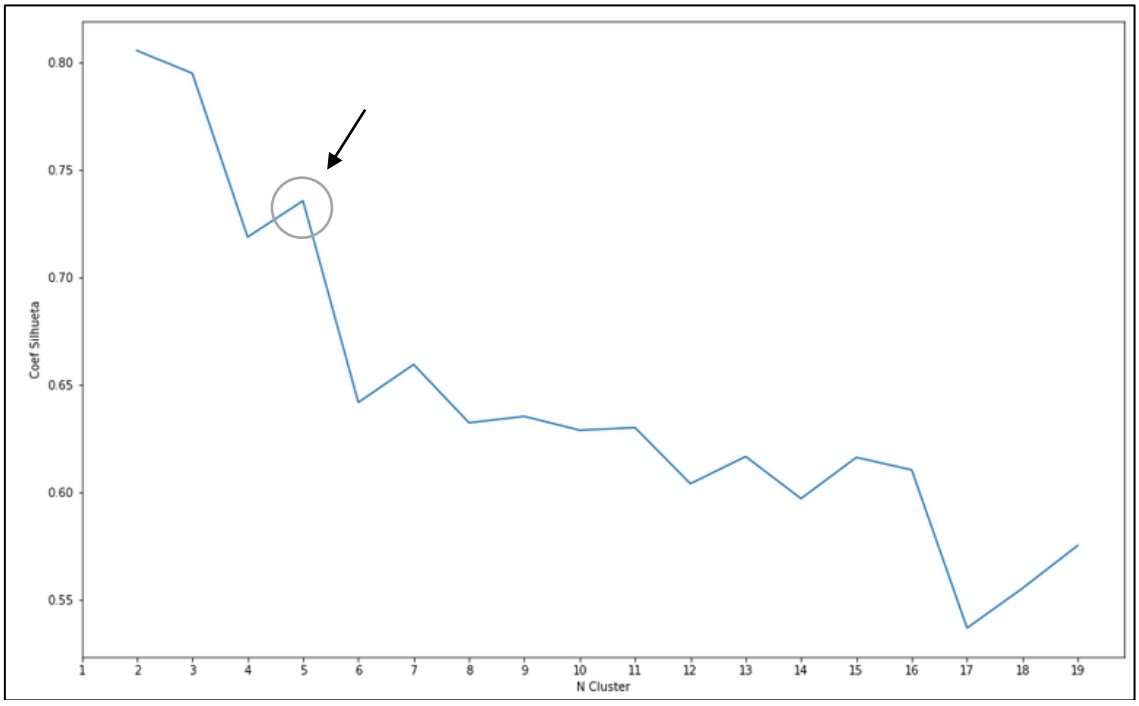

Figura 3 - Coef. Silhueta Câmbio Comercial PF - K-Means

Baseado pelos valores das métricas analisadas, foi determinado que o agrupamento com 5 clusters foi o que melhor performou, adotando este valor como o número de agrupamentos para o grupo de clientes pessoa física que realizaram câmbio comercial. Com um SSE de 573,16 e um Coeficiente de Silhueta de 0,73.

Ao analisar o mesmo grupo de dados porém utilizando o algoritmo DBSCAN, foram obtidos os resultados apresentados na Figura 4, utilizando uma combinação de parâmetros previamente testados.

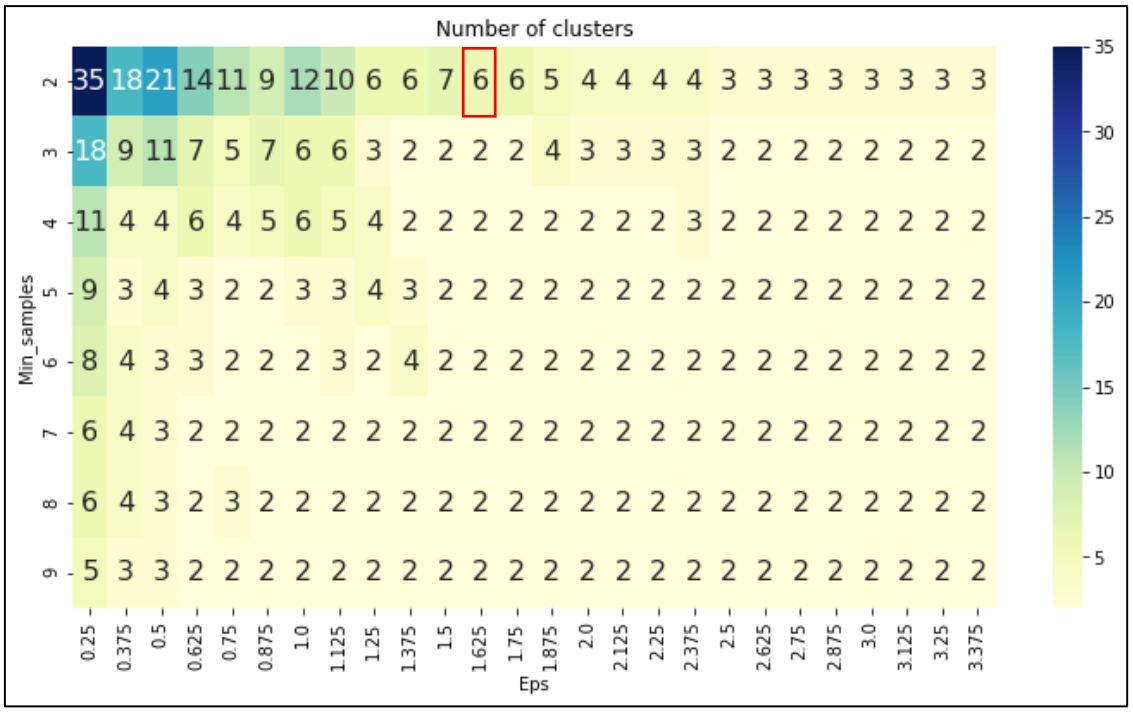

Figura 4 - Iterações DBSCAN Câmbio Comercial PF.

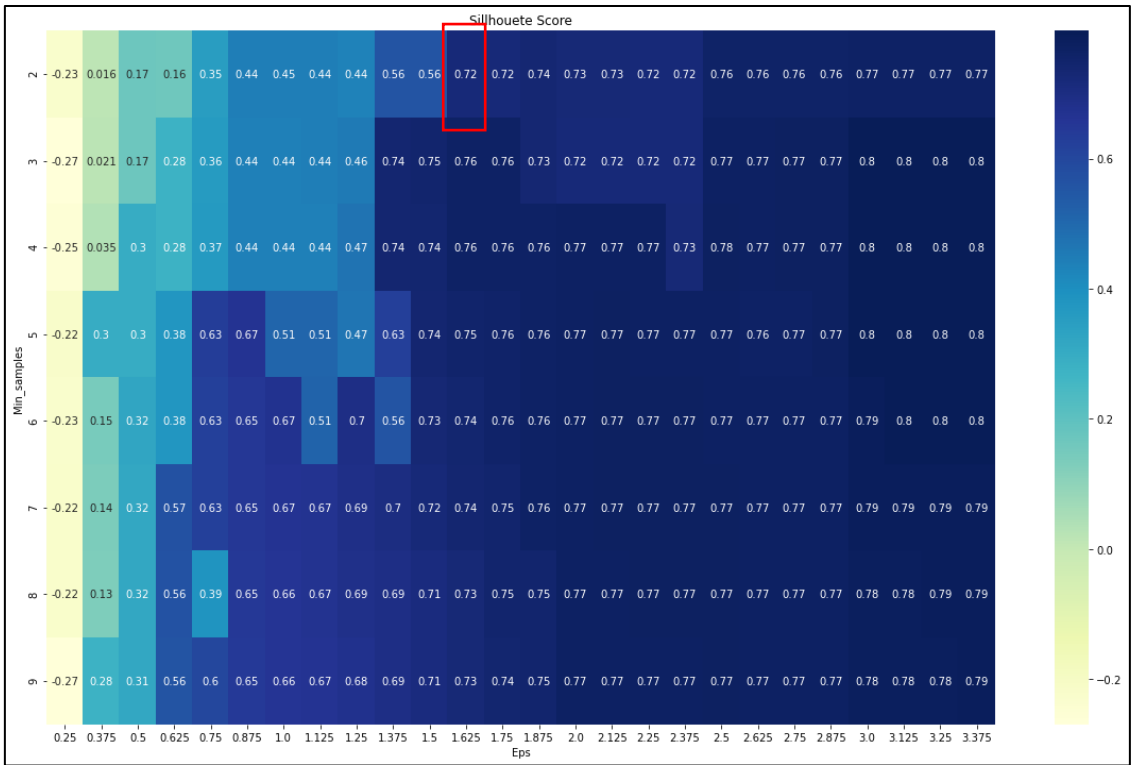

Figura 5 - Coef. Silhueta DBSCAN Câmbio Comercial PF.

As melhores métricas utilizando o algoritmo DBSCAN foi com 6 clusters, para obter um resultado e um cluster comparável com o obtido pelo K-Means, utilizando como parâmetros ***ε = 1.625*** e ***minPts = 2***, que apresentou um valor de Coeficiente de Silhueta aceitável, de 0,72, praticamente semelhante ao K-Means.

Comparando o conteúdo dos clusters de cada algoritmo, obteve-se os resultados apresentados na Tabela 4.

Nº Cluster	**Observações K-Means**	**Observações DBSCAN**
1	5	10 (outliers)
2	478	547
3	43	2
4	14	2
5	24	2
6	-	2

Tabela 4 - Distribuição clusters Câmbio Comercial PF.

Embora as métricas do DBSCAN para a devida base de dados sejam praticamente semelhantes e o algoritmo possibilite a identificação de outliers, podemos observar que a distribuição não é plausível, pois a grande maioria dos dados está concentrado em um único cluster, assim sendo, podemos adotar o algoritmo K-Means como o melhor algoritmo para a situação estudada.

Os atributos de cada grupo estão descritos na Tabela 5, apresentando como destaque as características dos cluster 1, 4 e 6.

Cluster Nº	Quantidade de Operações	Valor em USD Total	Valor em R$ Total	Taxa Média do USD	Valor Médio em R$	Valor Médio em USD	Qtde. Obs.
1	-0.135142	2.419807	2.615401	5.444020	7.225023	6.765599	5
2	-0.235694	-0.301596	-0.295637	4.768837	-0.284482	-0.296751	478
3	0.034646	0.814748	0.740183	4.859749	1.579346	1.780015	43
4	2.724369	4.892188	4.957723	5.181202	1.994695	1.959877	14
5	2.957591	0.985287	0.918494	4.714911	0.084371	0.086671	24

Tabela 5 - Atributos de cada cluster - Câmbio Comercial PF - K-Means.

Assim, mesmo com os valores escalonados, podemos observar que os Clientes dos Clusters 4 e 5 são os que mais efetuaram operações, porém os do Cluster 4 são os que efetuaram operações com maiores montantes, e do Cluster 1 apresentam a maior média no valor por operação, assim como os presentes no Cluster 2 são os que menos operaram, e, se operaram, foram com valores baixos, mesmo possuindo um maior número de observações.

4.2 CLIENTES CÂMBIO COMERCIAL PESSOA JURÍDICA

Assim como no item anterior, as Figuras 6 e 7 demonstram os resultados obtidos utilizando o algoritmo K-Means para a base de dados de câmbio comercial para clientes pessoa jurídica, comparando as métricas SSE e Coeficiente de Silhueta para cada loop do algoritmo:

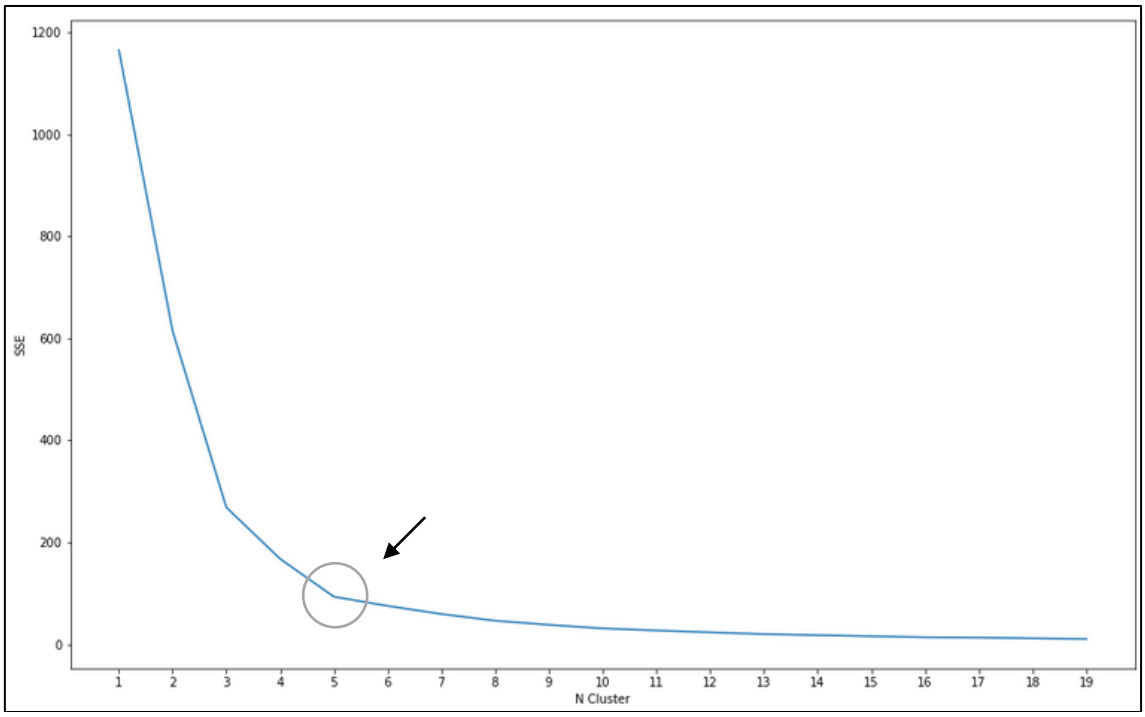

Figura 6 - SSE Câmbio Comercial PJ - K-Means

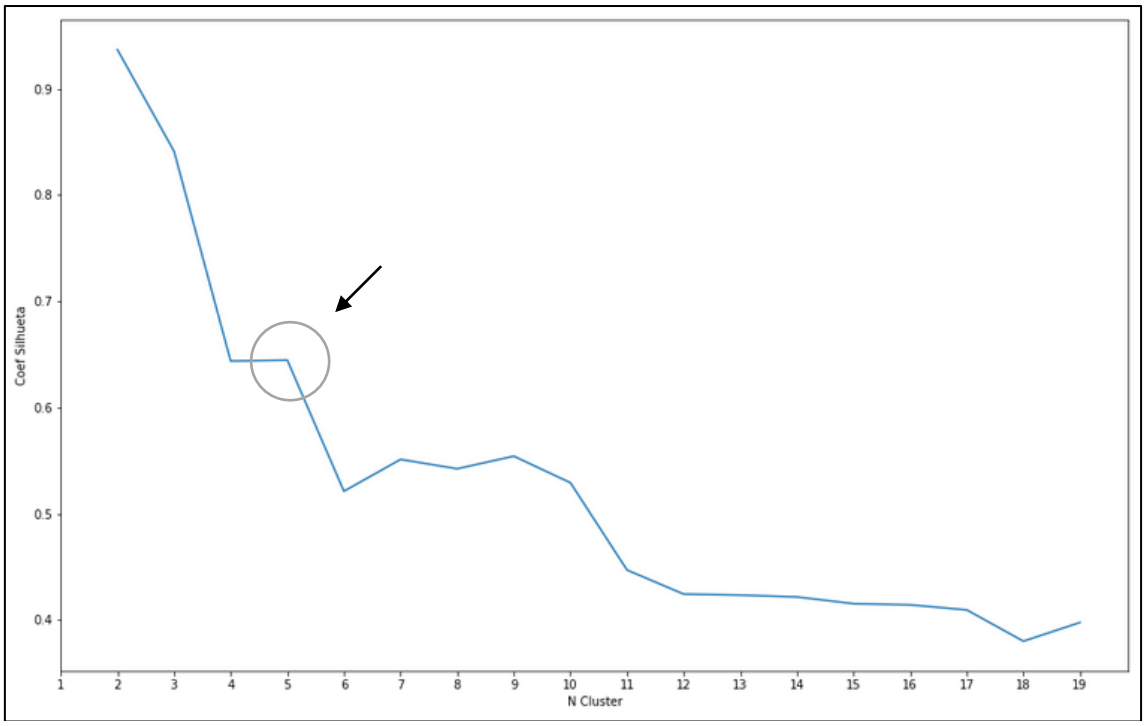

Figura 7 - Coef. Silhueta Câmbio Comercial PJ - K-Means

Coincidentemente, tendo como base os valores das métricas analisadas, foi determinado que o agrupamento com 5 clusters foi o que melhor performou, adotando este valor como o número de agrupamentos para o grupo de clientes pessoa física que realizaram câmbio comercial. Com um SSE de 93,31 e um Coeficiente de Silhueta de 0,64, performando com coeficiente de silhueta inferior comparado ao do grupo de Pessoas Físicas.

Ao analisar o mesmo grupo de dados porém utilizando o algoritmo DBSCAN, foram obtidos os resultados, apresentados nas Figuras 8 e 9, utilizando uma combinação de parâmetros previamente testados.

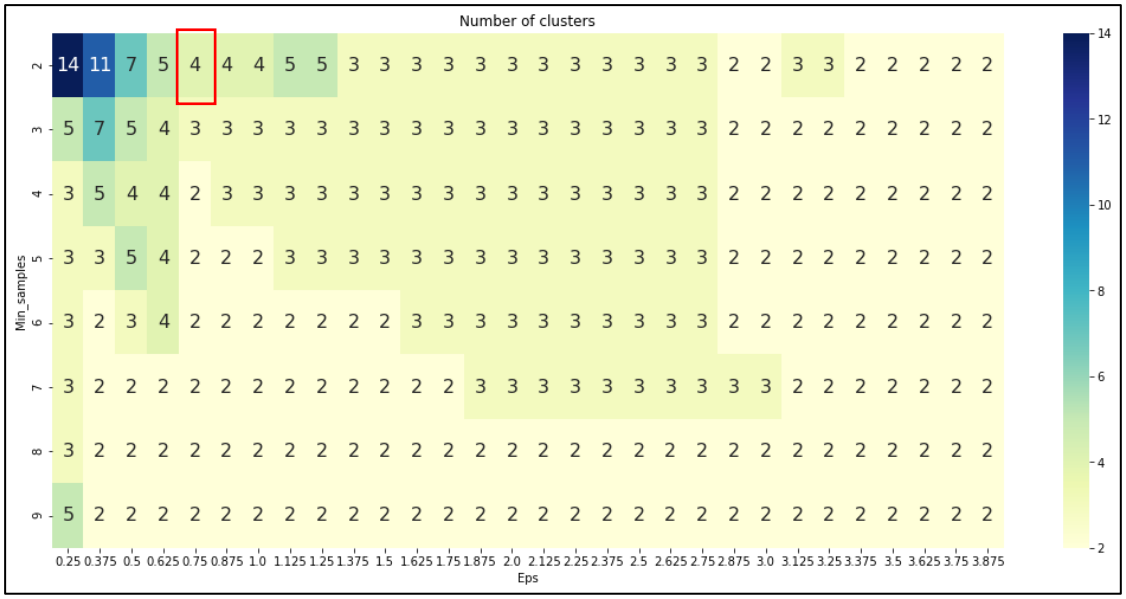

Figura 8 - Iterações DBSCAN Câmbio Comercial PJ.

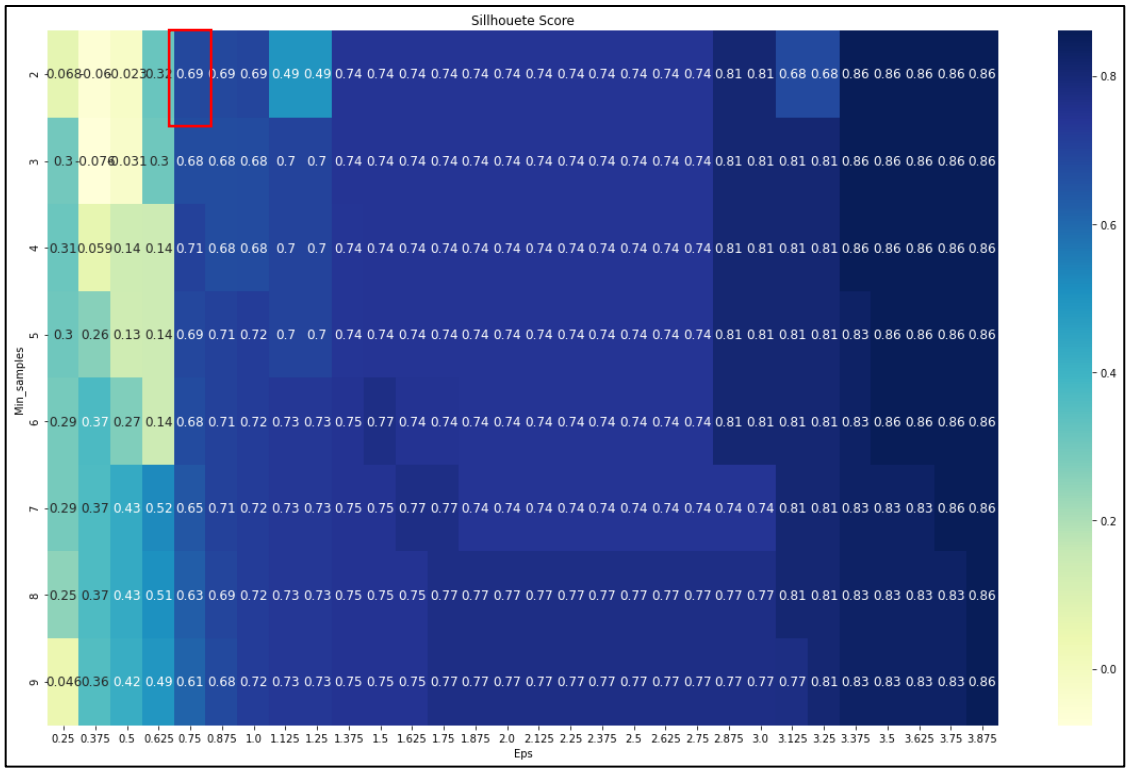

Figura 9 - Coef. Silhueta DBSCAN Câmbio Comercial PJ.

A melhor métrica utilizando o algoritmo DBSCAN foi com 4 clusters, utilizando como parâmetros ***ε = 0.75*** e ***minPts = 2***, gerando um Coeficiente de Silhueta de 0.69, superior ao do K-Means.

A Tabela 6 compara o conteúdo dos clusters de cada algoritmo. Desta vez, as métricas geradas pelo algoritmo DBSCAN apresentam valores mais elevados e possibilita a identificação de outliers, porém, podemos observar que a distribuição não é plausível, pois a grande maioria dos dados está concentrado em um único cluster, assim sendo, podemos adotar o algoritmo K-Means como o melhor algoritmo para a situação estudada.

Nº Cluster	**Observações K-Means**	**Observações DBSCAN**
1	3	10 (outliers)
2	1	217
3	39	4
4	7	2
5	183	-

Tabela 6 - Distribuição clusters Câmbio Comercial PJ.

Os atributos de cada grupo estão detalhados na Tabela 7. Pode se observar a discrepância nos atributos de cada grupo, mesmo com somente uma observação, sendo possivelmente um outlier, o cliente segregado no cluster 2 foi o que mais realizou operações assim como os com maior montante, porém, os encontrados no cluster 4 são os que apresentam uma maior média por operação, os clientes encontrados no Cluster 5, por mais que seja o com maior número de observações, são os que menos operaram e com

os menores valores.

Cluster N°	Quantidade de Operações	Valor em USD Total	Valor em R$ Total	Taxa Média do USD	Valor Médio em R$	Valor Médio em USD	Qtde. Obs.
1	0.228151	3.864917	4.247112	5.131744	1.465890	1.435519	3
2	15.181634	12.861093	12.322715	4.298791	-0.564316	-0.590268	1
3	-0.067133	0.183574	0.192839	4.840395	0.707101	0.787428	39
4	-0.100406	0.071920	0.122292	5.252120	4.721964	4.613841	7
5	-0.068552	-0.175512	-0.182737	4.773047	-0.352263	-0.364606	183

Tabela 7 - Atributos de cada cluster - Câmbio Comercial PJ - K-Means.

5 CONCLUSÃO

O mercado de câmbio é de fato um mercado altamente diversificado, tornando difícil até o trabalho para algoritmos de classificação amplamente utilizados como K-Means e DBSCAN, porém não impossível. Devido ao volume e alta diversidade, esse trabalho só se torna possível com o emprego computacional de mecanismos inteligentes para processamento de classificação. Esta pesquisa analisou, utilizando diferentes técnicas, a possibilidade de classificar duas bases de dados, utilizando atributos e características de clientes, deste mercado.

Como as bases de dados estudadas não possuem dados e atributos muito diversificados, sendo praticamente atributos referentes ao montante de operações ou vezes que o mesmo realizou operações, a classificação possui um viés baseado nestes atributos.

O algoritmo DBSCAN, por mais que possua alguns diferenciais e possa apresentar métricas melhores do que o K-Means, não apresentou um resultado satisfatório, mesmo com uma ampla diferenciação dos parâmetros testados, concentrou os dados em praticamente somente um cluster massivo, não se tornando uma classificação e clusterização satisfatórios ou que seja possível uma análise ou resultado conclusivos.

Por outro lado, o K-Means, mesmo apresentando métricas de valores mais ‘pobres’ e sendo deficiente na identificação de outliers, apresentou resultados conclusivos, segregando clientes com base em suas características de operação.

Por mais que o uso de uma grande quantidade de variáveis, assim como seu escalonamento, impacte negativamente na visualização gráfica dos dados segregados, o uso das duas métricas descritas neste trabalho se demonstraram suficientes para o julgamento da eficácia dos testes realizados utilizando os dois algoritmos, assim como na tomada de decisões.

Este trabalho demonstrou a dificuldade de classificar clientes deste tipo de mercado, evidenciando a necessidade de captar dados mais completos dos clientes, afim de tornar mais impactante o trabalho de classificação, auxiliando na tomada de decisões da diretoria de uma empresa do ramo assim como possibilitando a previsão das características operacionais do cliente.

Como trabalhos futuros, sugere-se a captação de dados mais completos, o uso de bases de dados de outras instituições, com dados demográficos diferentes dos da instituição utilizada, assim como o uso de uma gama de algoritmos maior, como a implementação de um algoritmo hierárquico,

por exemplo.

REFERÊNCIAS

ARYUNI, M.; MADYATMADJA, E. D.; MIRANDA, E. **Customer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering**. 2018.

BRAGA, A. DE P.; LUDERMIR, T. B.; CARVALHO, A. C. P. de L. F. **Redes neurais artificiais: teoria e aplicações**. 2007.

GAHLAUT, A.; KUMAR SINGH, P. **Prediction analysis of risky credit using Data mining classification models**. 2017.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data an Introduction to Cluster Analysis**. 1990.

METZ, J. **Interpretação de clusters gerados por algoritmos de clustering hierárquico**. 2006.

ROSSI, P. **Textos Avulsos - N° 5 - Outubro**. 2010.

ROUSSEEUW, P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis - Journal of Computational and Applied Mathematics**. 1987.

SCHUBERT, E. et al. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. **ACM Transactions on Database Systems**, v. 42, n. 3, 1 jul. 2017.

STEINLEY, D. **K-means clustering: A half-century synthesis.** British Journal of Mathematical and Statistical Psychology, v. 59, n. 1, p. 1–34, maio 2006.

WANG, S.; PETROUNIAS, I. **Big data analysis on demographic characteristics of Chinese mobile banking users**. Proceedings - 2017 IEEE 19th Conference on Business Informatics, CBI 2017. Anais...Institute of Electrical and Electronics Engineers Inc., 17 ago. 2017.