EVENTO ON-LINE
ConBRepro
XIII CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
IA nas Engenharias
29 nov. a 01 de dezembro 2023

Descoberta de Padrões de Comportamento do Cliente em Dados Mistos por Meio de Clusterização

Leonardo Monteiro Ribas
Engenharia de Produção – Universidade Federal do Paraná
Luiz Gustavo Daemme
Engenharia de Produção – Universidade Federal do Paraná
Mariana Kleina
Engenharia de Produção – Universidade Federal do Paraná

Resumo: Este estudo explora a aplicação de algoritmos de clusterização na segmentação de clientes em um cenário de comércio digital em crescimento. Com quase 100 mil pedidos de diversos *marketplaces*, o objetivo é fornecer *insights* para personalização de ofertas e aprimoramento da experiência do cliente. O artigo utiliza os algoritmos de clusterização K-Medoids e DBSCAN, com a métrica de distância de Gower para dados mistos. A avaliação dos resultados é feita por meio do índice de validação Coeficiente de Silhueta. Os resultados destacam a influência significativa da categoria do produto na decisão de compra dos clientes. A pesquisa demonstra a aplicabilidade da clusterização em dados mistos.

Palavras-chave: Clusterização, Segmentação de clientes, Dados mistos, K-Medoids, DBSCAN.

Discovery Customer Behavior Patterns in Mixed Data Through Clustering

Abstract: This study explores the application of clustering algorithms in customer segmentation within a growing digital commerce setting. With nearly 100,000 orders from various marketplaces, the aim is to provide insights for offer personalization and customer experience enhancement. The article employs the clustering algorithms K-Medoids and DBSCAN, along with the Gower distance metric for mixed data. Evaluation of the results is conducted using the Silhouette Coefficient validation index. The findings emphasize the substantial influence of product category on customer purchase decisions. The research demonstrates the applicability of clustering in mixed data.

Keywords: Clustering, Customer segmentation, Mixed data, K-Medoids, DBSCAN.

1. Introdução

O crescimento explosivo do comércio digital nas últimas décadas tem gerado uma enorme quantidade de dados de pedidos e clientes. Compreender e segmentar eficazmente esses dados tornou-se crucial para empresas que buscam otimizar suas estratégias em um mercado com tantos desafios e oportunidades (LIN *et al.*, 2019). Neste estudo, a atenção está concentrada na clusterização de dados como uma ferramenta poderosa para

categorizar e segmentar clientes a nível nacional com base em produtos encomendados e suas faixas de preço.

O objetivo é explorar como a clusterização de dados pode ser aplicada a um conjunto de dados de quase 100 mil pedidos em diversos *marketplaces*. A base de dados é formada por variáveis mistas, que neste estudo abrangem variáveis contínuas e binárias. Ao identificar grupos semelhantes de pedidos, pretende-se oferecer *insights* valiosos para personalização de ofertas e o aprimoramento da experiência do cliente (BARMAN; CHOWDHURY, 2019).

Este artigo está dividido em cinco seções, contando com esta. A próxima seção apresenta a revisão teórica que sustenta a pesquisa. Em seguida, é detalhada a metodologia utilizada para conduzir o estudo e, posteriormente, são discutidos os resultados obtidos e as avaliações feitas com base nessas descobertas. Na seção subsequente, tem-se as conclusões derivadas desses resultados, resumindo as principais contribuições deste trabalho.

2. Referencial Teórico

2.1 Algoritmos de Clusterização

Os algoritmos de clusterização são técnicas essenciais de análise de dados que objetivam identificar grupos naturais dentro de conjuntos de dados. Esses grupos, os chamados clusters, são formados com base na similaridade entre os elementos de dados. Neste trabalho, serão abordadas duas técnicas de clusterização: os algoritmos K-Medoids e DBSCAN.

2.1.1 K-Medoids

O K-Medoids é um algoritmo baseado em centroides: os pontos mais centrais em cada cluster e que os representam. Ao contrário do K-Means, método similar mais comumente utilizado, o K-Medoids utiliza pontos de dados reais como centroides (os medoides), o que o torna menos sensível a *outliers*. Além disso, este algoritmo é mais flexível, permitindo o uso de distâncias não-euclidianas em seus cálculos (DE ASSIS; DE SOUZA, 2011).

Embora haja diversos algoritmos para a clusterização via K-Medoids, o Particionamento em Torno de Medoides (PAM), proposto por Kaufman e Rousseeuw (1990), é considerado o mais poderoso dentre os existentes (PATEL; SINGH, 2013).

Reynolds *et al.* (2004) resumem o algoritmo da seguinte forma:

a) Selecione k objetos aleatoriamente para serem medoides iniciais dos clusters;

b) Atribua cada objeto ao seu medoide mais próximo;

c) Recalcule as posições dos k medoides;

d) Repita as etapas *b* e *c* até que os medoides se tornem fixos.

2.1.2 DBSCAN

O DBSCAN (*Density Based Spatial Clustering of Application with Noise*) é um algoritmo de clusterização baseado na densidade de pontos, especialmente adequado na detecção de clusters de diferentes densidades dentro de conjuntos de dados, ou em dados muito ruidosos (ESTER *et al.*, 1996).

O algoritmo requer a definição de dois parâmetros principais: o raio de vizinhança (*Eps*) e o número mínimo de pontos (*MinPts*) para formar um cluster. Os pontos de dados são classificados como centrais, de fronteira ou de ruído, dependendo de sua densidade local em relação a esses parâmetros. O processo de clusterização envolve a identificação de

pontos centrais e a expansão de clusters a partir deles, conectando pontos alcançáveis dentro do raio de vizinhança *Eps*. Pontos que não pertencem a nenhum cluster são considerados ruídos.

2.2 Distância de Gower

Quando tratando dados com variáveis mistas, não é possível utilizar métricas como a distância Euclidiana, que só lidam com dados contínuos (HUANG, 1998). A distância de Gower supre essa necessidade, sendo capaz de lidar com conjuntos de dados mistos (GOWER, 1971).

A distância é calculada pela equação (1).

$$S_{ij} = \frac{\sum_{k=1}^{p} W_k S_k}{\sum_{k=1}^{p} W_k} \quad (1)$$

Em que S_{ij} é a distância entre os elementos x_i e x_j, com $i \neq j$. Se a *k*-ésima variável é qualitativa, tem-se S_k dado pela equação (2).

$$S_k = \begin{cases} 0, se\ x_{ki} = x_{kj} \\ 1, se\ x_{ki} \neq x_{kj} \end{cases} \quad (2)$$

Caso a *k*-ésima variável for quantitativa, S_k será dado pela equação (3).

$$S_k = \frac{|x_{ki} - x_{kj}|}{\max(x_k) - \min(x_k)} \quad (3)$$

Em que:

— k: 1, 2, ..., p;

— p: número total de variáveis;

— x_{ki}: é o valor da *k*-ésima variável para o elemento i;

— i: 1, 2, ..., n;

— n: número de observações;

— W_k: é igual a 1 (um) quando se tem os valores da *k*-ésima variável para ambos os elementos e 0 (zero), quando não se tem os valores da *k*-ésima variável para quaisquer dos dois elementos.

2.3 Coeficiente de Silhueta

O Coeficiente de Silhueta é uma métrica utilizada para avaliar a qualidade dos clusters obtidos por algoritmos de clusterização, ao medir o quão bem separados e coesos os clusters estão. Para cada observação é definido um índice que varia de -1 a 1, e a média do índice de todas as observações no cluster é o que indica a sua homogeneidade. No Quadro 1, tem-se a demonstração de como podem ser interpretados os diferentes valores assumidos pelo coeficiente (KAUFMAN; ROUSSEEUW, 1990).

Quadro 1 – Interpretação do Coeficiente de Silhueta

Coeficiente de Silhueta	Interpretação
0,71 a 1,00	Estrutura forte
0,51 a 0,70	Estrutura razoável
0,26 a 0,50	Estrutura fraca
Menor que 0,25	Nenhuma estrutura

Fonte: Kaufman e Rousseeuw (1990)

O índice é calculado utilizando a distância média *intra* e *inter* clusters, como explicitado na equação (4):

$$Silhueta = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (4)$$

Em que:

— a_i: é a distância média entre a observação i às demais observações do cluster;

— b_i: é a distância média entre a observação i e todas as observações do cluster mais próximo.

Para este trabalho, a métrica utilizada para o cálculo do Coeficiente de Silhueta foi a Distância de Gower.

2.4 Transformações

2.4.1 Transformação Box-Cox

A transformação Box-Cox (BOX; COX, 1964) é uma técnica estatística que visa transformar a variável *target* para apresentar um comportamento mais próximo à normalidade, facilitando a construção de testes de hipótese, cálculo para número de amostras, intervalos de confiança e afins. A equação (5) define a transformação Box-Cox.

$$y(\lambda)\begin{cases}\frac{y^{\lambda}-1}{\lambda}, se\ \lambda \neq 0 \\ \log y\,,\ \ se\ \lambda = 0\end{cases} \quad (5)$$

Em que:

— y: é a variável original (*target*) que se deseja transformar;

— λ: é o parâmetro da transformação, estimado a partir dos dados ou inserido manualmente.

2.4.2 Transformação Yeo-Johnson

A transformação Yeo-Johnson é uma extensão da transformação Box-Cox, proposta por Yeo e Johnson (2000), que permite tratar dados que incluem valores negativos e zero. A equação (6) expressa a transformação Yeo-Johnson.

$$y(\lambda)\begin{cases}\frac{(y+1)^{\lambda}-1}{\lambda}, se\ \lambda \neq 0, y \geq 0 \\ \log(y+1)\,,\ \ se\ \lambda = 0, y \geq 0 \\ \frac{(1-y)^{2-\lambda}-1}{\lambda-2}, se\ \lambda \neq 2, y < 0 \\ -\log(1-y)\,,\ \ se\ \lambda = 2, y < 0\end{cases} \quad (6)$$

Em que:

— y: é a variável original (*target*) que se deseja transformar;

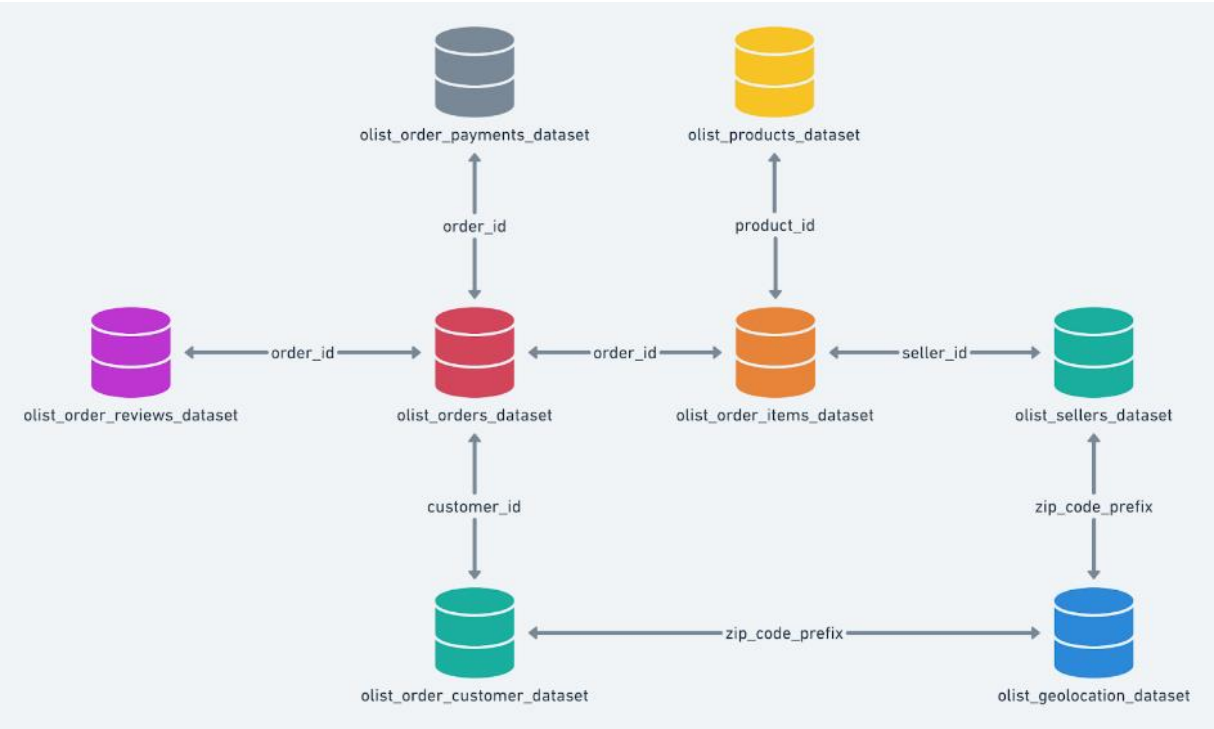
— λ: é o parâmetro da transformação, estimado a partir dos dados ou inserido manualmente.

3. Metodologia

A metodologia aqui apresentada foi aplicada em um conjunto de dados extraído do *website Kaggle*, de quase 100 mil encomendas reais feitas entre 2016 e 2018 na plataforma brasileira de *e-commerce Olist Store*, que integra diversos *marketplaces* no país. O objetivo é agrupar esses diversos pedidos e seus clientes de forma a tornar possível a elaboração de um *marketing* de relacionamento específico para cada grupo.

Inicialmente, os dados estavam espalhados em oito conjuntos de dados distintos, interligados entre si por diversas variáveis, conforme indicado na Figura 1. Em um preparo inicial dos dados, foi necessário realizar junções externas (ROCKOFF, 2016) a fim de criar um conjunto de dados com todas as variáveis de interesse para a clusterização dos pedidos, detalhados no Quadro 2.

Figura 1 – Relação entre os conjuntos de dados iniciais



Fonte: *Kaggle* (2021)

Quadro 2 – Detalhamento das variáveis

Nome da Variável	Tipo de Variável	Descrição/Categorias
Número de itens	Numérica	Número de produtos comprados
Preço	Numérica	Preço dos produtos comprados em R$
Frete	Numérica	Preço do frete em R$
Valor total	Numérica	Valor total da compra em R$
Categoria do produto	Categórica	Moda e acessórios
		Arte e entretenimento
		Eletrônicos e tecnologia
		Casa e decoração
		Lazer e hobbies
		Construção e ferramentas
		Livros e papelaria
		Alimentação

		Outros
Região de residência do cliente	Categórica	Centro-Oeste
		Nordeste
		Norte
		Sudeste
		Sul

Fonte: Os autores (2023)

Após obter um conjunto único de dados, pode-se observar a presença de dados numéricos e categóricos. Desta forma, antes da clusterização, faz-se necessário tratar os dados de modo que seja possível aplicar a distância de Gower.

O atributo “Categoria do produto” foi dicotomizado, ou seja, cada uma de suas categorias foi transformada em uma variável binária (*dummy*), de valor 0 (ausência) ou 1 (presença). Vale notar: é possível que uma mesma observação tenha valor 1 (presença) em mais de uma categoria, como mostrado na Tabela 1. Portanto, das nove categorias da variável, formou-se nove variáveis *dummies*.

Tabela 1 – Exemplo das categorias do atributo “Categoria do produto” dicotomizadas em *dummies*

Categoria(s)	Moda e acessórios	Arte e entretenimento	...	Outros
Moda e acessórios	1	0	...	0
Arte e entretenimento	0	1	...	0
Lazer e hobbies	0	0	...	0
Moda e acessórios, Outros	1	0	...	1

Fonte: Os autores (2023)

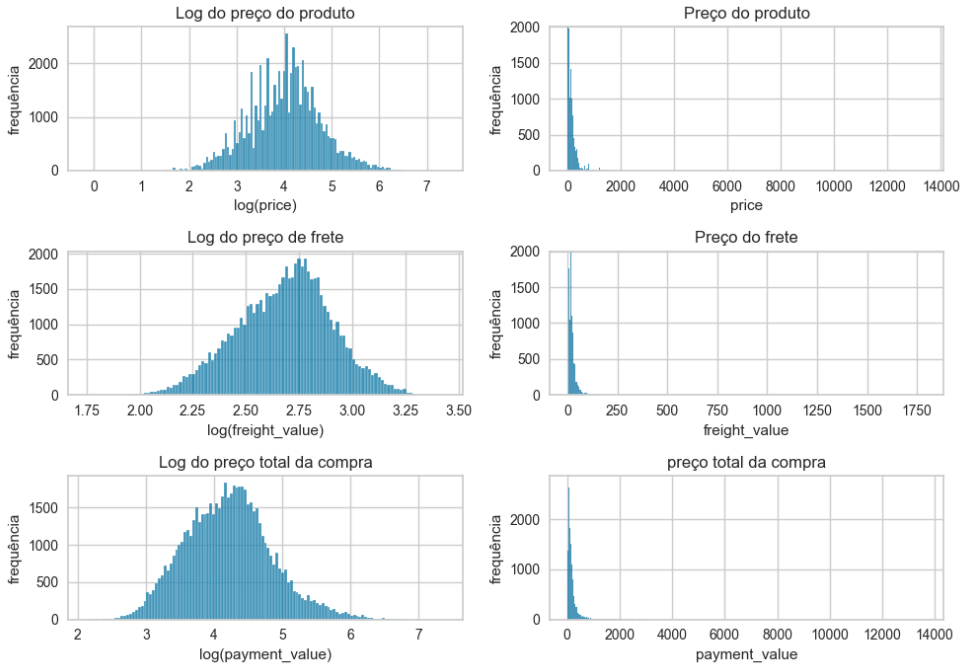
O atributo “Região de residência do cliente” não passou pelo mesmo processo, pois não será utilizado para a aplicação dos algoritmos de clusterização, apenas para a separação do conjunto de dados em estratos, como será abordado mais adiante nesta seção.

Devido a suas distribuições assimétricas à direita, os dados numéricos “Preço” e “Valor total” passaram por uma transformação Box-Cox a fim de tomarem uma aparência assumidamente normal. Pela mesma razão e com o mesmo objetivo, os dados do atributo "Frete" foram submetidos a uma transformação Yeo-Johnson, capaz de lidar com os valores zero em algumas observações. O comparativo dos dados antes e depois das transformações pode ser visto na Figura 2.

Após isso, todos os dados numéricos foram padronizados, isto é, os valores foram subtraídos da média e divididos pelo desvio padrão (MONTGOMERY, 2016), e as observações acima ou abaixo de 3 desvios-padrão em qualquer das variáveis, consideradas *outliers* (MONTGOMERY, 2016), foram removidas da população. Então, as observações restantes foram retornadas a seus valores pré-padronização.

Com a versão final do conjunto de dados, extensivamente tratado e com 13 variáveis, já há todas as informações necessárias para se calcular a matriz de distância de Gower. O tamanho do conjunto de dados, porém, com 100 mil observações, exigiria um tempo de processamento que tornaria esse cálculo inviável.

Figura 2 – Comparativo entre os dados numéricos tratados (à esquerda) e dados originais (à direita)



Fonte: Os autores (2023)

A fim de mitigar esse problema, buscou-se formas de reduzir o tamanho do conjunto de dados. Inicialmente, foram eliminadas as linhas duplicadas — ou seja, que continham dados iguais e que não acrescentavam informação ao conjunto. Em seguida, foi retirada uma amostra da população, utilizando o processo de amostragem estratificada proporcional. Neste caso, as observações foram estratificadas segundo a variável "Região da residência do cliente", com o intuito de manter a proporcionalidade dos dados, detalhada na Tabela 2.

Tabela 2 – Proporção dos dados entre as regiões do Brasil

Região	Proporção
Centro-Oeste	7,54%
Nordeste	11,91%
Norte	2,66%
Sudeste	61,53%
Sul	15,71%
Total	100,00%

Fonte: Os autores (2023)

Dentro de cada estrato, mantendo um intervalo de confiança de 99% e uma margem de erro de 2 pontos percentuais para mais ou para menos, e seguindo a proporção supracitada, foi obtida uma amostra total de 9.182 observações.

Após isso, foi retirada a variável categórica “Região de residência do cliente”, pois não seria utilizada na aplicação dos algoritmos de clusterização.

Utilizando a amostra como o novo conjunto de dados principal, calculou-se a matriz de distância de Gower, exemplificada na Tabela 3, que é usada como entrada para os algoritmos de clusterização.

Tabela 3 – Distância de Gower aplicada nas cinco primeiras observações

Observação	1	2	3	4	5
1	0,000000	0,167471	0,154687	0,154773	0,169574
2	0,167471	0,000000	0,166676	0,168398	0,002610
3	0,154687	0,166676	0,000000	0,155568	0,168733
4	0,154773	0,168398	0,155568	0,000000	0,170430
5	0,169574	0,002610	0,168733	0,170430	0,000000

Fonte: Os autores (2023)

A matriz de distância foi então utilizada como entrada pelo algoritmo K-Medoids (PAM) para a determinação dos clusters. Importante ressaltar que a escolha do algoritmo de clusterização K-Medoids (em detrimento ao K-Means, por exemplo) se deu pelo fato da permissibilidade do uso de distâncias não-euclidianas. O parâmetro de entrada exigido pelo algoritmo, o número k de clusters, foi definido mediante a aplicação do Coeficiente de Silhueta para os resultados do algoritmo em um intervalo de k entre 5 e 15 clusters. Este foi o intervalo utilizado devido à interpretabilidade e à aplicabilidade dos clusters — como serão utilizados para criar estratégias de marketing, fugir desse intervalo pode dificultar sua aplicação.

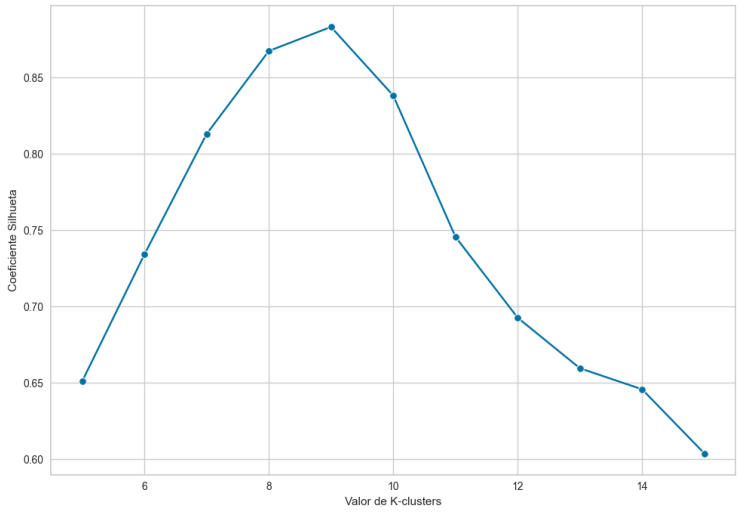
Da mesma maneira, a matriz de distância foi utilizada como entrada pelo algoritmo DBSCAN. O parâmetro *MinPts* foi calculado a partir da relação $MinPts = 2 \times D$ (SANDER *et al*., 1998), sendo D a dimensão do conjunto de dados. Portanto, $MinPts = 26$. O parâmetro *Eps* foi definido por meio da aplicação do Coeficiente de Silhueta para os resultados do DBSCAN em um intervalo de *Eps* entre 0,0025 e 0,02, variando em uma escala de 0,0025.

4. Resultados e Avaliação

Para se determinar qual o melhor agrupamento dentro dos algoritmos, utilizou-se o Coeficiente de Silhueta.

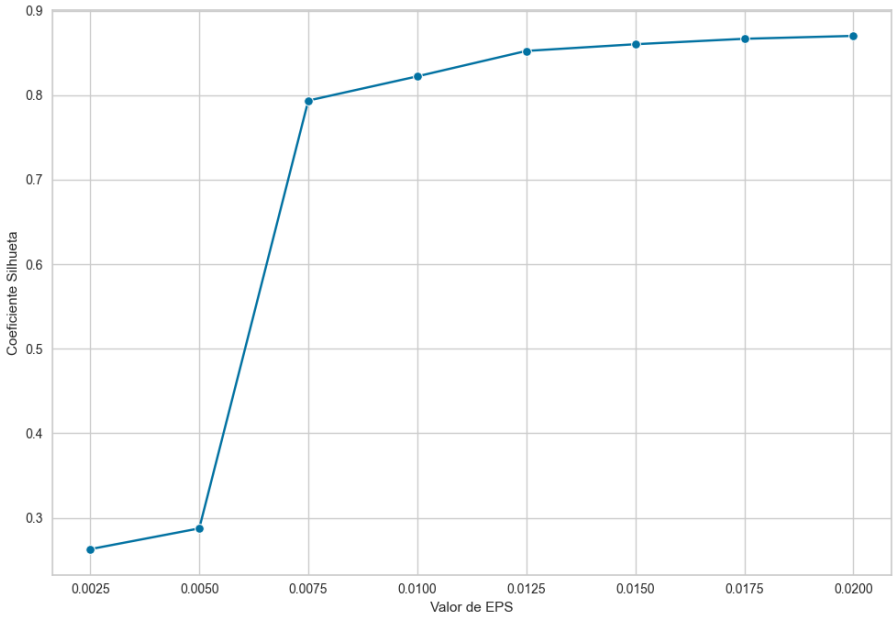
Dentre os testes para os algoritmos K-Medoids e DBSCAN, conforme indicado nas Figuras 3 e 4, respectivamente, os resultados mais adequados (ou seja, aqueles com o maior Coeficiente de Silhueta) são respectivamente os que utilizaram $k = 9$ e $Eps = 0{,}02$ (que também gerou 9 clusters).

Figura 3 – Coeficiente de Silhueta aplicado aos resultados do K-Medoids para cada k



Fonte: Os autores (2023)

Figura 4 – Coeficiente de Silhueta aplicado aos resultados do DBSCAN para cada *Eps*



Fonte: Os autores (2023)

Na Tabela 4, explicita-se um comparativo final entre o melhor resultado dentro de cada um dos algoritmos. Ambos os resultados indicam uma estrutura de clusters bastante significativa.

Tabela 4 – Comparativo entre K-Medoids e DBSCAN

	K-Medoids	DBSCAN
Coeficiente de Silhueta	0,883196	0,870065

Fonte: Os autores (2023)

Como o Coeficiente de Silhueta do resultado obtido do algoritmo K-Medoids foi o mais próximo de 1, este foi o agrupamento empregado adiante, em uma avaliação descritiva das características de cada cluster gerado.

A Tabela 5 demonstra que a quantidade de clientes agrupados em cada cluster não é homogênea, havendo clusters mais ou menos representativos. Isso se dá devido à forma como as variáveis *dummies* se distribuíram dentro de cada grupo, detalhada nas Tabelas 6 e 7.

Tabela 5 – Quantidade de observações por cluster

Cluster	Quantidade de Observações	Proporção
1	1.938	21,11%
2	2.280	24,83%
3	1.258	13,70%
4	1.287	14,02%
5	1.001	10,90%
6	520	5,67%
7	505	5,50%
8	308	3,35%
9	85	0,92%
Total	9.182	100,00%

Fonte: Os autores (2023)

Tabela 6 – Alocação das variáveis *dummies* por cluster

Cluster	Moda e Acessórios	Arte e Entretenimento	Eletrônicos e tecnologia	Casa e decoração	Lazer e *hobbies*
1	2	1	1	3	3
2	2	0	2	2.280	3
3	0	0	2	0	1.258
4	1.287	0	0	2	0
5	0	1	1.001	0	1
6	1	520	0	2	0
7	0	0	2	8	1
8	0	0	0	0	0
9	0	0	0	0	0

Fonte: Os autores (2023)

Tabela 7 – Alocação das variáveis *dummies* por cluster (continuação)

Cluster	Construção e ferramentas	Livros e papelaria	Alimentação	Outros
1	2	0	0	1.938
2	1	1	2	6
3	0	1	0	0
4	0	3	1	1
5	0	0	0	2
6	0	0	0	8
7	505	0	0	4
8	0	308	0	1
9	0	0	85	0

Fonte: Os autores (2023)

Nota-se uma clara segmentação dos dados dentro de cada cluster. Cada uma das nove categorias apresentou maior participação em cada um dos nove grupos, embora nenhuma delas tenha se limitado exclusivamente a um único cluster.

A Tabela 8 apresenta a média das demais variáveis usadas na clusterização, que são as variáveis contínuas. Nota-se que o número médio de itens comprados não difere muito nos 9 clusters. Já o preço médio dos itens comprados varia de R$ 86,50 (cluster 9) a R$ 197,94 (cluster 6), indicando que a amostra gasta menos com “Alimentação” e mais com “Arte e Entretenimento”. O preço médio de frete é ligeiramente maior para os clusters 2 e 7, que são os clusters formados pelas categorias “Casa e decoração” e “Construção e ferramentas”, respectivamente. Finalmente, o valor total médio da compra (valor dos itens + frete) tem a mesma interpretação da variável preço médio dos itens.

Tabela 8 – Resumo dos clusters para as demais variáveis

Cluster	Média do número de itens	Preço médio dos itens	Preço médio do frete	Valor total médio da compra
1	1,12	R$ 163,59	R$ 25,17	R$ 188,76
2	1,27	R$ 138,96	R$ 32,24	R$ 171,20
3	1,13	R$ 137,67	R$ 25,65	R$ 163,32
4	1,15	R$ 145,79	R$ 23,21	R$ 169,00
5	1,21	R$ 172,65	R$ 25,95	R$ 198,60
6	1,11	R$ 197,94	R$ 27,24	R$ 225,18
7	1,27	R$ 180,81	R$ 32,47	R$ 213,28
8	1,16	R$ 106,10	R$ 22,59	R$ 128,69
9	1,32	R$ 86,59	R$ 24,88	R$ 111,47

Fonte: Os autores (2023)

Com isso, fica evidente que a variável de maior impacto na decisão de compra é a “Categoria do produto”. Esta análise revela que, quase que independentemente das variações em outras características, como “Valor total” e “Número de itens”, a variável categórica se destaca como o principal fator de influência na escolha dos clientes.

5. Conclusões

Este artigo apresenta a aplicação de dois algoritmos de clusterização para o agrupamento de clientes, quando os dados de entrada apresentam variáveis mistas. Assim, demonstra-se como os processos de clusterização, em união à métrica de Gower, podem ser aplicados em uma base de dados híbrida. É apresentado também o tipo de tratamento que deve ser aplicado ao conjunto de dados para que não haja perdas no momento de aplicação dos algoritmos.

Como resultado, as clusterizações apesentaram 9 grupos de clientes para ambos os algoritmos utilizados, com característica de segmentação baseado na categoria dos produtos comprados. Com estes agrupamentos, é possível elaborar campanhas de *marketing* pautadas nas categorias dos produtos, tendo em mente que as outras variáveis têm menos significância.

Para trabalhos futuros, tendo ciência de que a categoria do produto, a nível macro, enviesa a clusterização, recomenda-se trabalhar com produtos específicos e analisar relacionamentos de associação para recomendação de compra do cliente.

Referências

BARMAN, D.; CHOWDHURY, N. A novel approach for the customer segmentation using clustering through self-organizing map. **International Journal of Business Analytics**, v. 6, n. 2, p. 23–45, 2019.

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 26, n. 2, p. 211-243, 1964.

DE ASSIS, E. C.; DE SOUZA, R. M. C. R. A K-medoids clustering algorithm for mixed feature-type symbolic data. **Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics**, p. 527–531, 2011.

ESTER, M.; KRIEGEL, H. P.; SANDER, X.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **KDD**. 1996. p. 226-231.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, p. 857-871, 1971.

HUANG, Z.. Extensions to the k-means algorithm for clustering large data sets with categorical values. **Data mining and knowledge discovery**, v. 2, n. 3, p. 283-304, 1998.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data.** [s.l.] Hoboken: Wiley Online Library, 1990.

LIN, Q.; ZHANG, H.; WANG, X.; XUE, Y.; LIU, H.; GONG, C. A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers. **IEEE ACCESS**, v. 7, p. 28696–28711, 2019.

MONTGOMERY, D. C. **Introdução ao Controle Estatístico da Qualidade**. Rio de Janeiro: Ed. LTC, 2016.

PATEL, A.; SINGH, P. New Approach for K-mean and K-medoids Algorithm. **International Journal of Computer Applications Technology and Research**, v. 2, n. 1, p. 1-5, 2013.

REYNOLDS, A. P.; RICHARDS, G.; RAYWARD-SMITH, V. J. The application of k-medoids and pam to the clustering of rules. In: **Intelligent Data Engineering and Automated Learning–IDEAL 2004: 5th International Conference, Exeter, UK. August 25-27, 2004. Proceedings 5**. Springer Berlin Heidelberg, 2004. p. 173-178.

ROCKOFF, Larry. **The language of SQL:** 2nd Edition. United States of America: Ed. Addison-Wesley Professional, 2021.

SANDER, J.; ESTER, M.; KRIEGEL, H. P.; XU, X. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. **Data mining and knowledge discovery**, v. 2, p. 169-194, 1998.

YEO, I. K.; JOHNSON, R. A. A new family of power transformations to improve normality or symmetry. **Biometrika**, v. 87, n. 4, p. 954-959, 2000.