Segmentação de clientes em E-commerce utilizando K-means para Personalização e Fidelização

Lucas Fortolan Sampaio
Engenharia da Computação
Fundação Hermínio Ometto (FHO)
Araras, Brasil
lucas.fortolan@alunos.fho.edu.br

***Resumo*—O crescimento do e-commerce tem gerado grande volume de dados sobre consumidores, mas muitas empresas ainda enfrentam desafios para transformá-los em insights estratégicos de retenção e fidelização. Este trabalho propõe o uso de Inteligência Artificial Não Supervisionada, especificamente o algoritmo K-means, para segmentar clientes com base em histórico de compras. O objetivo é identificar grupos de clientes de alto valor e engajamento, permitindo ações personalizadas de marketing, aumento do ticket médio e melhoria do relacionamento com o cliente, otimizando o retorno sobre investimento. O método proposto envolve o pré-processamento e análise de dados de navegação e histórico de compras de usuários, utilizando o algoritmo de clusterização K-means para identificar grupos de clientes com características semelhantes. Os resultados esperados incluem a identificação de segmentos de consumidores com diferentes níveis de valor e engajamento, permitindo o desenvolvimento de ações personalizadas de marketing e retenção. Conclui-se que a aplicação do K-means representa uma abordagem eficaz para otimizar estratégias de fidelização, melhorar o relacionamento com o cliente e maximizar o retorno sobre o investimento em e-commerce.**

***Palavras-chave*—inteligência artificial não supervisionada, k-means, e-commerce.**

I. Introdução

Segundo a Fundação Getulio Vargas (FGV) [8], a participação das vendas online no faturamento total do varejo praticamente dobrou desde o período pré-pandêmico, passando de 9,2% para 17,8% em fevereiro de 2025, o maior percentual registrado desde junho de 2021. Esse avanço reflete uma mudança significativa no comportamento dos consumidores e na adaptação das empresas ao ambiente digital.

Por meio do crescimento do comércio eletrônico (e-commerce) no Brasil, observa-se que existem grandes volumes de dados comportamentais disponíveis nessas plataformas, embora muitas empresas não os utilizem de forma estratégica. Tais empresas enfrentam dificuldades para identificar segmentos de clientes com alto potencial de valor, prever mudanças no comportamento de compra e aplicar ações de marketing realmente personalizadas. A análise tradicional de dados não consegue acompanhar o volume e a complexidade das informações geradas continuamente, deixando padrões relevantes ocultos, entretanto, a análise de comportamento de clientes é essencial para empresas que desejam aumentar a fidelização, melhorar o retorno das campanhas e alocar recursos de forma mais eficiente.

Diante desse contexto, este trabalho propõe a aplicação do algoritmo K-Means para segmentação de clientes. A aplicação de técnicas de Inteligência Artificial para retenção de clientes permite que empresas de *e-commerce* aumentem a eficiência de suas estratégias de marketing, reduzam custos associados à aquisição de novos clientes e melhorem o relacionamento com usuários recorrentes. Esse impacto reflete-se em benefícios diretos para consumidores, que passam a receber ofertas e recomendações de produtos personalizados, e para o mercado, considerando que cerca de 20% dos clientes geram 80% do faturamento. A identificação desse grupo de alto valor é essencial para maximizar o retorno sobre investimento e promover a sustentabilidade financeira das plataformas de *e-commerce*.

Por fim, o objetivo central deste trabalho consiste em aplicar técnicas de Inteligência Artificial não supervisionada para realizar a segmentação de clientes B2C em um *e-commerce*, identificando padrões ocultos nos dados de compra e extraindo *insights* estratégicos para personalização, retenção e aumento da conversão, por meio do tratamento e análise dos dados utilizando princípios e métodos da Ciência de Dados. De forma mais detalhada, os objetivos específicos incluem: realizar o pré-processamento de dados de clientes e transações de *e-commerce* (histórico de compras, frequência, *ticket* médio, etc.) extraídos de um banco de terceiros para análise e segmentação; aplicar técnicas de aprendizado de máquina não supervisionado para segmentação de clientes e análise de padrões de comportamento de compra, incluindo frequência de aquisição, sazonalidade; gerar *insights* estratégicos que apoiem ações de engajamento e fidelização de clientes; e avaliar a qualidade dos agrupamentos obtidos, mapeando possíveis limitações e desafios da abordagem, como alta dimensionalidade dos dados e necessidade de interpretação dos *clusters*.

II. Trabalhos Relacionados

Diversos estudos têm explorado a aplicação de técnicas de Inteligência Artificial e mineração de dados em diferentes contextos de análise. A seguir, serão apresentados alguns trabalhos que se relacionam com o tema proposto neste estudo.

Silva [3] estruturou seu projeto segundo a metodologia CRISP-DM, apresentando o modelo LRFMVP (*Length, Recency, Frequency, Monetary, Variety, Periodicity*) como uma evolução do modelo LRFM, ao incorporar novas dimensões

comportamentais. Para a segmentação de clientes, as variáveis derivadas do modelo LRFMVP foram utilizadas como entradas do algoritmo de clusterização K-Means. O Método do Cotovelo (*Elbow Method*) foi aplicado para determinar o número ideal de clusters. O autor também aplicou o algoritmo Apriori para identificar regras de associação de compra, obtendo melhoria na confiança média das regras de 93,67% para 94,38% e identificando três clusters de clientes valiosos, sendo um composto por 219 clientes com indicadores acima da média.

Benzi [2] analisou o comportamento de compra em um e-commerce brasileiro utilizando o modelo RFV (*Recência, Frequência e Valor*) para caracterizar clientes. A autora comparou os algoritmos K-Means, K-Medoids e DBSCAN, avaliando a qualidade dos clusters pelo *Silhouette Score*, e aplicou PCA para reduzir a dimensionalidade dos dados. O K-Means apresentou o melhor desempenho (0,56), enquanto o DBSCAN revelou agrupamentos mais coerentes com padrões de pagamento e frequência. A segmentação identificou clientes ocasionais que utilizam boleto e clientes frequentes com meios de pagamento variados.

III. Material e Método

Para fins de análise acadêmica e científica, será utilizada a base de dados disponibilizada no Kaggle[1], referente a uma empresa localizada no Reino Unido, que fornece registros detalhados sobre o comportamento de usuários em ambientes de comércio eletrônico. O conjunto de dados contém 541.909 registros entre 1º de dezembro de 2010 e 9 de dezembro de 2011, abrangendo informações transacionais como identificação do cliente, número da fatura, data da compra, quantidade de itens e preço unitário.

TABLE I
Descrição das colunas do *E-Commerce Data*

Coluna	Tipo de dado	Descrição
InvoiceNo	String/Numérico	Número único da fatura que identifica cada transação. Valores começando com 'C' indicam cancelamentos.
StockCode	String	Código único de cada produto, utilizado para identificação no estoque.
Description	String	Nome ou descrição detalhada do produto.
Quantity	Inteiro	Quantidade de itens adquiridos na transação. Pode ser negativa em casos de devolução.
InvoiceDate	Data/Hora	Data e hora em que a transação foi realizada.
UnitPrice	Float	Preço unitário do produto em libras esterlinas (£).
CustomerID	Inteiro	Identificador único do cliente que realizou a compra.
Country	String	País do cliente, indicando a origem da transação.

O presente estudo busca compreender padrões de comportamento e identificar segmentos de clientes, com o intuito de apoiar estratégias de retenção e engajamento em plataformas de e-commerce.

Todas as etapas de pré-processamento e modelagem foram executadas utilizando bibliotecas consolidadas do ecossistema Python, cada uma desempenhando funções específicas no tratamento e análise dos dados:

- **Pandas**: utilizada para leitura, manipulação e limpeza dos dados transacionais. Sua estrutura tabular permitiu a criação eficiente da matriz RFM, bem como o tratamento de valores ausentes, duplicidades e inconsistências.
- **NumPy**: empregada para operações numéricas de baixo nível e suporte às transformações matemáticas, garantindo maior desempenho em cálculos vetorizados.
- **Scikit-learn** (StandardScaler, KMeans, silhouette_score):
  - **StandardScaler**: responsável pela padronização das variáveis RFM, evitando distorções decorrentes de diferentes escalas entre recência, frequência e valor monetário.
  - **KMeans**: algoritmo utilizado para a clusterização não supervisionada dos clientes.
  - **silhouette_score**: métrica adotada para avaliar a coesão e separação dos clusters, assegurando qualidade e consistência na segmentação obtida.
- **Matplotlib**: responsável pela construção de gráficos de apoio, incluindo o Método do Cotovelo e visualizações das distribuições dos clusters no espaço das variáveis.
- **Seaborn**: utilizada para visualizações exploratórias mais robustas, como *pairplots* em escala logarítmica, facilitando a interpretação visual dos padrões de segmentação.

Para isso, no pré-processamento dos dados, engloba as etapas como limpeza, normalização e transformação, a fim de adequá-los ao uso em técnicas de análise e aprendizado de máquina. Esse processo inclui o tratamento de valores ausentes, a padronização de formatos e a codificação de variáveis categóricas, garantindo a consistência e a qualidade necessárias para a análise.

Com o intuito de enriquecer as variáveis de entrada, foi criada a variável derivada `TotalPrice`, obtida pela multiplicação de *Quantity* e *UnitPrice*, representando o valor total de cada transação. Essa métrica permite capturar o impacto econômico de cada compra, tornando o modelo mais sensível ao comportamento de gasto dos clientes.

Na sequência, aplicou-se a técnica RFM (Recência, Frequência e Valor Monetário) para mensurar o comportamento de compra de cada cliente. Han, Pei e Kamber [9] descrevem o modelo RFM sendo uma das abordagens mais tradicionais para análise de comportamento de clientes, sendo amplamente utilizado em marketing de relacionamento por sua simplicidade e eficácia. Essa técnica avalia três dimensões fundamentais: recência, que representa o tempo decorrido desde a última compra do cliente; frequência, que indica o número de compras realizadas em determinado período; e valor monetário, que expressa o total gasto pelo cliente. As métricas RFM foram calculadas individualmente para cada

[1]Plataforma online voltada à ciência de dados e aprendizado de máquina, oferecendo datasets e ambientes de execução para análises reprodutíveis.

cliente e utilizadas como variáveis de entrada para a etapa de modelagem. Para uniformizar as escalas e evitar distorções entre as variáveis, aplicou-se a padronização por meio do método *StandardScaler*, da biblioteca `scikit-learn`.

A etapa seguinte consistiu na aplicação do algoritmo de clusterização K-Means, um método não supervisionado amplamente utilizado para agrupar dados em subconjuntos homogêneos, com base na similaridade entre observações (MacQueen [10]). O número ótimo de clusters (k) foi determinado pelo Método do Cotovelo (*Elbow Method*), que identifica o ponto de equilíbrio entre o número de grupos formados e a variância explicada pelo modelo. Após a definição de k, o algoritmo K-Means foi aplicado às variáveis RFM padronizadas, resultando na segmentação dos clientes em grupos com comportamentos de compra semelhantes.

Para avaliar a qualidade dos agrupamentos obtidos, utilizou-se o Índice de Silhouette, métrica proposta por Rousseeuw [11], que mede o grau de coesão interna e separação entre os clusters formados. Valores próximos de 1 indicam alta similaridade dentro dos grupos e boa distinção entre eles, validando a consistência dos resultados, e valores negativos sugerem que pontos foram mal alocados. A análise dos clusters foi complementada com visualizações gráficas, incluindo *pairplots* e gráficos de dispersão.

IV. Resultados

Após o tratamento de preparação dos dados a aplicação das métricas RFM permitiu a criação de três variáveis derivadas: Recência, Frequência e Valor Monetário, padronizadas utilizando o método StandardScaler.

Em seguida, aplicou-se o algoritmo de clusterização K-Means. A definição do número de agrupamentos baseou-se no Método do Cotovelo, conforme visualizado na figura 1, observou-se uma redução acentuada da inércia entre $k = 2$ e $k = 3$, seguida de uma queda ainda significativa ao atingir $k = 4$. A partir de quatro clusters, o ganho marginal decorrente da adição de novos grupos torna-se reduzido, indicando um ponto de equilíbrio entre a qualidade da separação e a complexidade do modelo.

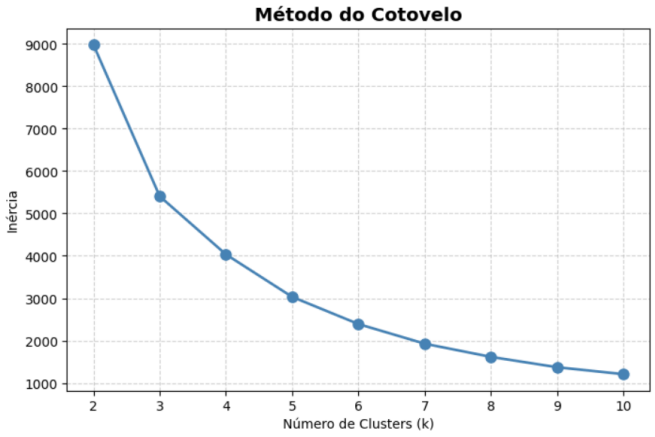

Fig. 1. Método do Cotovelo

Para reforçar essa definição, foi calculado o Índice de Silhouette, que apresentou média de 0,61, valor considerado satisfatório para modelos de clusterização em dados reais, indicando boa separação entre os grupos e coerência interna das estruturas formadas.

A figura 2 apresenta a segmentação dos clientes após a análise RFM com K-Means, utilizando Recência e Valor Monetário para visualizar os resultados. Cada cor representa um cluster distinto, enquanto os símbolos em "X" indicam os centróides, ou seja, a posição média de cada grupo.

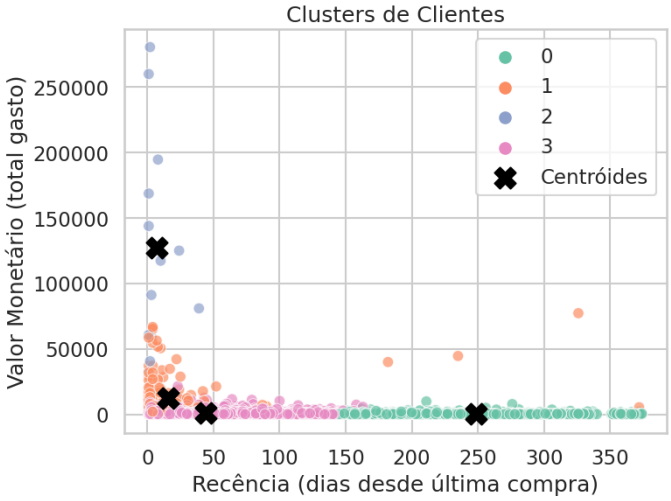

Fig. 2. Clusters de Clientes

Com a definição de $k = 4$, os clientes foram agrupados e as métricas médias de cada cluster foram calculadas, possibilitando a interpretação dos perfis obtidos. A figura 3 apresenta os dados de Recência (Recency), Frequência (Frequency) e Valor Monetário (Monetary) para cada grupo, bem como sua participação proporcional no total de clientes e na receita gerada.

	Clientes	% Clientes	Recency	Freq. Média (ano)	Monetary	Valor Total (£)	% Receita
0	1062	24.48	248.5600	1.60	£476.33	£505,863.04	5.69
1	211	4.86	15.6700	22.00	£12,435.09	£2,623,803.29	29.52
2	13	0.30	7.3800	82.50	£127,187.96	£1,653,443.47	18.60
3	3052	70.36	43.9200	3.70	£1,344.72	£4,104,099.09	46.18
Total	4338	100.00	78.8825	27.45	£2,048.69	£8,887,208.89	100.00

Fig. 3. Estatísticas descritivas dos quatro clusters formados pelo algoritmo K-Means

A análise descritiva dos clusters revela padrões de comportamento bastante distintos entre os clientes (figura 3). Observa-se que o Cluster 3 concentra a maior parte da base (70,36%), caracterizando-se por um perfil de compra moderado, com gasto médio reduzido, porém representando 46,18% da receita total devido ao seu volume absoluto.

Em contraste, o Cluster 2 reúne um grupo altamente seleto de clientes estratégicos (apenas 0,30%), mas com o maior

valor monetário médio (£127.187,96), sendo responsável por 18,60% da receita total. Trata-se de um segmento de alto valor agregado, essencial para manutenção e direcionamento de ações personalizadas.

O Cluster 1 destaca-se como o perfil de clientes fiéis, apresentando alta frequência de compras e ticket médio elevado, contribuindo de forma significativa para o faturamento global (29,52%).

Por outro lado, o Cluster 0 demanda maior atenção gerencial. Apesar de representar 24,48% dos clientes, possui a maior recência média (248 dias), evidenciando indícios de desengajamento e risco elevado de evasão (churn). Esse grupo se mostra prioritário para estratégias de reativação e retenção.

Esses resultados fornecem insumos estratégicos para ações de marketing direcionadas. Clientes do *Cluster 3* tendem a responder melhor a programas de fidelização, enquanto o *Cluster 0* demanda estratégias de reativação. O *Cluster 1*, apesar de apresentar um volume relevante de clientes, possui retorno médio reduzido, nesse caso, recomenda-se a aplicação de estratégias voltadas para o aumento do *ticket* médio, como ações de *cross-selling* e recomendações de produtos complementares às compras já realizadas pelo cliente, ampliando o valor capturado por transação.

Assim, considerando o princípio de Pareto, os Clusters 1 e 2 configuram-se como os segmentos prioritários para alocação de recursos de marketing e ações de relacionamento, por apresentarem maior retorno proporcional sobre o esforço de investimento.

O código-fonte implementado para este estudo está disponibilizado publicamente em repositório GitHub [12], garantindo transparência metodológica e possibilitando a reprodução e extensão dos resultados por outros pesquisadores.

V. Referências

[1] Acconcia, *Elaboração de TCCs*, Medium, 8 ago. 2025. Disponível em: https://acconcia.medium.com/elabora%C3%A7%C3%A3o-de-tccs-6d308b5fc7da. Acesso em: 21 set. 2025.

[2] G. Benzi, *Análise do comportamento de compra dos clientes no e-commerce brasileiro utilizando técnicas de mineração de dados*, Trabalho de Conclusão de Curso (MBA em Ciência de Dados), Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo – USP, São Carlos, 2023. Disponível em: https://sites.icmc.usp.br/apneto/MBA/TCC_MBA_2023_Gabriella.pdf. Acesso em: 21 set. 2025.

[3] J. C. V. da Silva, *Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com Inteligência Artificial*, Dissertação de Mestrado, Escola de Engenharia, Universidade do Minho, Braga, Portugal, jul. 2022. Disponível em: https://repositorium.uminho.pt/server/api/core/bitstreams/41560ca5-b9e4-4bd2-ace8-2d6aec16a34a/content. Acesso em: 26 set. 2025.

[4] L. Igual, S. Seguí, *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, 2ª ed., Springer Nature Switzerland AG, 2024. ISBN 978-3-031-48955-6.

[5] M. M. Maia, S. Fernandes, *K-means na análise de características socioeconômicas de candidatos ao ensino superior*, Econômica, v. 19, n. 1, 2021. Disponível em: https://periodicos.ufersa.edu.br/ecop/article/view/11168/10877. Acesso em: 26 set. 2025.

[6] J. Santos, M. S. Lima, J. P. Costa, *Segmentação via machine learning: proposta de clusterização de consumidores do e-commerce de uma empresa multinacional do varejo esportivo*, Revista HOLOS, v. 9, n. 1, p. 1–15, 2023. Disponível em: https://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/12032/3522. Acesso em: 26 set. 2025.

[7] Y. Peng, J. P. F. Silva, M. H. Nagata, *Um guia rápido sobre os conceitos fundamentais em ciência de dados: Como começar, como fazer certo e com o que tomar cuidado*, ResearchGate, 2025. Acesso em: 26 set. 2025.

[8] FGV IBRE. *Sondagem do Comércio: Indicador de Vendas Online*. Rio de Janeiro: Instituto Brasileiro de Economia (IBRE), Fundação Getulio Vargas, 7 out. 2025. Disponível em: https://portalibre.fgv.br/system/files/2025-10/pressrelease-indicador-de-vendas-online-fgv-ibre.pdf. Acesso em: 10 nov. 2025.

[9] J. Han, J. Pei, M. Kamber. *Data Mining: Concepts and Techniques*. Cambridge, MA: Morgan Kaufmann, 2018.

[10] J. B. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations*. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, p. 281–297. Berkeley, CA: University of California Press, 1967.

[11] P. J. Rousseeuw. *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65. Elsevier, 1987.

[12] FORTOLAN, L. *Aplicação do K-means no Ecommerce*. GitHub, 2025. Disponível em: https://github.com/LucasFortolan/Aplicacao_K-means_E-commerce. Acesso em: 24 nov. 2025.