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. Para isso, será realizado o pré-processamento dos dados, englobando 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 é 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 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. A análise dos clusters foi complementada com visualizações gráficas, incluindo *pairplots* e gráficos de dispersão, os quais facilitaram a interpretação das relações entre as dimensões RFM.

Por fim, os resultados foram sintetizados em uma tabela de resumo, contendo indicadores médios de recência, frequência e valor monetário por cluster, bem como a participação percentual de clientes e de receita em cada grupo. Essa análise possibilitou a identificação de diferentes perfis de clientes, como clientes VIP, clientes regulares e clientes inativos.

[1] A Kaggle é uma plataforma global de ciência de dados, aprendizado de máquina e e análise de dados que oferece conjuntos de dados.

Assim, a metodologia proposta integrou análise quantitativa, estatística descritiva e técnicas de aprendizado de máquina, estruturando-se nas seguintes etapas: coleta e preparação dos dados, cálculo das métricas RFM, padronização das variáveis, aplicação do algoritmo K-Means e avaliação dos resultados. Essa abordagem possibilitou uma segmentação precisa e fundamentada, oferecendo subsídios para a formulação de estratégias de marketing personalizadas e ações de fidelização de clientes no contexto do e-commerce analisado.

IV. Resultados Obtidos

Após o tratamento e 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 para garantir comparabilidade entre as dimensões. Em seguida, foi implementado o algoritmo de clusterização K-Means, com o número de agrupamentos definido por meio do Método do Cotovelo, que indicou quatro clusters como configuração ideal, equilibrando a variância intra e intergrupos.

O Índice de Silhouette apresentou valor médio de 0,61, indicando boa separação entre os clusters e coerência interna entre os perfis identificados. A partir da análise descritiva dos grupos, observou-se que:

- **Cluster 1:** clientes ocasionais, baixa frequência e baixo valor monetário (45% da base);
- **Cluster 2:** clientes regulares, compras recorrentes e valor intermediário (30%);
- **Cluster 3:** clientes de alto valor, alta frequência e grandes gastos (15%);
- **Cluster 4:** clientes inativos, longos intervalos desde a última compra (10%).

A visualização dos clusters em plano bidimensional, após redução de dimensionalidade via PCA, reforçou a distinção entre os grupos e a coerência dos padrões identificados. Os resultados demonstram a eficácia da metodologia na segmentação de clientes e na geração de subsídios estratégicos para retenção, fidelização e campanhas direcionadas.

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.