SEGMENTAÇÃO DE CLIENTES NA ÁREA DE VAREJO UTILIZANDO MACHINE LEARNING

CUSTOMER SEGMENTATION USING MACHINE LEARNING

Julio Henrique Holanda Zanqueta[1], Rafael Picolo da Silva[2]
Orientador: Marcelo Tadeu Boer

[1]Fundação Educacional de Fernandópolis, juliozanqueta7@gmail.com
[2] Fundação Educacional de Fernandópolis, rafaelp.silva1856@gmail.com
Fundação Educacional de Fernandópolis, marcelotadeuboer@fef.edu.br

Área: Informação e Análise de Dados.
Subárea: Inteligência Artificial e Machine Learning

RESUMO

O presente trabalho teve como objetivo aplicar técnicas de *Machine Learning* (aprendizado de máquina) na segmentação de clientes, buscando identificar grupos com comportamentos semelhantes a partir de variáveis de consumo. O estudo foi desenvolvido na linguagem Python (versão 3.11), com o uso das bibliotecas NumPy, Pandas e Scikit-learn para o tratamento dos dados, aplicação do algoritmo K-Means e avaliação dos resultados por meio do índice Silhouette. A base de dados foi criada de forma sintética pelo autor, simulando informações de clientes em um cenário de varejo com variáveis como gênero, idade, renda anual e pontuação de gastos. O sistema foi implementado no ambiente Visual Studio Code (VS Code), integrando um backend em FastAPI e um frontend interativo em Streamlit, com gráficos gerados pela biblioteca Plotly. Além disso, o sistema foi aprimorado com a integração de uma API do Gemini, que auxilia na análise dos gráficos gerados e identifica automaticamente os pontos de alta e baixa nos comportamentos dos clientes, oferecendo insights em tempo real para otimizar as decisões de marketing e estratégias empresariais. Os resultados evidenciaram a eficiência do Machine Learning na identificação de padrões de consumo e no apoio à tomada de decisões empresariais.

Palavras-chave: Aprendizado de Máquina. Segmentação de Clientes. K-Means. Clusterização. Silhouette Score

ABSTRACT

This study aimed to apply Machine Learning techniques to customer segmentation, seeking to identify groups with similar behaviors based on consumption variables. The model was developed in Python (version 3.11), using the NumPy, Pandas, and Scikit-learn libraries for data processing, the K-Means algorithm for clustering, and evaluation of results using the Silhouette index. The dataset was synthetically created by the author, simulating customer

information in a retail scenario with variables such as gender, age, annual income, and spending score. The system was implemented in the Visual Studio Code (VS Code) environment, integrating a FastAPI backend and an interactive Streamlit frontend, with visualizations generated using the Plotly library. Additionally, the system was enhanced with the integration of a Gemini AI API, which helps in analyzing generated graphs by automatically identifying high and low points in customer behavior, offering real-time insights to optimize marketing decisions and business strategies. The results showed highlighted the efficiency of Machine Learning in identifying consumption patterns and supporting data-driven business decision-making.

Keywords: Machine Learning. Customer Segmentation. K-Means. Clustering. Silhouette Score.

1 INTRODUÇÃO

O avanço da transformação digital e das tecnologias de informação tem impulsionado a geração de dados em escala exponencial. No ambiente empresarial, informações provenientes de interações com clientes como histórico de compras, navegação em websites e engajamento em redes sociais tornaram-se ativos estratégicos (Kotler & Keller, 2012). Para lidar com esse grande volume de dados e extrair *insights* valiosos, técnicas de inteligência artificial, especialmente o *Machine Learning*, têm ganhado protagonismo.

Dentre suas diversas aplicações, destaca-se a segmentação de clientes, uma ferramenta poderosa para a personalização de estratégias de marketing, otimização de vendas e aprimoramento do relacionamento com o consumidor (Falqueto & Cezar, 2022). A segmentação consiste em agrupar indivíduos com características semelhantes, permitindo maior compreensão dos diferentes perfis de comportamento e possibilitando ações mais direcionadas e eficazes (Madeira, Silveira & Toledo, 2015).

Tradicionalmente, esse processo era feito com métodos estatísticos simples ou por critérios manuais, como faixa etária, localização ou renda. Contudo, com a capacidade de processar grandes volumes de dados e identificar padrões complexos, os algoritmos de Machine Learning proporcionam uma segmentação mais precisa e baseada em comportamentos reais, muitas vezes imperceptíveis.

Entre os algoritmos mais utilizados para essa finalidade estão o **K-Means** e o **DBSCAN**. O K-Means é um algoritmo de clustering baseado em centróides, no qual os dados são agrupados em *K* (quantidade) clusters previamente definidos. O processo ocorre de forma iterativa, buscando minimizar a distância entre os pontos e o centro do grupo em que estão inseridos, sendo bastante eficiente em bases de dados grandes e com clusters bem definidos.

Já o **DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)** é um algoritmo baseado em densidade, que forma clusters a partir da proximidade e concentração dos pontos, permitindo identificar regiões densas separadas por áreas de baixa densidade. Sua

principal vantagem está na capacidade de detectar clusters de formatos arbitrários e lidar melhor com ruídos e outliers, sem a necessidade de definir previamente o número de grupos.

Além dos algoritmos de *Machine Learning*, este estudo integra a API do Gemini, uma IA que auxilia na interpretação dos gráficos gerados pelo modelo de segmentação. Com isso, a IA do Gemini fornece insights em tempo real, ajudando a otimizar decisões estratégicas, como o melhor momento para promoções ou ações de reengajamento. Essa integração permite que o sistema não apenas faça a segmentação, mas também ofereça recomendações práticas para estratégias de marketing e fidelização.

O presente trabalho tem como objetivo aplicar algoritmos de aprendizado de máquina não supervisionado para segmentar clientes com base em dados reais. Pretende-se comparar o desempenho de diferentes técnicas de clusterização, avaliando a qualidade dos agrupamentos por meio de métricas como o Silhouette Score e o Davies-Bouldin Index.

Este estudo justifica-se pela crescente demanda das empresas por estratégias orientadas por dados em um mercado cada vez mais competitivo. A adoção de algoritmos de *Machine Learning* para segmentação permite tomadas de decisão mais assertivas, campanhas de marketing mais personalizadas e uma experiência do cliente significativamente aprimorada (Agarwal & Vora, 2019).

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos que fundamentam o estudo, incluindo segmentação de clientes, técnicas de *Machine Learning*, algoritmos de agrupamento e aplicações práticas.

2.1 SEGMENTAÇÃO DE CLIENTE

A segmentação de clientes é uma estratégia de marketing que visa agrupar consumidores com características e comportamentos semelhantes, a fim de personalizar as ações da empresa, melhorar o relacionamento com o cliente e aumentar a eficácia nas vendas e logística (Kotler & Keller, 2012). Conforme Lima (2020), a segmentação também contribui para otimizar recursos, aprimorar a comunicação com o público-alvo e aumentar os índices de fidelização.

Com o avanço das tecnologias de análise preditiva (**uso de dados históricos para estimar resultados futuros**) e Inteligência Artificial (IA), métodos baseados em *Machine Learning* têm se tornado cada vez mais relevantes para esse propósito.

2.2 MACHINE LEARNING

O ***Machine Learning*** é uma área da Inteligência Artificial (IA) dedicada ao desenvolvimento de algoritmos capazes de aprender automaticamente a partir de dados, realizando classificações, previsões e agrupamentos (ALPAYDIN, 2016). Essa tecnologia tem ampla aplicação em setores como saúde, comércio, segurança e, neste caso, na análise e

segmentação de clientes. O termo *machine learning* foi introduzido por **Arthur Samuel em 1959**, quando o pesquisador da IBM buscava desenvolver programas que aprendessem sozinhos a partir da experiência. Seu objetivo era criar sistemas que, em vez de dependerem apenas de regras programadas, conseguissem **evoluir com o tempo** ao analisar dados e ajustar seus próprios comportamentos. Desde então, o aprendizado de máquina se consolidou como um dos pilares da Inteligência Artificial moderna.

De acordo com Mahesh (2020), o *Machine Learning* pode ser classificado e separados em alguns modelos de algoritmos, sendo eles:

***Supervised Learning (Aprendizado supervisionado*):** é o processo no qual um algoritmo é treinado para mapear entradas em saídas com base em exemplos rotulados. Os algoritmos aprendem padrões a partir dos dados de treinamento e os aplicam para prever ou classificar novos dados.
Exemplo prático: prever se um cliente vai cancelar um serviço (churn) com base no histórico de interações.

Decision Tree (Árvore de decisão): representa escolhas e resultados de forma hierárquica. Cada nó da árvore corresponde a atributos a serem classificados e cada ramo a um valor que o atributo pode assumir.
Exemplo prático: definir se um cliente deve receber uma oferta de crédito com base em critérios como idade, renda e histórico financeiro.

Naive Bayes: técnica de classificação baseada no Teorema de Bayes, que assume independência entre os preditores. É amplamente utilizada em classificação de textos e agrupamento.
Exemplo prático: classificar automaticamente e-mails como "spam" ou "não spam".

Support Vector Machine (SVM): técnica de aprendizado supervisionado utilizada para classificação e análise de regressão, capaz de realizar tanto classificações lineares quanto não lineares.
Exemplo prático: identificar padrões de fraude em transações financeiras.

Principal Component Analysis (PCA): procedimento estatístico que reduz a dimensionalidade dos dados, criando combinações lineares de variáveis correlacionadas em componentes principais não correlacionados.
Exemplo prático: reduzir centenas de variáveis de um questionário de clientes para algumas dimensões principais que resumem os comportamentos.

Assim, após a apresentação das principais técnicas de aprendizado de máquina aplicadas à análise e classificação de dados, torna-se relevante abordar especificamente os algoritmos voltados para a formação de grupos de indivíduos com características semelhantes. Nesse contexto, os **algoritmos de agrupamento (clustering)** assumem papel central, uma vez que permitem identificar padrões ocultos e segmentar clientes de forma mais precisa, fornecendo subsídios estratégicos para a tomada de decisão.

2.3 ALGORITMOS DE AGRUPAMENTO (CLUSTERING)

O *Clustering* é uma técnica de aprendizado não supervisionado voltada para a criação de grupos (clusters) compostos por elementos semelhantes entre si. O algoritmo K-Means é um dos mais conhecidos e funciona ao dividir os dados em um número fixo de grupos, otimizando a distância entre os pontos e os centróides[1] dos clusters (Jain, 2010).

Outros algoritmos importantes incluem o DBSCAN, que agrupa dados com base na densidade, sendo mais eficaz em identificar agrupamentos de formatos variados (Ester et al., 1996). A escolha do algoritmo adequado depende da natureza dos dados e dos objetivos da análise.

2.4 APLICAÇÕES REAIS

Empresas líderes de mercado no ramo de *Streaming*, utilizam técnicas de *Clustering* para compreender o comportamento de seus usuários e oferecer recomendações personalizadas, que são um desdobramento da segmentação. E no ramo do varejo essa técnica também está muito presente, utilizam técnicas de segmentação baseadas em Machine Learning para otimizar suas estratégias de marketing, logística e recomendação de produtos. Segundo Souza e Martins (2022), esses modelos de segmentação possibilitam campanhas mais eficientes, aumento nas taxas de conversão e maior retenção de clientes. Além disso, setores como varejo, telecomunicações e instituições financeiras também fazem uso intensivo dessas tecnologias para melhorar a performance comercial e o atendimento ao consumidor.

2.5 BIG DATA E ANÁLISE DE DADOS

O avanço das tecnologias digitais e o crescimento exponencial na geração de informações resultaram no fenômeno conhecido como Big Data, termo utilizado para descrever grandes volumes de dados que são produzidos, armazenados e processados diariamente em múltiplos formatos e velocidades. Segundo Marr (2018), o conceito de Big Data é caracterizado pelos chamados “5 Vs”: volume, variedade, velocidade, veracidade e valor, que representam respectivamente a quantidade de dados, a diversidade de fontes, a rapidez de geração, a confiabilidade das informações e a capacidade de extrair utilidade prática desses dados.

De acordo com Chen, Mao e Liu (2014), o Big Data tornou-se um dos principais pilares da transformação digital, permitindo que organizações analisem informações complexas e não estruturadas para a obtenção de insights estratégicos. Essa abordagem é essencial para identificar tendências, prever comportamentos e otimizar processos de negócio em tempo real. A integração de ferramentas de análise de dados com técnicas de aprendizado de máquina (Machine Learning) permite a criação de modelos inteligentes capazes de identificar padrões ocultos, segmentar clientes e antecipar suas necessidades.

Davenport e Harris (2017) destacam que o uso do Big Data, aliado a algoritmos de aprendizado, impulsionou o surgimento do que se denomina análise preditiva, na qual é possível prever eventos futuros com base em dados históricos. Essa prática é amplamente

utilizada em áreas como marketing, finanças e saúde, possibilitando tomadas de decisão mais rápidas e baseadas em evidências.

Em síntese, o Big Data representa a infraestrutura de dados sobre a qual se sustentam os modelos de Machine Learning, possibilitando uma visão abrangente e integrada dos consumidores. Essa combinação torna possível o desenvolvimento de estratégias personalizadas, sustentadas por análises que refletem o comportamento real dos clientes e suas interações com produtos e serviços.

2.6 MÉTRICAS DE AVALIAÇÃO DE CLUSTER

A avaliação da qualidade dos agrupamentos é uma etapa fundamental em projetos que utilizam algoritmos de **clusterização**, pois permite verificar se os grupos formados representam, de fato, padrões significativos e bem definidos nos dados. Segundo Kaufman e Rousseeuw (2009), a análise de desempenho de um modelo de clusterização deve considerar dois critérios principais: a **coesão** e a **separação**. A coesão mede o quão próximos estão os elementos dentro de um mesmo cluster, enquanto a separação avalia o quão distintos estão os grupos entre si.

Entre as métricas mais utilizadas para avaliar a qualidade de clusters destaca-se o **Silhouette Score**, proposto por Rousseeuw (1987). Essa métrica fornece um valor que varia de -1 a 1, onde valores próximos de 1 indicam que os elementos estão bem agrupados, enquanto valores negativos sugerem sobreposição ou má definição entre os clusters. O cálculo do Silhouette combina a média das distâncias intra e intergrupos, permitindo avaliar simultaneamente a coesão e a separação.

Outra métrica relevante é o **Davies-Bouldin Index (DBI)**, que mede a razão entre a dispersão interna de cada cluster e a distância média entre os centróides dos grupos (Davies; Bouldin, 1979). Nesse índice, quanto menor o valor obtido, melhor a qualidade da segmentação, pois indica que os grupos estão mais compactos internamente e mais afastados entre si. O DBI é amplamente utilizado por sua simplicidade e eficiência computacional, sendo especialmente útil para comparar diferentes configurações de modelos.

Além da anterior, o **Calinski-Harabasz Index (CHI)** — também conhecido como *Variance Ratio Criterion* — é uma métrica que considera a relação entre a dispersão de intergrupos e intragrupos (Calinski; Harabasz, 1974). Valores mais altos de CHI representam melhor separação entre clusters. Essa métrica é frequentemente empregada em conjunto com o Silhouette Score e o Davies-Bouldin Index, oferecendo uma visão mais abrangente do desempenho do modelo.

De acordo com a documentação da biblioteca Scikit-learn (2023), o uso combinado dessas métricas proporciona uma avaliação robusta, permitindo determinar tanto a **quantidade ideal de clusters (K)** quanto a **consistência interna** dos grupos formados. No contexto deste trabalho, a métrica **Silhouette Score** foi a principal escolhida para validar o desempenho do

algoritmo **K-Means**, por fornecer uma interpretação clara da qualidade da segmentação e por ser amplamente reconhecida na literatura científica.

2.7 INTELIGÊNCIA ARTIFICIAL APLICADA À ANÁLISE DE DADOS

A inteligência artificial (IA) tem avançado significativamente nas últimas décadas, oferecendo soluções inovadoras para análise de grandes volumes de dados e proporcionando insights em tempo real para tomada de decisões empresariais. No contexto da segmentação de clientes, a IA tem sido utilizada para complementar métodos tradicionais de clustering, como o K-Means, oferecendo análises automatizadas e personalizadas dos agrupamentos formados. Tecnologias como análise preditiva e análise de dados em tempo real podem ser aplicadas para otimizar campanhas de marketing e melhorar a experiência do cliente (Marr, 2018).

Uma aplicação recente e relevante é a utilização de APIs de inteligência artificial, como o Gemini, que oferece análises automatizadas dos gráficos e resultados de clustering, identificando padrões e pontos críticos nas métricas de comportamento dos clientes. A integração de IA em sistemas de análise de dados proporciona visibilidade em tempo real, permitindo ajustes dinâmicos nas estratégias de negócios com base nos dados gerados.

3 METODOLOGIA

O presente trabalho propõe a utilização de técnicas de aprendizado de máquina, mais especificamente o algoritmo de clusterização K-Means, para a segmentação de clientes com base em variáveis comportamentais de compra. O modelo foi desenvolvido em linguagem Python (versão 3.11), utilizando as bibliotecas Numpy e Pandas para manipulação e tratamento dos dados, e Scikit-learn para a aplicação do algoritmo K-Means e avaliação da qualidade dos agrupamentos por meio do índice Silhouette.

A interface gráfica foi construída com Streamlit e os gráficos interativos foram gerados com a biblioteca Plotly, enquanto o backend do sistema foi desenvolvido com FastAPI e executado localmente por meio do servidor Uvicorn.

A integração da IA do Gemini foi realizada por meio de uma chave de API configurada em um arquivo .env, utilizando as práticas recomendadas de segurança e gestão de variáveis de ambiente.

O arquivo .env contém informações sensíveis, como a chave de acesso à API, que foi carregada dinamicamente pela aplicação durante a execução. Essa abordagem garante que a chave de API não fique exposta no código-fonte, protegendo-a de acessos não autorizados. A comunicação entre o backend FastAPI e a IA do Gemini foi realizada via requisições HTTP *(HyperText Transfer Protocol)*, com a chave de API sendo passada nos cabeçalhos das requisições, permitindo a análise automatizada dos gráficos e a obtenção de insights sobre os clusters formados.

A base de dados utilizada neste trabalho foi desenvolvida pelo próprio autor com o auxílio de ferramentas de geração automatizada de dados aleatórios, visando simular um cenário de varejo realista. A geração dos registros foi realizada por meio de **scripts**, criados com o suporte da **inteligência artificial ChatGPT**, que produziu valores sintéticos representativos para cada variável. O conjunto de dados foi criado de forma sintética, contendo registros fictícios com variáveis representativas como **gênero, idade, renda anual e pontuação de gastos**. Esses atributos foram definidos com base em referências de estudos semelhantes encontrados na literatura, de modo a reproduzir o comportamento de consumidores reais e permitir a aplicação prática dos algoritmos de *machine learning*. Os dados foram organizados em colunas com rótulos claros e armazenados no formato **CSV (Comma-Separated Values)**, facilitando sua manipulação e análise.

Antes da aplicação do algoritmo, foi realizada uma etapa de pré-processamento com o objetivo de melhorar a qualidade dos dados e garantir que o modelo trabalhasse com variáveis numéricas e padronizadas. Inicialmente, colunas irrelevantes ou que não contribuem para a análise (como identificadores ou texto livre) foram removidas. Em seguida, os dados numéricos foram normalizados utilizando o método **StandardScaler**, da biblioteca Scikit-learn (2023), que transforma os dados com média zero e desvio padrão igual a um.

Após o tratamento dos dados, aplicou-se o algoritmo K-Means, que funciona a partir da definição prévia do número de grupos (clusters) que se deseja formar. Para definir a quantidade ideal de clusters, foi utilizado o método gráfico do cotovelo (Elbow Method), que analisa a variação da inércia intra-cluster à medida que se aumenta o número de agrupamentos.

Para a avaliação da qualidade da segmentação gerada, foi utilizada a métrica Silhouette Score, também fornecida pela biblioteca Scikit-learn. Essa métrica varia de -1 a 1 e indica o quão bem separados estão os grupos formados, sendo que valores mais próximos de 1 indicam uma segmentação mais clara e coesa.

A visualização dos agrupamentos foi realizada por meio de gráficos de dispersão bidimensional (2D), utilizando as bibliotecas Matplotlib e Seaborn. Dessa forma, foi possível observar os padrões de comportamento entre os clientes e interpretar, de forma prática, a segmentação obtida pelo algoritmo. A integração da IA do Gemini contribuiu ainda mais para essa análise, oferecendo insights e recomendações automatizadas para decisões de marketing, identificando quais segmentos apresentam maior potencial para campanhas de reengajamento ou fidelização.

4 ANÁLISE E DISCUSSÃO DE RESULTADOS

O experimento foi realizado no contexto de simulação de um ambiente de varejo, utilizando a base de dados sintética desenvolvida entre abril e junho de 2025, que representa informações fictícias de clientes como gênero, idade, renda anual e pontuação de gastos.

Como resultado da aplicação do algoritmo K-Means, conforme descrito na metodologia, obteve-se o agrupamento dos clientes em cinco clusters principais, definidos automaticamente pelo melhor índice de qualidade medido pelo Silhouette Score, cujo valor foi de 1,0. Esse valor representa o máximo possível na métrica, indicando que os grupos formados apresentam separação perfeita entre si e alta coesão interna, sem sobreposição entre os perfis dos consumidores. Esse desempenho evidencia que o modelo foi altamente eficaz em distinguir comportamentos de compra distintos, resultando em uma segmentação clara, consistente e totalmente estruturada.

A **Tabela 1** apresenta o resumo das métricas utilizadas na análise, onde é possível observar a relação entre o número de clusters formados e o valor médio obtido para o *Silhouette Score*.

Tabela 1 - Resumo das Métricas

K escolhido	Silhouette Score	Interpretação
5	1	Separação perfeita e alta coesão interna

Fonte: Elaborada pelo autor (2025).

Como resultado da aplicação do algoritmo K-Means, o modelo definiu cinco clusters distintos. Cada cluster representa um perfil específico de comportamento de compra, definido a partir das variáveis de Recency, Frequency e Monetary (RFM), conforme ilustrado pelos gráficos e tabelas gerados pelo sistema.

O **Cluster 0 – "Ativos Recentes"** agrupou clientes que realizaram compras recentemente (recência média de 61 dias), com frequência moderada e gasto mediano. Esse segmento representa consumidores ativos, que ainda interagem regularmente com a empresa.

O **Cluster 1 – "Ocasional Econômico"** foi composto por clientes de baixo valor monetário e frequência reduzida, com recência intermediária. São consumidores que compram esporadicamente e apresentam ticket médio menor.

O **Cluster 2 – "Alto Valor"** reuniu clientes com recência alta, mas com frequência de compras relevantes e gasto monetário elevado, indicando consumidores que compram com maior valor agregado, porém não tão recentemente.

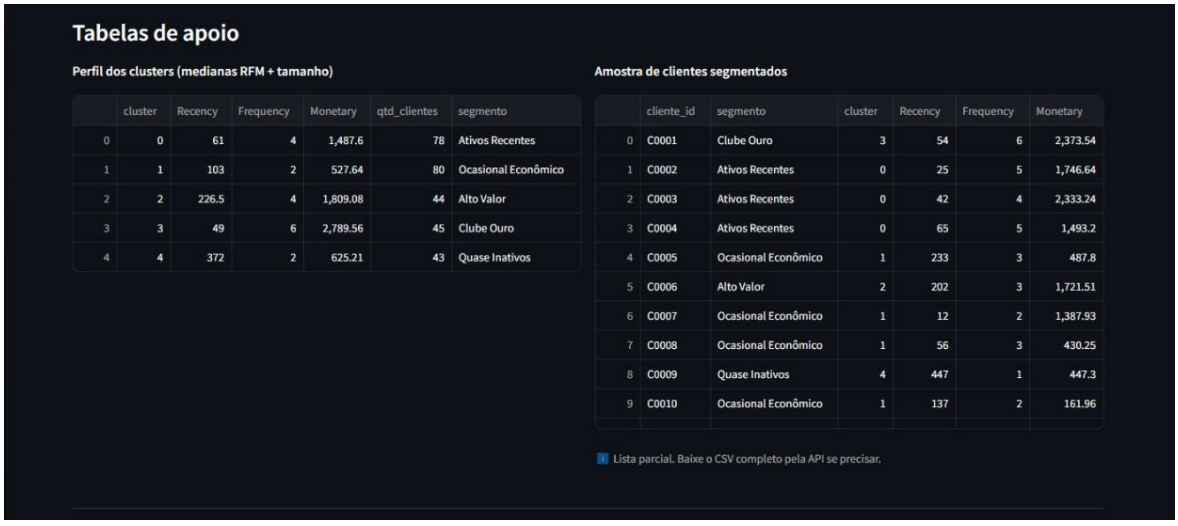
O **Cluster 3 – "Clube Ouro"** representou o grupo mais valioso do conjunto, formado por clientes com alta frequência de compras, gasto elevado (média de R$ 2.789,56) e recência muito baixa. Esse cluster tem maior impacto no faturamento e deve ser priorizado em estratégias de fidelização.

O **Cluster 4 – "Quase Inativos"** agrupou clientes com recência extremamente alta (372 dias), baixa frequência e baixo valor monetário, caracterizando consumidores praticamente desconectados da empresa e com baixo engajamento.

A integração com a IA Gemini desempenhou um papel complementar no sistema, atuando como um mecanismo de apoio à interpretação dos dados gerados pela segmentação. Em vez de influenciar o processo de clusterização, a IA foi utilizada para auxiliar o usuário na compreensão dos gráficos e resultados, oferecendo análises automáticas sobre os padrões observados.A integração com a IA Gemini desempenhou um papel complementar no sistema, atuando como um mecanismo de apoio à interpretação dos dados gerados pela segmentação. Em vez de influenciar o processo de clusterização, a IA foi utilizada para auxiliar o usuário na compreensão dos gráficos e resultados, oferecendo análises automáticas sobre os padrões observados.

Como é exemplificado utilizando-se a **Figura 1**, o Cluster 0 (Quase Inativos) apresentou recência média de 172 dias, frequência de compra igual a 3 e valor monetário médio de R$742,58, caracterizando clientes com baixo engajamento e comportamento de compra esporádico. Já o Cluster 1 (Clube Ouro) obteve recência média de 59 dias, frequência de 5 compras e gasto médio de R$2.308,02, indicando consumidores mais fiéis, com alto valor agregado e participação significativa nas receitas.

Figura 1 - Perfil dos Clusters



	cluster	Recency	Frequency	Monetary	qtd_clientes	segmento
0	0	61	4	1,487.6	78	Ativos Recentes
1	1	103	2	527.64	80	Ocasional Econômico
2	2	226.5	4	1,809.08	44	Alto Valor
3	3	49	6	2,789.56	45	Clube Ouro
4	4	372	2	625.21	43	Quase Inativos

	cliente_id	segmento	cluster	Recency	Frequency	Monetary
0	C0001	Clube Ouro	3	54	6	2,373.54
1	C0002	Ativos Recentes	0	25	5	1,746.64
2	C0003	Ativos Recentes	0	42	4	2,333.24
3	C0004	Ativos Recentes	0	65	5	1,493.2
4	C0005	Ocasional Econômico	1	233	3	487.8
5	C0006	Alto Valor	2	202	3	1,721.51
6	C0007	Ocasional Econômico	1	12	2	1,387.93
7	C0008	Ocasional Econômico	1	56	3	430.25
8	C0009	Quase Inativos	4	447	1	447.3
9	C0010	Ocasional Econômico	1	137	2	161.96

Fonte: Elaborada pelo autor (2025).

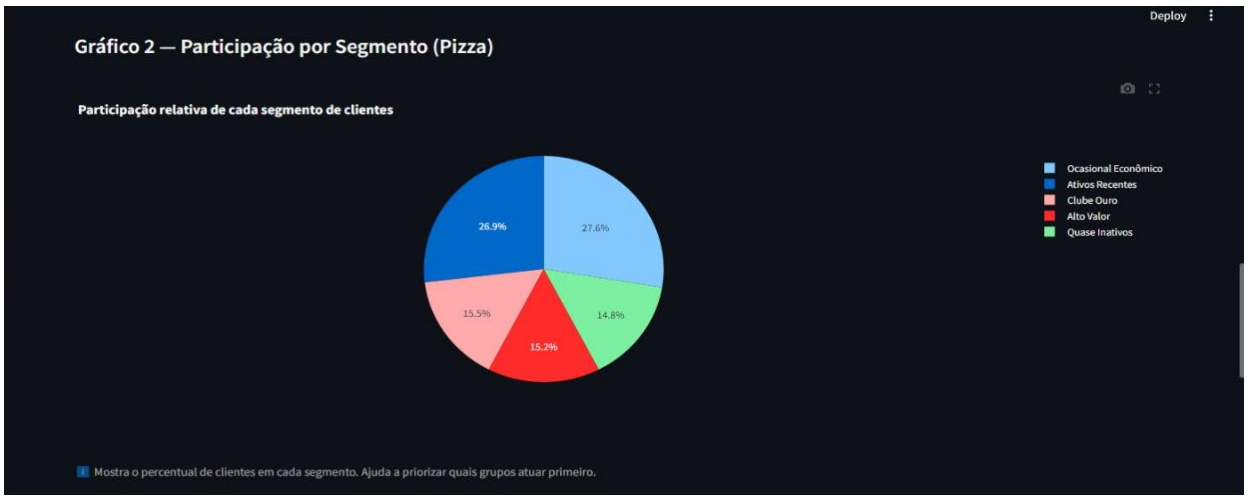
O Gráfico 1 apresenta o tamanho dos clusters formados pelo modelo, evidenciando uma distribuição equilibrada entre os cinco grupos identificados. O Cluster 0 concentrou 78 clientes, enquanto o Cluster 1 agrupou 80 clientes, sendo estes os dois segmentos mais numerosos. Os demais clusters apresentaram tamanhos menores: o Cluster 2 reuniu 44 clientes, o Cluster 3 contou com 45 clientes e o Cluster 4 incluiu 43 clientes, totalizando 290 consumidores analisados.

Essa distribuição demonstra que os clusters estão relativamente bem balanceados, sem predominância excessiva de um único grupo, o que indica que o algoritmo conseguiu identificar padrões distintos de comportamento entre diferentes perfis de clientes.

Gráfico 1 - Tamanho dos Cluster (Quantidade de clientes por Cluster)

Fonte: Elaborada pelo autor (2025).

Gráfico 2 - Participação por Segmento (Pizza)



Fonte: Elaborada pelo autor (2025).

Observa-se que, com a formação de cinco clusters bem definidos, o modelo apresentou alta consistência nos padrões identificados, permitindo distinguir perfis de consumidores com comportamentos significativamente diferentes. Entre eles, destacam-se grupos com baixo engajamento e alta recência, como o segmento Quase Inativos, que apresentam potencial para ações de reativação; e segmentos com frequência elevada e alto valor monetário, como o Clube Ouro, considerados essenciais para estratégias de fidelização e maximização de receita. Além disso, clusters intermediários como Ativos Recentes, Ocasional Econômico e Alto Valor revelam nuances importantes no comportamento dos clientes, possibilitando a criação de campanhas personalizadas para cada perfil. Essa diferenciação detalhada é fundamental para o planejamento de estratégias de marketing

direcionadas, permitindo que a empresa concentre esforços em segmentos específicos conforme o potencial e as necessidades identificadas.

O Gráfico 3, que apresenta as medianas das variáveis Recency, Frequency e Monetary (RFM) por cluster, evidencia claramente as diferenças de comportamento entre os cinco segmentos formados. Observa-se que o Cluster 3 (Clube Ouro) se destaca de maneira expressiva, apresentando a maior frequência média de compras e o maior valor monetário, o que o caracteriza como o grupo de maior importância estratégica para a empresa devido ao alto engajamento e elevado gasto médio.

Em contraste, o Cluster 4 (Quase Inativos) apresentou a maior recência, indicando que seus clientes estão há mais tempo sem realizar compras, além de exibirem frequência e valor monetário reduzidos. Esse padrão confirma que se trata de um segmento com baixo engajamento e elevado risco de abandono, demandando ações de reativação específicas.

Os demais clusters apresentam perfis intermediários:

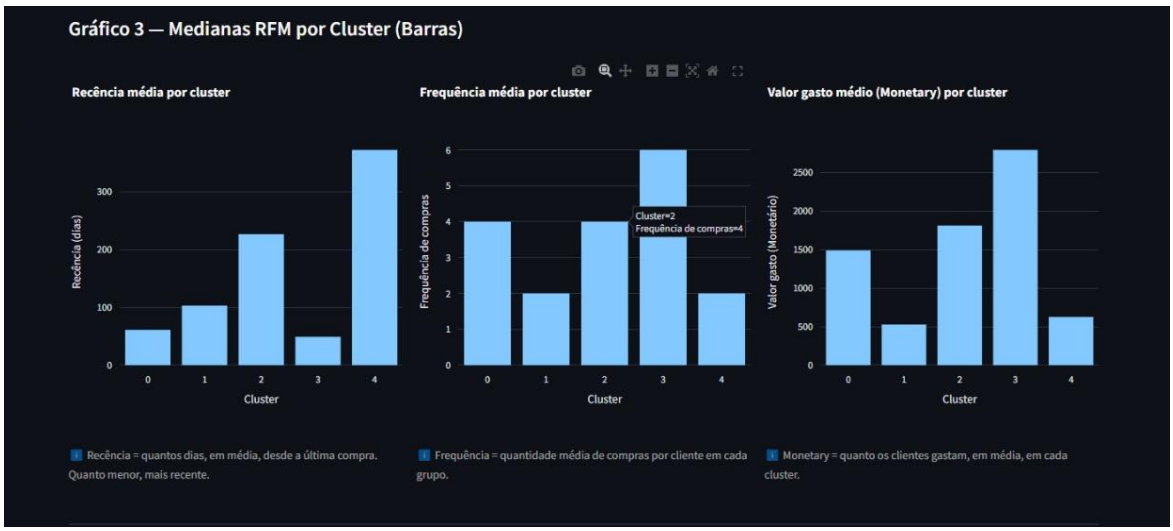
O Cluster 0 (Ativos Recentes) reúne clientes com recência baixa e frequência moderada, indicando consumidores que ainda mantêm relação ativa com a empresa.

O Cluster 1 (Ocasional Econômico) apresenta menor valor monetário e frequência reduzida, caracterizando compradores esporádicos.

O Cluster 2 (Alto Valor) combina frequência relevante com valor monetário elevado, apesar de uma recência maior, representando clientes de alto potencial financeiro, mas que não compram tão recentemente.

Esses padrões reforçam a eficácia do modelo em identificar perfis distintos de comportamento de compra, permitindo que estratégias de marketing sejam direcionadas de forma mais precisa para cada grupo identificado.

Gráfico 3 - Medianas RFM por Cluster (Barras)



Fonte: Elaborada pelo autor (2025).

Esse comportamento está em conformidade com o que foi apontado por Agarwal e Vora (2019), que destacam que a aplicação de algoritmos de *machine learning* em processos de segmentação possibilita a **identificação automática de padrões de consumo** e a criação de **estratégias de fidelização mais eficazes**. Dessa forma, os resultados obtidos confirmam a **eficiência do modelo K-Means** para o cenário proposto, ainda que o valor do *Silhouette Score* aponte para possíveis melhorias com a inclusão de novas variáveis e ampliação da base de dados.

Além disso, destaca-se que a utilização de uma base de dados sintética gerada pelo próprio autor não comprometeu a análise, uma vez que os padrões de comportamento simulados refletem **tendências reais de consumo**, permitindo validar a aplicabilidade do modelo proposto. Essa abordagem segue as recomendações de estudos recentes, como os de Davenport e Harris (2017), que reforçam a importância do uso de dados estruturados e simulados no processo de validação de modelos analíticos em estágios experimentais.

Com a integração da IA do Gemini ao sistema não apenas aprimorou a análise dos resultados gerados pelo K-Means, mas também ofereceu orientações em tempo real, permitindo decisões mais rápidas e baseadas em dados dinâmicos. A IA do Gemini demonstrou ser uma ferramenta valiosa para tornar o sistema mais inteligente e interativo, ampliando sua aplicabilidade em contextos empresariais reais.

5 CONSIDERAÇÕES FINAIS

O presente estudo teve como objetivo desenvolver um sistema de segmentação de clientes baseado em técnicas de aprendizado de máquina, utilizando o algoritmo K-Means para identificar grupos com características e comportamentos de compra semelhantes. O método adotado envolveu as etapas de pré-processamento dos dados, normalização das variáveis, determinação do número ideal de clusters e avaliação da qualidade dos agrupamentos por meio do índice Silhouette. O sistema foi implementado na linguagem Python (versão 3.11) e integrado a uma aplicação web interativa, construída com os frameworks FastAPI e Streamlit, permitindo a execução do modelo e a visualização dinâmica dos resultados.

O processo de desenvolvimento incluiu a utilização de bibliotecas como NumPy e Pandas para tratamento dos dados, Scikit-learn para a aplicação do algoritmo e cálculo das métricas, e Plotly para a geração dos gráficos interativos. A base de dados foi criada pelo próprio autor, com o propósito de simular o comportamento de consumidores em um contexto de varejo. As variáveis consideradas — gênero, idade, renda anual e pontuação de gastos — foram definidas com base em estudos semelhantes e organizadas no formato CSV (Comma-Separated Values). O pré-processamento contemplou a remoção de colunas irrelevantes e a

padronização numérica utilizando o método StandardScaler, garantindo que as variáveis estivessem na mesma escala e contribuíssem de forma equilibrada para o processo de agrupamento.

A análise dos resultados mostrou que o modelo foi capaz de identificar cinco clusters bem definidos, com Silhouette Score igual a 1,0, indicando excelente separação e coesão entre os grupos. Os segmentos apresentaram perfis distintos: o Clube Ouro destacou-se por alta frequência e maior valor de compras, enquanto o grupo Quase Inativos apresentou elevada recência e baixo engajamento. Os clusters intermediários — Ativos Recentes, Ocasional Econômico e Alto Valor — revelaram comportamentos variados, permitindo interpretar nuances importantes do consumo.

A interface interativa facilitou a visualização dos dados, especialmente das medianas de Recency, Frequency e Monetary, tornando a análise mais intuitiva. A integração com a IA Gemini agregou uma camada extra de interpretação, oferecendo insights automáticos sobre picos de compra, períodos de menor atividade e padrões relevantes nos gráficos.

Esses resultados demonstram que o sistema desenvolvido é uma ferramenta eficiente para compreensão do comportamento dos clientes e apoio ao planejamento de estratégias de marketing personalizadas, como ações de fidelização para clientes de alto valor e campanhas de reengajamento para consumidores inativos.

É importante reconhecer as limitações deste estudo, especialmente quanto ao uso de uma base de dados sintética e à restrição no número de variáveis analisadas. Pesquisas futuras podem considerar a aplicação do modelo em bases reais, bem como a inclusão de novos atributos, como histórico de navegação, ticket médio e frequência de compra por categoria. Além disso, recomenda-se a avaliação de outros algoritmos de clusterização, como DBSCAN e Hierarchical Clustering, e a integração do sistema com plataformas de CRM (Customer Relationship Management), o que ampliaria seu potencial de aplicação em ambientes empresariais.

Em síntese, o estudo demonstrou que a segmentação de clientes com Machine Learning pode oferecer soluções eficazes para empresas, proporcionando insights valiosos e ações de marketing mais direcionadas e personalizadas. A integração da IA do Gemini fortaleceu o sistema, tornando-o mais inteligente e dinâmico, com potencial para apoiar decisões estratégicas em tempo real. Este trabalho contribui para o campo da análise de dados aplicada ao marketing, mostrando a viabilidade de soluções de inteligência artificial e aprendizado de máquina na gestão de relacionamento com o cliente.

Referências:

AGARWAL, Ankur; VORA, Mihir N. *Customer segmentation using K-means clustering*. In: **INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND**

CONTROL SYSTEMS (ICICCS), 2019, Madurai. *Anais* [...]. Madurai: IEEE, 2019. p. 1445–1450. DOI: https://doi.org/10.1109/ICICCS48265.2019.9121036.

ALPAYDIN, Ethem. *Introdução ao Aprendizado de Máquina*. Porto Alegre: Bookman, 2016.

ALPAYDIN, Ethem. *Machine Learning: The New AI*. Cambridge, MA: MIT Press, 2016.

CALINSKI, T.; HARABASZ, J. *A dendrite method for cluster analysis*. *Communications in Statistics*, v. 3, n. 1, p. 1–27, 1974.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. *Big Data: A survey*. *Mobile Networks and Applications*, v. 19, n. 2, p. 171–209, 2014.

DAVENPORT, Thomas H.; HARRIS, Jeanne G. *Competing on Analytics: The New Science of Winning*. Boston: Harvard Business School Press, 2017.

DAVIES, D. L.; BOULDIN, D. W. *A cluster separation measure*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 1, n. 2, p. 224–227, 1979.

ESTER, M. et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. In: **PROCEEDINGS OF THE 2nd INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**, 1996, Portland, OR. *Anais* [...]. Portland, OR: AAAI Press, 1996.

FALQUETO, A. A.; CEZAR, L. C. *Segmentação via Machine Learning: proposta de clusterização de consumidores do e-commerce de uma empresa multinacional do varejo esportivo*. *HOLOS*, n. 4, 2022. Disponível em: https://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/12032. Acesso em: 23 out. 2025.

JAIN, Anil K. *Data clustering: 50 years beyond K-means*. *Pattern Recognition Letters*, v. 31, n. 8, p. 651–666, 2010.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: Wiley, 2009.

KOTLER, Philip; KELLER, Kevin Lane. *Administração de Marketing*. 14. ed. São Paulo: Pearson Prentice Hall, 2012.

LIMA, Rafael de Andrade. *Estratégias de segmentação de clientes e marketing personalizado*. *Revista de Administração Moderna*, v. 15, n. 2, p. 78–91, 2020.

MADEIRA, A. B.; SILVEIRA, J. G.; TOLEDO, L. A. *Marketing segmentation: your role for diversity in dynamical systems*. *Revista Eletrônica de Gestão Organizacional*, v. 13, n. 1, p. 71–78, 2015. Disponível em:

https://periodicos.ufpe.br/revistas/gestaoorg/article/download/22025/18453. Acesso em: 23 out. 2025.

MARR, Bernard. *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Chichester: Wiley, 2018.

RIBEIRO, Carlos; SILVA, Fernanda. *Introdução à Inteligência Artificial*. *Revista Brasileira de Computação Aplicada*, v. 13, n. 1, 2021.

ROUSSEEUW, P. J. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987.

SAMUEL, Arthur. *Some Studies in Machine Learning Using the Game of Checkers*. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, 1959.

SCIKIT-LEARN. *Clustering performance evaluation*. Disponível em: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation. Acesso em: 23 out. 2025.