**Análise de satisfação e segmentação de clientes em comércio eletrônico utilizando Aprendizado de Máquina.**

**Satisfaction analysis and customer segmentation in e-commerce using Machine Learning.**

Leite, Franciele L.

[francieledelimaleite@gmail.com](mailto:francieledelimaleite@gmail.com)

Centro Universitário Facens - Sorocaba, SP, Brasil

Submetido em: 10/03/2023. Aceito em: --

# RESUMO

Para empresas focadas no cliente, a satisfação é ao mesmo tempo uma meta e uma ferramenta de *marketing*, onde o seu intuito é ajudar a atingir seus objetivos. O *marketing* está presente em todos os âmbitos de uma organização porque ajuda a conduzir a tomada de decisões. O propósito deste trabalho é desenvolver uma estratégia de segmentação baseada no comportamento de compra de clientes de um *e-commerce* brasileiro. Visto que cada cliente possui diferentes necessidades, a segmentação ajudará a direcionar os esforços de *marketing* para resultados mais assertivos. A abordagem utilizada faz uso do conceito de frequência, recência e valor monetário, comumente chamados de RFM, aliada posteriormente à clusterização dos dados. A partir da tabela de pedidos por cliente, gerou-se as métricas RFM para cada cliente único. As métricas foram submetidas a um processo de clusterização que utilizou os dois principais métodos particionais: k-médias e k-medóides. Os métodos foram escolhidos devido à escalabilidade dos algoritmos aliada à interpretabilidade do resultado. Em ambos os métodos foi possível identificar grupos equivalentes de clientes: clientes ativos de baixo valor e baixa frequência, clientes recorrentes, clientes fiéis e clientes de alto valor. Este trabalho também visa analisar as *reviews* do *e-commerce* e identificar onde as empresas podem melhorar para que seus consumidores tenham uma experiência de compra satisfatória. Identificou-se que o atraso na entrega era o maior motivo para a insatisfação dos clientes, e a solução apontada para resolver o atraso foi a troca de horário de despacho da mercadoria para a transportadora.

**Palavras-chave:** Marketing. Clusterização de dados. K-médias. K-medóides.

# ABSTRACT

For customer-focused companies, satisfaction is both a goal and a marketing tool, where its purpose is to help achieve their goals. Marketing is present in all areas of an organization because it helps drive decision-making. The purpose of this work is to develop a segmentation strategy based on the purchasing behavior of Brazilian e-commerce customers. Since each customer has different needs, segmentation helps direct marketing efforts towards more assertive results. The approach used makes use of the concept of frequency, recency and value received, commonly called RFM, followed after data clustering. From the orders per customer table, RFM metrics were generated for each unique customer. The metrics were observed in a clustering process that used the two main partition methods: k-means and k-medoids. The methods were chosen due to the scalability of the algorithms combined with the interpretability of the result. In the methods, it was possible to identify equivalent groups of customers: low-value and low-frequency active customers, recurring customers, loyal customers and high-value customers. This work also aims to analyze e-commerce evaluations and identify where companies can improve so that their consumers have a satisfactory shopping experience. It was identified that the delay in delivery was the main reason for customer dissatisfaction, and the solution suggested to resolve the delay was to change the shipping time of the goods to the carrier.

**Keywords:** Marketing. Data clustering. Machine learning.K-means. K-medoids.

**1 INTRODUÇÃO**

Quando as empresas falham em atender as expectativas do cliente, é comum observar queda nos indicadores de retenção e fidelidade, o que tem impacto direto na receita do negócio. Em contrapartida, exceder as expectativas pode gerar custos operacionais maiores do que os necessários e comprometer o retorno de investir na experiência do consumidor. A geração de valor é máxima quando a empresa entrega a experiência que está alinhada com a expectativa do cliente. Portanto, essa relação deve ser vista como uma decisão de investimento (MOSADDEGH et al., 2021; STAHL; MATZLER; HINTERHUBER, 2003).

Drucker (1998, p. 53), afirma que “[...] Muitas empresas compreendem o custo de obter um cliente, mas não entendem o custo de perder um cliente. De fato, custa de cinco a seis vezes mais conseguir um cliente novo do que reatar negócios com um cliente antigo”. Ou seja, nenhuma organização, seja lá qual for seu ramo, suporta perder clientes constantemente. Por isso é importante sempre estar atento aos fatores que levam a essa queda nos indicadores de retenção. Hoje em dia, graças aos recursos da Ciência de Dados e das ferramentas de Aprendizado de Máquina, é possível ter acesso a *insights* valiosos, facilitando a tomada de decisões.

A segmentação de mercado tem sido considerada um dos conceitos mais fundamentais do *marketing* moderno (WIND, 1978). Este conceito introduz a ideia de que a demanda de determinado mercado é heterogênea, todavia, as empresas podem aumentar sua lucratividade ao calibrar estratégias comerciais que considerem seu mercado consumidor como grupos cujos elementos possuem características similares.

O aprendizado de máquina encontra estrutura e significado para grandes volumes de dados, de forma rápida e eficiente ajudando na tomada de decisões, análise de investimentos e diversificação de estratégias. De acordo com Nisbet, Elder & Miner (2009, p.17), o uso de algoritmos de aprendizado de máquina para encontrar padrões tênues de relacionamento entre elementos de dados em grandes conjuntos de dados, pode levar a ações em ordem de obter algum benefício de alguma forma (diagnóstico, lucro, detecção, etc).

Uma das formas de fazer segmentação de mercado por meio de métodos de aprendizado de máquina é usá-los na análise de dados gerados pelo modelo RFM (recência, frequência e valor monetário). Esse tipo de modelo é bastante popular na pesquisa de *marketing* pois traz informações acerca do comportamento do cliente. É muito valioso para prever a resposta aos estímulos de *marketing,* e pode aumentar os lucros de uma empresa em curto prazo. Apesar de fornecer intuições importantes, apenas o cálculo das métricas RFM não permite a identificação de segmentos relevantes. Este problema é contornado através da utilização de métodos de aprendizado de máquina não supervisionados. Especificamente, este estudo utiliza os métodos de clusterização particionais k-médias e k-medóides para realizar a tarefa de agrupamento dos dados.

Segundo Philip Kotler (2000), um dos principais pensadores de *marketing* de todos os tempos, a satisfação do cliente consiste na sensação de prazer ou desapontamento resultante da comparação do desempenho, percebido de um produto em relação às suas expectativas. Sendo assim, a empresa de hoje precisa ter em mente que clientes satisfeitos permanecem na empresa por mais tempo, compram mais e falam favoravelmente da organização.

A satisfação do cliente é tão importante para o *marketing* quanto a segmentação, pois é um indicador de desempenho que mensura o quanto uma empresa consegue atender as expectativas do consumidor. Além do mais, ajuda a entender o nível de retenção, pois clientes satisfeitos tendem a se manter ativos. Em *e-commerces* é mais fácil mensurar a satisfação, pois temos acesso aos comentários da *reviews.*

**2 REFERENCIAL TEÓRICO**

A segmentação é representada por um ajuste racional e mais preciso do produto e do esforço de *marketing* aos requisitos do consumidor. Através da segmentação, as empresas dividem mercados grandes e heterogêneos em um número de mercados menores e homogêneos que podem mais facilmente serem atingidos.

**2.1 Segmentação de mercado**

Kotler (1980, p. 213), define que os compradores em qualquer mercado diferem em seus desejos, recursos, localizações, atitudes de compra e práticas de compra. As empresas capitalizam a heterogeneidade das necessidades identificando grupos de clientes que possuem conjuntos de necessidades relativamente homogêneos. Apesar de extensões consideráveis, exploração de variáveis e uma abundância de pesquisas sobre o tema, a essência da teoria da segmentação por quase 50 anos tem sido a noção de heterogeneidade das necessidades do cliente (BLOCKER; FLINT, 2007). De acordo com Veloso (2008, p.37), o conceito de segmentação pode ser resumido em três princípios básicos:

1. É um processo de agregação de consumidores baseado em determinadas características.
2. O agrupamento de consumidores permite identificar grupos que possuem necessidades e desejos similares.
3. Por possuírem características similares, os grupos responderão de maneira similar aos estímulos de marketing.

Os resultados da segmentação podem ser utilizados para embasar um amplo conjunto de decisões de negócios. Tynan e Drayton (1978, p. 346, apud Lunn) apontam que a segmentação de mercado é normalmente empregada para solucionar os seguintes problemas:

1. Definir um mercado: a segmentação auxiliará na compreensão do ambiente a partir da perspectiva do consumidor. Por exemplo, produtos que são vistos como concorrentes pela perspectiva do fabricante podem não ser considerados desta forma do ponto de vista do cliente.
2. Justificar princípios de ação para marcas e produtos: direciona estratégias para aumentar a retenção de clientes, converter compradores de outras marcas ou atrair um novo grupo de compradores para os produtos e serviços ofertados.
3. Posicionar as marcas e os produtos: diante dos inúmeros segmentos de mercado existentes, as empresas precisam direcionar seus esforços para os grupos que oferecem maior potencial de retorno.
4. Identificar lacunas no mercado: a segmentação de mercado permite também identificar grupos de consumidores cujas necessidades não estão sendo atendidas. Essas necessidades podem ser satisfeitas a partir do lançamento de novos produtos ou pela modificação de um produto ou serviço existente.

A segmentação de mercado é um instrumento poderoso para direcionar decisões de negócio. Os resultados da análise permitem que as empresas melhorem o posicionamento de marcas e produtos perante o consumidor.

**2.2 RFM**

Para atingir melhores níveis de retenção de clientes, uma empresa precisa alocar eficientemente seus recursos de *marketing*. Nesse sentido, o método de RFM foi introduzido por Hughes (1994) como ferramenta para avaliar o valor de determinado consumidor para uma empresa. Este modelo é usado em dados transacionais com base em um período de análise e tem como objetivo identificar os clientes mais importantes de uma base de dados através de atributos como a recência, que é o tempo decorrido desde a última interação, a frequência, que é o número de interações em um determinado intervalo de tempo e valor, que é o valor monetário envolvido nas transações no período selecionado.

**2.3 Clusterização**

O aprendizado de máquina pode ser catalogado em aprendizado supervisionado e aprendizado não supervisionado. O aprendizado supervisionado ocorre quando o modelo aprende a partir de dados pré definidos, ou seja, os métodos contidos nesse tipo de aprendizagem, inferem a função a partir de dados rotulados (BERTOLINI et al., 2021), diferente do aprendizado não supervisionado que identifica semelhanças nos dados e reage conforme a presença ou ausência dessas semelhanças

Quando realizada uma clusterização, as observações de um conjunto de dados são agrupadas, dividindo-os em grupos distintos para que as observações dentro de cada grupo sejam bastante semelhantes entre si. Clusterização de dados se refere a um amplo conjunto de técnicas de Mineração de Dados com o intuito de agrupar os dados segundo alguma métrica de associação. De acordo com Han & Kamber (2011, p.8), Mineração de Dados é definido como o processo de descoberta de padrões e conhecimentos interessantes de grandes quantidades de dados. Suas fontes podem incluir bancos de dados, *data warehouses*, a internet, outros repositórios de informações ou dados que são transmitidos dinamicamente para um sistema. Frank (2002) reforça que a maioria das técnicas de Mineração de Dados desenvolveram-se em um campo conhecido hoje como Aprendizado de Máquina.

O funcionamento do modelo de segmentação RFM e suas aplicações encontradas na literatura são comumente combinadas com outras técnicas de clusterização como k-médias e métodos hierárquicos. A clusterização é bastante conhecida na literatura de análise de dados por sua vasta aplicação em questões como segmentação de clientes, classificação e análise de tendências. Este é um problema não supervisionado porque tenta-se descobrir a estrutura (neste caso grupos distintos) com base em um conjunto de dados (GARETH et al., 2013).

**3 MATERIAIS E MÉTODOS**

O conjunto de dados utilizados nesta pesquisa foi cedido pela Olist, empresa brasileira fundada em 2015 que atua no segmento de *e-commerce*. Dentre as soluções oferecidas está a Olist Store. De acordo com o próprio site da empresa, o Olist Store é uma solução de vendas que une tecnologia de ponta com inteligência de mercado para aumentar o faturamento de lojistas nos *marketplaces*. Os dados estão disponíveis no Kaggle, uma plataforma de aprendizado e prática de *Data Science*. Os arquivos disponibilizados contém informações reais de 100 mil pedidos realizados entre 2016 a 2018 em vários marketplaces do Brasil. De maneira geral, é possível encontrar informações de pedidos, pagamentos, clientes e produtos (OLIST; SIONEK, 2018).

Como ferramenta para manipulação dos dados, este trabalho utiliza a linguagem de programação Python, pois deriva de um ativo ecossistema de pacotes e bibliotecas para análise de dados.

Para a análise das *reviews* foi utilizado a biblioteca *WordCloud*. Com ela temos uma forma visual de plotar a frequência das palavras em determinado texto. Logo após identificar o motivo da insatisfação dos consumidores, que nesse caso foi o atraso na entrega das mercadorias, foi proposto descobrir uma forma de contornar a situação.

**3.1 Métodos de clusterização**

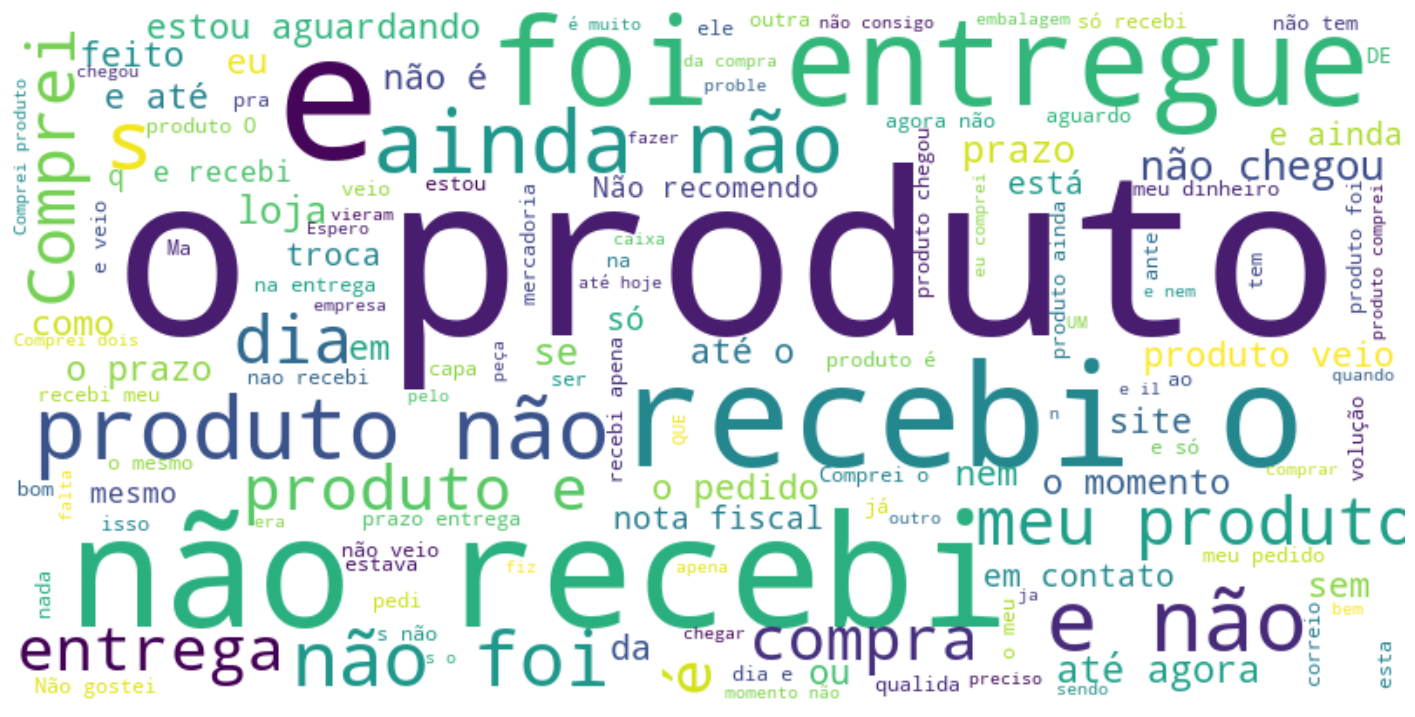
Foram comparados dois métodos de clusterização diferentes, k-médias e k-medóides. K-médias é uma técnica clássica de particionamento de *clustering* que agrupa o conjunto de dados de n objetos em k grupos. Dado um conjunto inicial de centróides, o algoritmo prossegue alternando entre duas etapas:

1. Etapa de atribuição: atribui cada observação ao grupo com a média mais próxima.
2. Etapa de atualização: Calcula as novas médias para ser o centróide das observações no grupo.

O algoritmo k-médias é mais sensível a outliers, em relação ao k-medóides Han & Kamber (2011, p.454). O k-medóides é um [algoritmo](http://en.wikipedia.org/wiki/Algorithm) [de agrupamento](http://en.wikipedia.org/wiki/Data_clustering) relacionado ao algoritmo [k-m](http://en.wikipedia.org/wiki/K-means)édias e ao algoritmo *medoidshift*. Enquanto o k-médias tenta minimizar o [erro quadrático](http://en.wikipedia.org/wiki/Mean_square_error) total, o k-medóides minimiza a soma das dissimilaridades entre os pontos rotulados como pertencentes a um grupo e um ponto designado como o centro desse grupo, por isso o k-medóides pode ser mais robusto a ruído e *outliers* em comparação com [k](http://en.wikipedia.org/wiki/K-means)-médias porque minimiza uma soma de dissimilaridades gerais aos pares em vez de uma soma de distâncias euclidianas quadradas. A desvantagem do k-medóides é que utiliza de uma abordagem computacionalmente mais custosa, com complexidade computacional de O(N²KT), onde N é o número de amostras, T é o número de iterações e K é o número de grupos. Isso o torna mais adequado para conjuntos de dados menores em comparação com k-médias que é O(NKT) (SCIKIT LEARN, 2019).

**3.2 Análise de *reviews***

Após uma análise exploratória nos dados das *reviews,* com a ajuda de gráficos e tabelas, chegou-se a conclusão que 77,1% dos clientes parecem satisfeitos com a compra (4 acima), enquanto 22,9% estão insatisfeitos (3 abaixo). Para entender melhor o motivo da insatisfação dessa porcentagem de clientes foi feito uso de uma poderosa ferramenta de Ciência de Dados usado quando se trabalha com textos para criação de nuvem de palavras, a biblioteca *WorldCloud.* Foi identificado que o maior motivo de insatisfação dos clientes era o atraso nas entregas ou até mesmo a falta dela.



Estudando a fundo os dados da logística, gerou-se uma hipótese de que haveria uma grande chance de reduzir a quantidade de entregas atrasadas simplesmente trocando os horários mais frequentes de despacho de mercadorias para a transportadora. Essa hipótese foi confirmada com o teste de significância p-valor.

Entende-se o p-valor como um indicador de probabilidade de que a estatística do teste tenha um valor extremo em relação ao valor observado quando a hipótese nula é verdadeira. Valores próximos a 0 indicam que há fortes evidências contra a hipótese nula, enquanto valores próximos de 1 representam que a probabilidade da hipótese nula é verdadeira (Dahiru, T, 2011).

Foi identificado nos dados de despacho das mercadorias que os horários de 14h, 15h e 19h não são os de melhores performances para entrega sem atraso. É provável que haja uma redução dos atrasos que existem se as mercadorias fossem entregues à transportadora/correios nos horários de 9h, 10h e 11h. Através dessas informações foi gerado dois grupos (grupo A e grupo B) com os diferentes horários para que fosse possível ser feita a estatística de teste e obter o p-valor. O resultado do p-valor foi igual a 0.0, o qual significa que é estatisticamente relevante a diferença de resultado nos atrasos de entregas entre os grupos A e B, não sendo essas diferenças dadas ao acaso, ou seja, a hipótese tem força estatística.

**3.3 Métricas**

Para determinar a quantidade de grupo central na análise de grupos foi utilizado o método do cotovelo, que fornece uma heurística para determinar um valor apropriado de grupos através da redução da soma da variância (HAN; KAMBER, 2006, pp. 486). Este procedimento é baseado na ideia de que aumentar o número de grupos permite capturá-los com objetos mais semelhantes entre si. Porém, o efeito marginal de reduzir a soma das variâncias dentro do grupo perde intensidade à medida que mais grupos são formados.

No algoritmo k-médias, definiu-se como parâmetro aceitável 5 grupos. Por outro lado, no k-medóides, por ser um algoritmo computacionalmente mais custoso, foi preciso executar 10 repetições com apenas 10% dos dados, escolhidos aleatoriamente para cada iteração para que o processo não consumisse mais memória do que o computador suporta. Ao final foi escolhido o resultado com o melhor *score* que definiu como parâmetro aceitável 6 grupos.

**4 RESULTADOS E DISCUSSÃO**

A partir dos grupos gerados por cada um dos modelos aplicados, realizou-se uma análise das características de cada grupo com intuito de identificar suas particularidades. Os resultados do agrupamento são apresentados nas Tabelas 1 e 2.

Tabela 1 - Média das métricas RFM para o modelo k-médias.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CLUSTER | MÉDIA RECÊNCIA | MÉDIA FREQUÊNCIA | MÉDIA VALOR | NÚMERO DE CLIENTES |
| 0 | 152.3 | 1.0 | 61.9 | 29763 |
| 1 | 121.2 | 1.0 | 225.7 | 24965 |
| 2 | 221.8 | 2.1 | 275.9 | 2736 |
| 3 | 368.4 | 1.0 | 255.6 | 17113 |
| 4 | 426.2 | 1.0 | 70.7 | 17875 |

Tabela 2 - Média das métricas RFM para o modelo k-medóides.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CLUSTER | MÉDIA RECÊNCIA | MÉDIA FREQUÊNCIA | MÉDIA VALOR | NÚMERO DE CLIENTES |
| 0 | 243.5 | 1.0 | 127.9 | 17973 |
| 1 | 396.9 | 1.0 | 55.6 | 15437 |
| 2 | 187.5 | 1.1 | 423.4 | 11283 |
| 3 | 455.7 | 1.0 | 201.9 | 12852 |
| 4 | 75.7 | 1.0 | 147.2 | 16236 |
| 5 | 127.5 | 1.0 | 50.9 | 18671 |

Como forma de facilitar a análise, cada grupo foi nomeado de acordo com seu principal atributo.

1. Ocasional ativo: são clientes de baixo valor, baixa frequência e baixa recência.
2. Ocasional inativo: são clientes de baixo valor, baixa frequência e alta recência.
3. Ativo de alto valor: são clientes de alto valor e baixa recência.
4. Ativo de baixo valor: são clientes de baixo valor e baixa recência.
5. Inativo de alto valor: são clientes de alto valor e alta recência.
6. Inativo de baixo valor: são clientes de baixo valor e alta recência.
7. Frequente: são clientes com maior frequência.

Estão representados na tabela 3 e 4.

Tabela 3 - Categoria de clientes k-médias.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Característica | Número de clientes | Porcentagem |
| 0 | Ocasional ativo | 29763 | 32,19% |
| 1 | Ativo de alto valor | 24965 | 27% |
| 2 | Frequente | 2736 | 2,96% |
| 3 | Inativo de alto valor | 17113 | 18,51% |
| 4 | Ocasional inativo | 17875 | 19,33% |

Tabela 4 - Categoria de clientes k-medóides.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Característica | Número de clientes | Porcentagem |
| 0 | Inativo de baixo valor | 17973 | 19,44% |
| 1 | Ocasional inativo | 15437 | 16,7% |
| 2 | Ativo de alto valor | 11283 | 12,2% |
| 3 | Inativo de alto valor | 12852 | 13,9% |
| 4 | Ativo de baixo valor | 16236 | 17,56% |
| 5 | Ocasional ativo | 18671 | 12,2% |

Nota-se que os resultados foram distintos, pois além do no modelo k-medóides possuir 6 grupos, identifica-se características que no k-médias não são encontradas e vice-versa.

O modelo k-médias, foi computacionalmente mais eficiente, o que permite que um número maior de modelos sejam treinados com diferentes parâmetros em um menor intervalo de tempo em relação ao k-medóides. Já o modelo k-medóides, mesmo sendo computacionalmente mais custoso, leva vantagem sobre o modelo k-médias por sua capacidade de lidar com *outliers,* como apontado por Han & Kamber (2011, p.454). Além disso, o k-medóides identificou duas características que o k-médias não encontrou, ativo de baixo valor e inativo de baixo valor, possuindo uma métrica considerável de dados, sendo 19,44% para inativo de baixo valor e 17,56% para ativo de baixo valor.

O modelo k-médias encontrou a característica frequência, essa característica não foi identificada pelo k-medóides. Além disso, a porcentagem de clientes com essa métrica é de apenas 2,96% dos dados.

**5 CONCLUSÃO**

Este trabalho surge da necessidade de entender a insatisfação de clientes de uma empresa de *e-commerce* e encontrar uma maneira de melhorar a experiência de compra. Surge também da necessidade de fornecer embasamento para o direcionamento de estratégias comerciais focalizadas em determinados segmentos de clientes. Sabe-se que os mercados são heterogêneos e, portanto, o processo de segmentação emerge como o primeiro passo para o desenvolvimento de uma estratégia de *marketing* eficaz.

Este estudo almejou identificar os principais segmentos de clientes baseado em métricas de recência, frequência e valor monetário, objetivo que foi alcançado ao final da análise. Optou-se pela utilização dos métodos de agrupamento particional k-médias e k-medóides devido à facilidade de implementação, escalabilidade e interpretabilidade dos resultados. Ambos os métodos foram extensivamente explorados na literatura neste tipo de problema, produzindo resultados satisfatórios.

Os métodos de agrupamentos adotados demandam a pré-definição do número de grupos. Com auxílio do método do cotovelo definiu-se cinco e seis grupos, que foram capazes de produzir resultados satisfatórios, isto é, alta diferenciação entre clusters e alta similaridade entre observações dentro de um mesmo grupo. Os resultados obtidos foram então analisados com intuito de aprofundar o conhecimento acerca do conjunto de métricas que define cada partição encontrada.

Para o objetivo de direcionar os esforços de *marketing* para resultados mais assertivos, o modelo k-medóides se mostrou mais eficiente, pois segmentou os clientes de maneira mais detalhada, facilitando a tomada de decisão para uma campanha de *marketing* mais direcionada.

Um resultado satisfatório também foi obtido na análise das *reviews*, onde, com a ajuda de bibliotecas adequadas para análise de dados, identificou-se o motivo da insatisfação dos clientes. A partir disso foi encontrada a solução para contornar o problema.

Ainda que os objetivos propostos tenham sido alcançados, há muito espaço para aprofundar a análise. Este estudo limitou-se a segmentar os clientes em relação às métricas RFM. No entanto, múltiplas bases de segmentação podem ser utilizadas para obter grupos cada vez mais bem definidos, o que viabiliza a construção de estratégias de negócio.

**REFERÊNCIAS**

ANITHA, Palaksha; PATIL, Malini M. RFM model for customer purchase behavior using K-Means algorithm. **Journal of King Saud University-Computer and Information Sciences**, Elsevier, 2019.

BALAKRISHNAN, PV Sundar et al. Comparative performance of the FSCL neural net and K-means algorithm for market segmentation. **European journal of operational research**, Elsevier, v. 93, n. 2, p. 346–357, 1996.

BERTOLINI, Massimo et al. Machine Learning for industrial applications: A

comprehensive literature review. **Expert Systems with Applications**, v. 175,

p. 114820, mar. 2021. DOI: 10.1016/j.eswa.2021.114820.

BLOCKER, Christopher P; FLINT, Daniel J. Customer segments as moving targets: integrating customer value dynamism into segment instability logic. **Industrial Marketing Management**, Elsevier, v. 36, n. 6, p. 810–822, 2007.

Dahiru, T. (2011). P-Value, a true test of statistical significance? a cautionary note. **Annals of Ibadan Postgraduate Medicine**, 6(1). doi:10.4314/aipm.v6i1.64038.

DRUCKER, Peter F. **O melhor de Peter Drucker:** a administração. São Paulo:

Nobel, 2002.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

GARETH, James et al. **An introduction to statistical learning: with applications in**

**R**. [S.l.]: Spinger, 2013.

HAN, Jiawei; KAMBER, Micheline. Data mining: concepts and techniques, 2nd.

**University of Illinois at Urbana Champaign: Morgan Kaufmann**, 2006.

HUGHES, Arthur Middleton. **Strategic database marketing : the masterplan for starting and managing a profitable, customer-based marketing program**. [S.l.]: Chicago, Ill. : Probus Pub. Co., 1994., 1994.

KAHAN, Ron. Using database marketing techniques to enhance your one-to-one marketing initiatives. **Journal of Consumer Marketing**, MCB UP Ltd, 1998.

KOTLER, P. **Administração de Marketing: Análise, planejamento, implementação e controle**. Tradução Ailton Bomfim Brandão- 5. ed. – 8 reimpr - São Paulo Atlas, 2009. p. 726.

KOTLER, P. **Administração de Marketing: A edição do novo milênio** - Traducao Bazan Tecnologia e Linguística, revisão técnica Arão Sapiro -10.ed - São Paulo: Prentice Hall,2000. p. 764.

KOTLER, P. **Principles of Marketing**. [S.l.]: Prentice-Hall, 1980. (The Prentice-Hall series in marketing). ISBN 9780137015573. Disponível em:

<https://books.google.com.br/books?id=%5C_HrT6bficY4C>.

MOSADDEGH, Abdolreza et al. Dynamics of customer segments: A predictor of

customer lifetime value. **Expert Systems with Applications**, Elsevier, v. 172,

p. 114606, 2021.

NISBET, Robert; ELDER, John; MINER, Gary D. **Handbook of statistical analysis and data mining applications**. [S.l.]: Academic press, 2009.

OLIST; SIONEK, André. **Brazilian E-Commerce Public Dataset by Olist**. [S.l.: s.n.], 2018. <https://www.kaggle.com/dsv/195341>.

SCIKIT LEARN, Extra. Clustering with KMedoids and Common-nearest-neighbors. **Scikit Learn**, 2019. Disponível em: <https://scikit-learn-extra.readthedocs.io/en/stable/modules/cluster.html#k-medoid>. Acesso em: 20 fev. 2023.

THISTED, Roanld A.. What is a P-value?. **Departments of Statistics and Health Studies The University of Chicago**, 2010.

TYNAN, A Caroline; DRAYTON, Jennifer. Market segmentation. **Journal of marketing management**, Taylor & Francis, v. 2, n. 3, p. 301–335, 1987.

VELOSO, Andres Rodriguez. **Estratégias de segmentação e posicionamento direcionadas para o mercado infantil**. 2008. Tese (Doutorado) – Universidade de São Paulo.

WEI, Jo-Ting; LIN, Shih-Yen; WU, Hsin-Hung. A review of the application of RFM model. **African Journal of Business Management**, v. 4, p. 4199–4206, 2010.

WIND, Yoram. Issues and advances in segmentation research. **Journal of marketing research**, SAGE Publications Sage CA: Los Angeles, CA, v. 15, n. 3, p. 317–337, 1978.

WITTEN, Ian H; FRANK, Eibe. Data mining: practical machine learning tools and techniques with Java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.