



ConBRepro

XIV CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO

I CIGELUBRA

Congresso Internacional de Gestão e Engenharia Luso-Brasileiro

04 a 06 de dezembro de 2024

Energias Limpas nas Engenharias

Clusterização de clientes utilizando variáveis de recência, frequência e valor

Sabrina Fedrizzi

Programa de Pós-Graduação em Engenharia de Produção - UFPR

Mariana Kleina

Programa de Pós-Graduação em Engenharia de Produção - UFPR

Resumo: Sabe-se que para uma empresa ser competitiva no mercado, ela precisa ter seu foco voltado para os clientes. Entretanto, clientes possuem comportamento de compra diferentes e não devem ser tratados de maneira igual pela equipe de *marketing* da empresa, e com a segmentação desses clientes é possível direcionar estratégias de *marketing* para os diferentes grupos. Assim, esse estudo propõe o uso da técnica de clusterização de dados K-Means, utilizando as variáveis dos clientes de recência, frequência e valor, para agrupar clientes com perfis semelhantes de compra, de uma loja online de varejo com sede no Reino Unido. Para a escolha do número adequado de grupos, foram utilizados os índices Silhueta e Dunn, os quais apontaram o número ideal de grupos como sendo 2. Os grupos formados tiveram características bastante distintas e resumidamente podem ser descritos como clientes menos valiosos ou desengajados e clientes mais valiosos ou engajados.

Palavras-chave: Clusterização, Clientes, Recência, Frequência, Valor.

Customer clustering using recency, frequency, and monetary value variables

Abstract: It is known that for a company to be competitive in the market, it needs to focus on its customers. However, customers exhibit different buying behaviors and should not be treated the same way by the company's marketing team. By segmenting these customers, it becomes possible to tailor marketing strategies to the different groups. Thus, this study proposes the use of the K-Means clustering technique, using customer variables such as recency, frequency, and monetary value, to group customers with similar purchasing profiles from an online retail store based in the United Kingdom. To determine the appropriate number of groups, the Silhouette and Dunn indices were used, both of which indicated the ideal number of groups as 2. The formed groups had significantly distinct characteristics and can be broadly described as less valuable or disengaged customers and more valuable or engaged customers.

Keywords: Clustering, Customers, Recency, Frequency, Monetary Value.

1. Introdução

Em um mercado altamente competitivo, uma empresa, independente do ramo de negócio, precisa entender o comportamento de compra dos seus clientes para conseguir realizar um tratamento diferenciado de acordo com o perfil de cada um e, somente assim, ser possível

manter a relação com seus consumidores e a sua rentabilidade a longo prazo. A nova economia está amplamente focada em uma melhor prestação de serviços e em um atendimento personalizado, sendo que a era atual é chamada de economia orientada para o cliente (ZOERAM; MAZIDI, 2018), a qual coloca as necessidades dos clientes em primeiro lugar para nortear as decisões do *marketing* do negócio.

Empresas podem ter milhares de clientes e é inviável tratar todos eles de maneira igualitária, mas também é inadequado tratá-los de maneira única. Sendo assim, uma estratégia que pode ser utilizada para a personalização no atendimento é a segmentação desses clientes em grupos que representem seus diferentes comportamentos de compra (KASEN *et al.*, 2024).

Para segmentar clientes de uma empresa, são necessários dados desses clientes e de suas compras. Possíveis informações que podem ser usadas para tal objetivo são oriundas do modelo de recência, frequência e valor (RFV), pois este avalia quão recentemente um cliente comprou, com que frequência faz compras e quanto gasta em média (WEI *et al.*, 2010). Estratégias baseadas nessas informações permitem que a equipe de Gestão de Relacionamento com o Cliente (*Customer Relationship Management* - CRM) ofereça serviços, descontos e campanhas publicitárias direcionadas a clientes específicos com base em suas preferências, visando um relacionamento de longo prazo e a consequente fidelização e rentabilidade, fatores essenciais para a continuidade dos empreendimentos (ANSHARI *et al.*, 2019).

Utilizando informações provenientes da análise RFV, a segmentação de clientes frequentemente é realizada por meio da clusterização de dados, que se refere ao processo de organizar dados de modo que clientes “semelhantes” fiquem no mesmo grupo (ou cluster) e os grupos formados sejam os mais distintos possíveis entre si (IKOTUN *et al.*, 2023).

Existem inúmeras técnicas de clusterização, sendo o K-Means a mais difundida na literatura (AHMED *et al.*, 2020). Esse método também é bastante implementado na prática, pois é simples e apresenta bom desempenho computacional mesmo em clusterizações que exigem a análise de grandes bases de dados (WU *et al.*, 2008).

Muitos estudos utilizam K-Means e as variáveis RFV para formar grupos de clientes com base em sua semelhança. Pode-se citar os trabalhos de Essayem *et al.* (2022) e Anitha e Patil (2019) que aplicam a metodologia a clientes de uma loja de varejo, Perdhana e Heikal (2024) melhoram a segmentação de clientes em serviços de transporte online, Demircioğlu *et al.* (2023) realizam a segmentação de clientes de uma empresa de entrega expressa, dentre muitos outros presentes na literatura.

Nesse sentido, o presente estudo tem como objetivo a aplicação da técnica de clusterização K-Means para segmentação de clientes de uma loja de varejo online, com sede no Reino Unido, entre os anos de 2010 e 2011. Para alocar clientes com base em sua semelhança de compra, serão utilizadas as informações de recência, frequência e valor desses clientes, e para se descobrir um número ideal de grupos serão utilizados os índices de Silhueta e Dunn. A principal ideia é que, com a aplicação da metodologia, aumentem-se as vendas e o lucro do negócio a partir da criação de estratégias personalizadas de publicidade e venda que tomem como base o agrupamento obtido e suas respectivas características, assim como sejam fornecidas percepções valiosas na previsão do comportamento de compra futura dos consumidores.

Este artigo está dividido em 5 seções. Esta primeira contextualiza o tema geral do trabalho, descrevendo a importância de segmentar clientes e da utilização de técnicas de análise RFV e de clusterização. A segunda seção traz uma revisão do conceito de RFV, bem como da técnica de clusterização K-Means e dos índices de validação da clusterização. A terceira

seção descreve a metodologia da pesquisa, onde é apresentada a base de dados, o tratamento realizado nela e como os índices RFV foram calculados. Na quarta seção são apresentados e discutidos os resultados da pesquisa. Por fim, na quinta seção são apresentadas as conclusões do trabalho.

2. Referencial Teórico

2.1 Análise RFV

A sigla RFV (RFM, no inglês) representa uma técnica para classificar os consumidores de acordo com o valor de três elementos essenciais baseados em suas características de compras (ANITHA; PATIL, 2019):

- Recência (R): período de tempo decorrido entre uma data de referência e a última compra do cliente; quanto mais recente, maior será a recência do cliente;
- Frequência (F): número de vezes que o cliente comprou algum produto em um determinado período de tempo;
- Valor (V): média ou valor total gasto pelo cliente durante determinado período de tempo.

O foco da análise RFV é avaliar a qualidade e o valor dos clientes com base nesses três elementos. Usando essa análise, é possível atribuir um *score* RFV para cada cliente, baseado em seu comportamento de compra e posteriormente classificá-lo (MOMTAZ *et al.*, 2013).

O modelo de análise RFV apresenta a grande vantagem de ser simples e de fácil compreensão para todos os envolvidos no processo, assim como sua eficiência e efetividade para a classificação da rentabilidade dos clientes de acordo com poucos aspectos (WEI *et al.*, 2010).

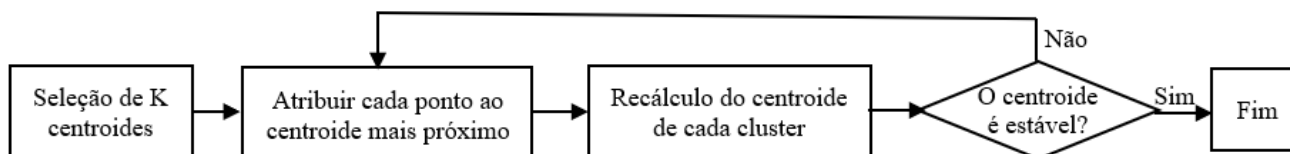
2.2 Método K-Means

O algoritmo K-Means tem como base a escolha de K centroides (ponto representativo de cada grupo) aleatórios (sendo K o número de clusters) e, a partir desses pontos, ajustá-los com base nos dados, a fim de encontrar o centro dos clusters (IKOTUN *et al.*, 2023). A Figura 1 ilustra o fluxograma do algoritmo K-Means.

Das (2023), em convergência com o apresentado na Figura 1, descreve as 4 etapas método K-Means, sendo:

- 1) Seleção aleatória no número de centroides iniciais (K);
- 2) Alocação de cada ponto ao cluster cujo centroide apresenta menor distância euclidiana até o ponto em questão;
- 3) Recálculo dos centroides de cada cluster;
- 4) Repetição das etapas 2 e 3 até que os centroides calculados na terceira etapa não sejam alterados ou até que se atinja um número máximo de iterações definido pelo usuário.

Figura 1 – Fluxograma das etapas do algoritmo K-Means



Fonte: Adaptado de Scaburi *et al.* (2017)

A grande dificuldade na aplicação do algoritmo K-Means, segundo Arthur e Vassilvitskii (2007), é a escolha adequada do número de clusters (K), pois ele afeta diretamente na

escolha dos centroides iniciais. Os autores ressaltam, ainda, que a normalização dos dados do conjunto abordado é imprescindível para a obtenção de clusters fortes.

2.3 Índices de validação da clusterização

Existem diversas métricas para avaliar quão boa é uma clusterização, dentre elas pode-se citar o coeficiente Silhueta e o índice Dunn. Tais métricas também são usadas para escolher adequadamente o número de grupos da clusterização.

A Silhueta é uma medida da distância entre os grupos, utilizada para determinar a qualidade dos clusters. Para cada observação é definido um coeficiente que compara a distância da observação ao cluster e pode ser calculado de acordo com a equação 1.

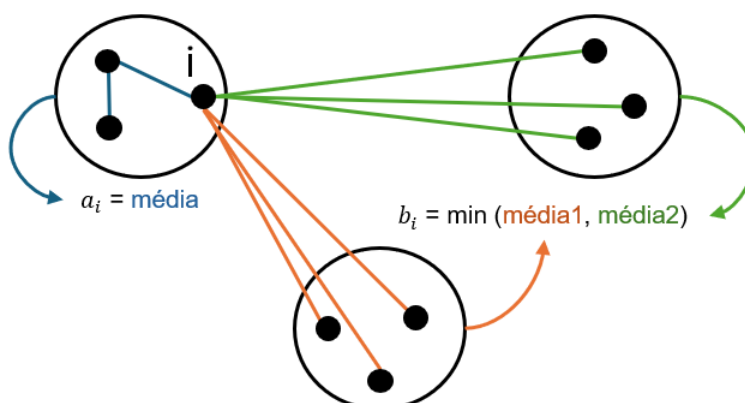
$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (1)$$

Onde:

- a_i : distância média entre o ponto i aos outros do mesmo cluster;
- b_i : menor distância média entre a observação i aos outros clusters.

A Figura 2 ilustra como os cálculos das distâncias são realizados.

Figura 2 – Distâncias Coeficiente Silhueta



Fonte: Adaptado de Annachhatre et al. (2015)

A média de todos os valores s_i calculados retornará um índice entre $[-1,1]$, representando a qualidade da clusterização geral, conforme explicitado pela Tabela 1. Assim, quanto mais perto de 1, melhor a qualidade do cluster, ou seja, sua heterogeneidade é menor (ROUSSEUW; KAUFMAN, 1990).

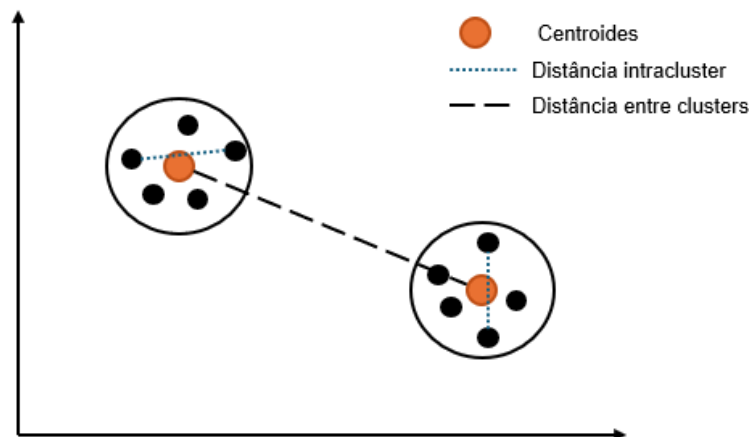
Tabela 1 – Interpretação do Coeficiente Silhueta	
Coeficiente de Silhueta	Interpretação
0,71 a 1,00	Estrutura forte
0,51 a 0,70	Estrutura razoável
0,26 a 0,50	Estrutura fraca
Menor que 0,25	Nenhuma estrutura

Fonte: Os autores (2024)

Já o Índice de Dunn mede a quão compacto é um cluster, ou seja, sua distância intracluster, assim como a separação entre os grupos formados, distância entre clusters. Ele é calculado

pela razão entre a menor distância entre observações que não estão no mesmo cluster e a maior distância intracluster (NCIR *et al.*, 2021). A Figura 3 ilustra como essas características podem ser visualizadas. O Índice de Dunn tem um valor entre zero e infinito e deve ser maximizado.

Figura 3 – Distâncias Índice de Dunn



Fonte: Adaptado de Ncir *et al.* (2021)

3. Metodologia

Todo o desenvolvimento do presente estudo se deu no *software* R. A base utilizada foi a *Online Retail II*, disponível no *UCI Machine Learning Repository*, e consiste em 541910 dados transacionais de uma loja online de varejo, com sede no Reino Unido, entre 01/12/2010 e 09/12/2011. A loja vende principalmente itens de presentes, que são vendidos principalmente para empresas.

A base utilizada contém 8 variáveis, dentre elas: o número identificador do cliente, a data da transação, a quantidade de cada item pedido, e o preço unitário do item. Porém, foi necessário fazer uma limpeza nos dados, pois havia linhas com valores ausentes e transações que resultaram em devolução.

Após essa limpeza, o conjunto de dados ficou com 397885 transações, feitas por 4338 clientes distintos. Também, foi criada uma variável, o valor monetário da transação, que consiste em multiplicar a quantidade de cada item pedido pelo valor unitário do item.

Para determinar os componentes RFV de cada cliente, utilizou-se:

- Para o cálculo da recência (R): a data mais recente do conjunto de dados, ou seja, a data da última transação realizada e subtraiu-se dela a data mais recente de compra de cada cliente;
- Para o cálculo da frequência (F): o número de transações que cada cliente fez;
- Para o cálculo do valor (V): o total gasto por cada cliente em todas as suas transações.

Com os valores de recência, frequência e valor de cada cliente, estes três elementos foram classificados em uma escala de 1 a 5, utilizando os quintis, onde:

- 1: menor valor (menos recente, menos frequente, menor valor monetário);
- 5: maior valor (mais recente, mais frequente, maior valor monetário).

Assim, um cliente com R=5, F=5 e V=5, resultando no *score* RFV=555 (concatenação dos três números) é o mais valioso para a companhia, bem como o cliente com *score* RFV=111 é o menos valioso.

Antes de aplicar a técnica de clusterização K-Means, realizou-se a padronização dos dados de recência, frequência e valor, por meio da fórmula z-score, na qual o valor normalizado é igual a razão entre o valor original subtraído da média do conjunto de dados e o desvio padrão do conjunto. Essa etapa é importante para garantir que os dados sejam analisados considerando uma escala padrão e assegurar a construção de clusters fortes.

Para escolher o número adequado de partições, na clusterização variou-se de 2 a 5 o valor de K, e foram calculados os índices Silhueta e Dunn para cada uma das 4 configurações. Por meio do valor ideal de K, obtido pela clusterização que retornou os maiores valores de Silhueta e Dunn, os grupos formados foram enfim analisados.

4. Resultados e Discussões

O algoritmo K-Means foi executado com o número de clusters (K) variando de 2 a 5, considerando 2 como o número mínimo possível para a geração de grupos e 5 como uma quantidade possível de ser analisada e factível em relação aos comportamentos dos consumidores. Os índices Silhueta e Dunn foram calculados para cada um dos casos e são apresentados na Tabela 2.

Tabela 2 – Valores de índices Silhueta e Dunn para diferentes valores de K

K	Silhueta	Dunn
2	0,8958	0,0525
3	0,5937	0,0003
4	0,6162	0,0003
5	0,5994	0,0008

Fonte: Os autores (2024)

Ambos os índices indicaram o melhor número de clusters como sendo 2, visto que o Silhueta é melhor quanto mais próximo de 1 e o índice Dunn deve ser maximizado. Portanto a análise da clusterização e a aplicação do método K-Means se deu com a separação dos clientes em dois grupos. A Tabela 3 apresenta alguns dados estatísticos acerca dos dois clusters formados.

Tabela 3 – Resumo dos clusters formados

Cluster	Número de Clientes	Média Recência	Média Frequência	Média Valor
1	4312	92,6 dias	3,9 vezes	£1549
2	26	5,5 dias	66,4 vezes	£85904

Fonte: Os autores (2024)

É possível notar que no Cluster 1 os clientes têm uma recência bem mais alta (aproximadamente 87 dias a mais em relação à média apresentada pelo Cluster 2), o que significa que não compram há bastante tempo, isto é, são menos recentes em termos de suas compras. Em relação a frequência, esses clientes têm o perfil de efetuarem poucas compras na empresa, o que fica claro se compararmos sua periodicidade em relação ao segundo grupo, o qual apresentou em torno de 62 transações a mais. E sobre o valor médio gasto nas aquisições realizadas no período considerado, o Cluster 1 apresentou valor expressivamente menor, representando cerca de 2% apenas do gasto do Cluster 2. Ou seja, além de comprarem com menor frequência, o perfil de compras do primeiro grupo é definido por compras menores ou de itens de preços mais baixos.

Dessa forma, os clientes representados pelo Cluster 1 provavelmente são menos valiosos para a empresa e podem estar em risco de *churn* ou podem ser clientes menos engajados.

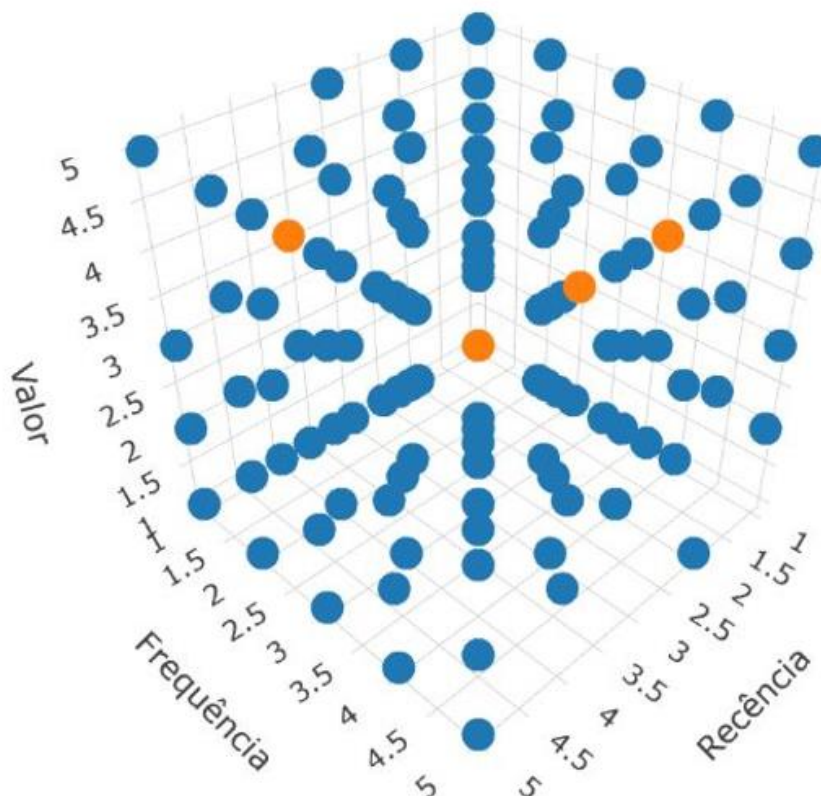
Entretanto, eles representam 99,4% dos clientes da empresa no período. Como o tamanho do Cluster 1 é significativo, os dados apresentados podem ser importantes para entender o perfil deste grupo e permitir que a companhia crie campanhas de reengajamento ou ofertas especiais utilizando uma abordagem personalizada e que gere curiosidade e interesse neste grupo para tentar trazê-los de volta.

Sobre o Cluster 2, a média da recência é baixa, significando que os clientes desse grupo fizeram compras recentemente. A média da frequência é alta, indicando que esses clientes fazem compras com frequência. E o valor gasto médio também é alto, indicando que esses clientes gastam mais dinheiro nas suas compras com maior quantidade de itens ou produtos de maior valor agregado.

Os clientes do segundo grupo podem ser classificados como mais valiosos para a loja ou mais engajados e são essenciais para a receita da empresa. Nesse sentido, campanhas de fidelização, estratégias de *cross-selling/up-selling* e até ofertas exclusivas podem ser feitas para mantê-los ativos. Contudo, esse grupo é composto por apenas 26 clientes, o que representa 0,6% do total de clientes da empresa no período.

A Figura 4 apresenta as variáveis recência, frequência e valor, posicionando cada cliente de acordo com seu cluster. Nota-se que, conforme já mencionado, o Cluster 1 (cor azul) é o mais representativo em termos de número de clientes, porém os *scores* RFV desses clientes são os mais variados possíveis. Já o Cluster 2 (cor laranja), com menor representatividade, agrupou 26 clientes, sendo que esses clientes têm *scores* RFV igual a 355, 455, 535 e 555.

Figura 4 – Valores RFV dos clientes que estão no Cluster 1 (azul) e Cluster 2 (laranja)



Fonte: Os autores (2024)

5. Conclusões

Esta pesquisa abordou a segmentação de clientes de uma base de dados de uma loja online de varejo. As variáveis utilizadas no estudo foram a recência, frequência e valor de

cada um dos clientes, de acordo com o método de análise RFV. A técnica utilizada para o agrupamento foi o bem conhecido algoritmo K-Means. Para a escolha do número de clusters, utilizou-se as métricas de Silhueta e Dunn, as quais indicaram que segmentar a base em 2 grupos era a melhor configuração.

Os dois grupos formados são bem distintos. O primeiro (Cluster 1) é composto pela ampla maioria dos clientes da empresa e representa clientes menos valiosos e que estão mais propensos a deixar de comprar ou que estão inativos na plataforma de compra, visto que foram definidos por menor recência, frequência e valor médio de compra. Já o segundo grupo (Cluster 2) é formado por uma pequena parcela dos clientes, que são os mais valiosos para a empresa, pois são compradores frequentes e recentes, com um gasto elevado, representando uma fonte estável para a companhia.

Os dados avaliados, os grupos formados e os *insights* trazidos pelo estudo podem ser usados para direcionar campanhas de *marketing* para cada grupo, visando aumentar o valor de *Lifetime Value* dos clientes e a perpetuidade do negócio. Por exemplo, o primeiro grupo pode ser alvo de campanhas de reativação ou promoções especiais em itens de menor valor agregado para tentar aumentar o engajamento e retomar consumidores com possível *churn*. Entender as características de compra deste grupo é bastante importante, pois estratégias personalizadas serão essenciais para empenhar os consumidores deste grupo na tentativa de torná-los mais valiosos e alterar a proporção entre os grupos vista no estudo.

Já as estratégias para o segundo grupo podem incluir ofertas exclusivas e campanhas de fidelização para estimular os clientes a aumentarem cada vez mais o valor médio gasto, intensificar a quantidade de compras realizadas e melhorar ainda mais o relacionamento com o negócio.

Para trabalhos futuros, sugere-se realizar análises mais detalhadas separando os clientes em mais grupos, mesmo que os indicadores informem que com 2 grupos a separação é melhor. Isto porque os grupos ficaram desbalanceados em relação ao número de clientes e a segmentação resultou em clientes mais engajados e clientes menos engajados, sendo que poderiam ter categorias intermediárias entre esses grupos. Essa redivisão dos grupos, considerando um número maior de clusters pode facilitar a segmentação dos clientes e, consequentemente, favorecer a criação de estratégias ainda mais personalizadas e com maior potencial para atingir os consumidores pretendidos. Outra sugestão é usar outras variáveis dos clientes, além da recência, frequência e valor, para que os grupos formados contenham mais informações e detalhes sobre as características de compra, gerando direcionamentos mais claros para as áreas gerenciais e de marketing.

Referências

- AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. **Electronics**, v. 9, n. 8, 1295, 2020.
- ANITHA, P.; PATIL, M. M. RFM model for customer purchase behavior using K-Means algorithm. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 5, p. 1785-1792, 2019.
- ANNACHHATRE, C.; AUSTIN, T. H.; STAMP, M. Hidden Markov models for malware classification. **Journal of Computer Virology and Hacking Techniques**, v. 11, p. 59-73, 2015.

ANSHARI, M.; ALMUNAWAR, M. N.; LIM, S. A.; AL-MUDIMIGH, A. Customer relationship management and big data enabled: Personalization & customization of services. **Applied Computing and Informatics**, v. 15, n 2. p. 94-101, 2019.

ARTHUR, D.; VASSILVITSKII, S. K-Means++: The Advantages of Careful Seeding. **Conference: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on 114 Discrete Algorithms, SODA 2007**, New Orleans, Louisiana, USA, January 7-9, 2007.

DAS, D.; KAYAL, P.; MAITI, M. A K-means clustering model for analyzing the Bitcoin extreme value returns. **Decision Analytics Journal**, v. 6, p. 100152, 2023.

DEMİRCİOĞLU, H. Ö.; ŞİMŞEK, B.; DİLMEN, S.; ÖZCAN, T. Customer Segmentation using RFM Analysis and Clustering Algorithms in Express Delivery Industry. **Eurasian Business & Economics Journal**, v. 32, p. 47-56, 2023.

ESSAYEM, W.; BACHTIAR, F. A; PRIHARSARI, D. Customer Clustering Based on RFM Features Using K-Means Algorithm. In: **IEEE International Conference on Cybernetics and Computational Intelligence**, Malang, Indonesia, pp. 23-27, 2022.

IKOTUN, A. M.; EZUGWU, A. E.; ABUALIGAH, L.; ABUHAJJA, B.; HEMING, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. **Information Sciences**, v. 622, p. 178-210, 2023.

KASEM, M. S.; HAMADA, M.; TAJ-EDDIN, I. Customer profiling, segmentation, and sales prediction using AI in direct marketing. **Neural Computing and Applications**, v. 36, p. 4995-5005, 2024.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: An Introduction to Cluster Analysis**. New Jersey: John Wiley & Sons, 1990.

MOMTAZ, N. J.; ALIZADEH, S.; VAGHEFI, M. S. A new model for assessment fast food customer behavior case study: An Iranian fast-food restaurant. **British Food Journal**, v. 115, n. 4, p. 601-613, 2013.

NCIR, C. E. B.; HAMZA, A.; BOUAGUE, W. Parallel and scalable Dunn Index for the validation of big data clusters. **Parallel Computing**, v.102, 102751, 2021.

PERDHANA, R. B.; HEIKAL, J. Enhancing Customer Segmentation in Online Transportation Services: A Comprehensive Approach Using K-Means Clustering and RFM Model. **Indonesian Interdisciplinary Journal of Sharia Economics**, v. 7, n. 2, p. 2849-2865, 2024.

SCABURI, A.; OLIVEIRA, M. A.; GENEST, L.; REUBREZ, E. Aplicação da metodologia k-means e do algoritmo de Little para roteirização de veículos no Projeto TIMM. In: **4º Congresso Internacional de Desenvolvimento da Engenharia Industrial**, UNISOCIESC, Joinville, 2017.

WEI, J. T.; LIN, S. Y.; WU, H. H. A review of the application of RFM model. **African Journal of Business Management**, v. 4, n. 19, p. 4199-4206, 2010.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z. H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and information systems**, v. 14, p. 1-37, 2008.

ZOERAM, A. A.; MAZIDI, A. K. A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System. **Iranian Journal of Management Studies**, v. 11, n. 2, p. 351-378, 2018.