UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO ACADÊMICO DO AGRESTE
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA PRODUÇÃO

DÉBORA PEREIRA DE MELO

TOPSIS-CKMEANS: uma nova abordagem para classificação de *clusters* em dados de segmentação de clientes

CARUARU

2025

DÉBORA PEREIRA DE MELO

TOPSIS-CKMEANS: uma nova abordagem para classificação de *clusters* em dados de segmentação de clientes

Dissertação apresentada ao Programa de Pós Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, Centro Acadêmico Do Agreste, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção.

Área de concentração: Otimização e Gestão da Produção

Orientador(a): Lucimário Gois de Oliveira Silva

CARUARU

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Melo, Débora Pereira de.
TOPSIS-CKmeans: uma nova abordagem para classificação de clusters em dados de segmentação de clientes / Débora Pereira de Melo. - Recife, 2025.
109f.: il.

Dissertação (Mestrado)- Universidade Federal de Pernambuco, Centro Acadêmico do Agreste, Programa de Pós-Graduação em Engenharia de Produção, 2025.
Orientação: Lucimário Gois de Oliveira Silva.

1. Mineração de dados; 2. RFM; 3. K-Means; 4. TOPSIS; 5. Clusterização Multicritério Ordinal. I. Silva, Lucimário Gois de Oliveira. II. Título.

UFPE-Biblioteca Central

DÉBORA PEREIRA DE MELO

TOPSIS-CKMEANS: uma nova abordagem para classificação de *clusters* em dados de segmentação de clientes

Dissertação apresentada ao Programa de Pós Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, Centro Acadêmico Do Agreste, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção.

Aprovado em: 21/08/25

BANCA EXAMINADORA

_____

Profº. Dr. Lucimário Gois de Oliveira Silva (Orientador)

Universidade Federal de Pernambuco

_____

Profº. Dr. Lúcio Câmara e Silva (Examinador Interno)

Universidade Federal de Pernambuco

_____

Profº. Dr. Fábio Sandro dos Santos (Examinador Externo)

Universidade Federal do Piauí

A Deus.

AGRADECIMENTOS

Agradeço a Deus primeiramente, pois foi com o apoio e ajuda do Pai celestial que consegui me adentrar ao meio da pós-graduação, o mestrado, me deu forças e blindou minha mente para enfrentar desafios que nunca enfrentei na minha carreira. Confesso que tudo que tenho aprendido tem sido permissão Dele. Quando eu pensava que estava sem nenhuma corda para me segurar, Ele colocava amigos, colegas, professores e familiares para dizer que eu estava em segurança. Tenho notado a sua participação em todo o meu caminhar, sendo Ele o único que preenche o vazio que sentimos muitas vezes em nossa vida, na trajetória de vida, nas melhores e piores fases. Quando me faltava um pai, Ele era o Pai, quando me faltava uma companhia ele era Aquele que preenchia esse espaço, além de ser a esperança de dias melhores. Soli deo Gloria.

À minha mãe que tem me dado muito apoio nessa pesquisa. Ela se assemelha à mulher virtuosa, ao qual seu valor excede o de rubis (Provérbios 31:10). Ela é para mim uma amiga, parceira e irmã da qual posso sempre contar. Tudo quanto fazemos é recíproco. Às vezes, quando me encontrava desesperada por qualquer coisa que me vinha, seja em relação à pesquisa, como em encontrar um título ou tema, seja em relação a minha vida pessoal, ela sempre esteve presente. Agradeço a Deus todos os dias por ela.

Ao meu orientador, Lucimário Gois, que me ajudou nessa trajetória desafiadora, estando, sempre que possível, disponível para tirar dúvidas e me orientar pelo melhor caminho, desde o planejamento do tema até a correções da minha escrita. Sou muito grata por tudo.

Aos meus colegas do mestrado, que me deram um apoio nas disciplinas novas que estavam diante de mim. Confesso que esse desafio de enfrentar novas áreas me impactou muito. Durante as aulas sentia que não tinha condições de lidar com desafios das disciplinas, mas tive colegas que sentiam parte dos meus temores e ajudávamos uns aos outros. Cheguei ao fim da pós, não como um sentimento de alívio, mas feliz que desbravei mais uma etapa da minha vida acadêmica e ganhei experiência nova. Agradeço aos meus colegas Josa, Laura, Flávio, Marcelo, Ednael, Larissa e Victor. Como todo mundo estava no mesmo barco, um apoiou o outro.

Ao pessoal do laboratório do GEEOC, que me deram boas risadas e contação de histórias. Ganhei muitas amizades.

Ao meu pai, *in memorian*, do qual herdei a sua alegria contagiante. Me ensinou que mesmo com o peso da vida, pequenas coisas podem alegrar o dia inteiro e grandes coisas podem transformar uma vida inteira. A felicidade deve ser algo intrínseco ao ser humano desde a sua infância, pois, quando crescer, nenhuma angústia será parte de sua personalidade.

À minha avó, *in memorian*, que partiu em maio de 2024. Ela trouxe para mim grandes ensinamentos, e um deles é que jamais devemos desprezar a história de cada pessoa. Carregamos bagagens que, para uns, podem ser desnecessárias, mas que moldam cada pedaço do nosso ser. Tenho registrado muitas de suas histórias para que um dia eu possa contar às próximas gerações.

Ao meu tio Reginaldo, carinhosamente chamado de tio Regi, por ser sempre engraçado e amoroso.

Ao CAPES pelo auxílio financeiro e contribuição para o desenvolvimento desse estudo.

“O homem que volta ao mesmo rio, nem o rio é o mesmo rio, nem o homem é o mesmo homem”.

Heráclito de Éfeso

RESUMO

A segmentação de clientes é uma área muito difundida no ramo empresarial, sendo muito utilizada no segmento D2C, (Direto para o Consumidor). É uma etapa importante para formação de estratégias de *marketing*, pois ajuda a segmentar clientes que tenham comportamento de compra semelhantes entre si, o que pode melhorar campanhas, direcionar o investimento em anúncios, fortalecer estratégias de empresas de E-*commerces* e elevar a experiência da clientela. Dentro dessa abordagem, o K-Means se destaca entre os métodos de clusterização por sua simplicidade, eficiência e implementação em várias situações. Contudo, apesar do método ser bastante utilizado na literatura, não garante que os *clusters* formados possam identificar os melhores clientes, sendo necessário uma análise *a posteriori*. Por isso, esse trabalho propõe o uso da clusterização multicritério ordinal para segmentação de clientes utilizando a abordagem RFM (Recência, Frequência e Valor Monetário). O modelo propõe uma combinação do método multicritério TOPSIS agregado com um método de agrupamento unidimensional. Essa abordagem garante que as classes formadas mantenham a ordenação original do método TOPSIS. Utilizando dados realísticos, o modelo desenvolvido foi comparado com o método tradicional K-Means, exibindo uma performance superior na separação dos melhores consumidores em relação às dimensões de segmentação. Através dos resultados, também é possível utilizar a nova metodologia para verificar a viabilidade do K-means para segmentação de clientes.

Palavras-chave: Mineração de dados. RFM. K-Means. TOPSIS. Clusterização Multicritério Ordinal.

ABSTRACT

Customer segmentation is a widely recognized area in the business field and is extensively applied in the D2C (Direct-to-Consumer) segment. It represents an important stage in the development of marketing strategies, as it helps group customers with similar purchasing behaviors, thereby improving campaigns, guiding advertising investments, strengthening e-commerce strategies, and enhancing the overall customer experience. Within this context, the K-Means algorithm stands out among clustering methods for its simplicity, efficiency, and applicability across various scenarios. However, although this method is widely adopted in the literature, it does not guarantee that the formed clusters accurately identify the best customers, thus requiring a posterior analysis. For this reason, this study proposes the use of ordinal multicriteria clustering for customer segmentation based on the RFM (Recency, Frequency, and Monetary value) approach. The proposed model combines the TOPSIS multicriteria method with a one-dimensional clustering technique, ensuring that the resulting classes preserve the original ranking established by TOPSIS. Using realistic data, the developed model was compared with the traditional K-Means method, demonstrating superior performance in distinguishing the most valuable customers across the segmentation dimensions. The results also show that the new methodology can be applied to assess the feasibility of using K-Means for customer segmentation.

Keywords: Data Mining. RFM. K-Means. TOPSIS. Multicriteria Ordered Clustering.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A segmentação de clientes, no campo do *marketing* e análise de dados, é essencial para as empresas de pequeno, médio e de grande porte no mercado, sejam atuantes no meio *on-line* ou físico. Traz uma rica gama de informações que podem resolver diversas lacunas em serviços e produtos, isso entendendo e atendendo diferentes consumidores (KUMAR, S. et al., 2025). Mais do que simplesmente vender e acumular clientes ao longo de períodos comerciais, a segmentação possibilita que as empresas e organizações construam abordagens mais estratégicas e personalizadas, solucionando as grandes deficiências no seu sistema de vendas e ganhando espaço nesse ambiente competitivo.

Os E-*commerces*, ambientes comerciais especialmente do segmento D2C (*Direct-to-Customer*), e *marketplaces*, têm crescido de maneira exponencial, impulsionados pelo uso de *sites* de terceiros como Mercado Livre, Amazon e outros. Nesses ambientes, têm se maximizado a experiência dos clientes, especialmente de vendedores em suas centrais de vendas, lhes apresentando várias ferramentas de logística. *Sites* próprios de varejo também têm se expandido como um canal de vendas eficiente e, especialmente a principal fonte de faturamento. Contudo, surge a necessidade de direcionar anúncios para os clientes certos, guiar as campanhas de vendas e principalmente, identificar os consumidores potenciais e otimizar processos produtivos (ABCOMM, 2024; BERALDO, 2024; GS1 BRASIL, 2024; RAMKUMAR et al., 2025).

Durante a pandemia de 2020, no período de isolamento, as pessoas passaram a ficar mais dependentes da *internet* em suas casas, que acarretou no alto crescimento do consumo por produtos *on-line*. Com o aumento de vendas em plataformas pela *internet*, surgiu uma grande complexidade de manusear os dados de vendas. Segmentá-los passou a ser um fundamento para entender quais benefícios se pode tirar com as informações de compra do público-alvo, não só para atrair novos clientes, mas de reter aqueles que já compraram nas lojas, como forma de melhorar a participação no mercado. Esse fator passou a ser até mais importante do que somente ter alta clientela, que se caracterizava por ter pouca experiência de compra e baixa frequência de transações (LING e WEILING, 2025; CHRISTY et al., 2021).

Entende-se que esse tipo de processo incluem melhorias na qualidade do produto até na oferta, mas com um diferencial de se adaptar às expectativas do seu

público-alvo. O conceito não se abarca em apenas vender mais ou extrair melhores propagandas do mercado, mas de elaborar um relacionamento com o consumidor que pode fortalecer sua posição no *ranking* competitivo, de modo significativo e duradouro. Em vez de focarem no *marketing* do seus produtos e na melhor apresentação, as empresas estão mirando em segmentos de mercado e satisfazendo seus consumidores, assim, podendo alavancar seus negócios (KOTLER, 2019).

Para Kotler (2019), existem alguns fatores que podem influenciar o comportamento dos consumidores, esses fatores podem ser:

- Culturais, onde se adequa a cultura, subcultura e classe social;
- Sociais, como grupos de referência, *status* e família;
- Pessoais, como idade, tipo de emprego e questões econômicas pessoais, estilo de vida e até o tipo de personalidade e autoimagem;
- Fatores psicológicos, que envolvem a motivação, percepção sobre determinadas situações, aprendizagem, crenças e atitudes.

As empresas buscam entender os padrões de compra para planejar estratégias aprimoradas de venda e *marketing*. Neste caso, a utilização desses fatores ajuda a nivelar e direcionar o caminho para atrair ainda mais seu público-alvo específico. Assim, esse método não só é um instrumento estratégico, mas uma escolha indispensável para a inovação no mercado atual.

Entre as diversas abordagens efetivas, alguns conceitos e algoritmos são utilizados para que se veja como cada um de seus clientes se comportam no âmbito comercial, o grau de envolvimento com a empresa, o que compram e por que compram determinados serviços e produtos. Além disso, quanto gastam e quantas vezes passam pela instituição comercial vale mais que outras informações demográficas sobre o cliente (HUGHES, 2012).

Entre os diferentes métodos de segmentação, destaca-se o método R (Recência), F (Frequência) e M (Monetário) que utiliza três dimensões geradas a partir do comportamento das compras dos clientes (HANDOJO et al., 2023). Assim, os clientes são segmentados a partir da análise conjunta das três variáveis, considerando, por exemplo, matrizes de classificação que são definidas a partir de intervalos das três variáveis (RAMKUMAR et al., 2025). Considerando que as informações, para gerar variáveis comportamentais de clientes, geralmente são oriundas de bancos de dados da empresa, é necessário o uso de ferramentas para extração do conhecimento acerca do comportamento do consumidor.

Nesse sentido, a Análise de Dados também tem a possibilidade de oferecer *insights* mais objetivos para otimizar o processo de segmentação, reduzir os custos de alguma operação e melhorar a eficiência dela. Outrossim, melhora o engajamento ao buscar compreender os anseios dos clientes e as suas preferências em relação a algum serviço, permitindo a tomada de decisão estratégica e eficiente. Desde a década de 1980, diversas técnicas avançadas de Análise de Dados vêm sendo desenvolvidas e aplicadas, como o *Data Warehouse*, OLAP (Sistema de Processamento Analítico *On-line*), aplicações da Mineração de dados, descoberta de padrões, correlação e etc., passaram a integrar o universo de processamento e pesquisa de dados (HAN et al., 2012). Desse modo, a Análise de Dados oferece um grande apoio no que diz respeito a encontrar anomalias e identificação de problemas em pouco tempo, ajudando em ações de prever comportamentos futuros e avaliar o desempenho de estratégias tomadas.

Em particular, além dessas funções, destacam-se os métodos preditivos, os testes de hipóteses, as inferências paramétricas e não paramétricas, a amostragem, dentre vários outros. Entre essas abordagens, a análise de *clusters* ocupa um papel relevante, pois, por meio do uso de algoritmos relacionados, permite entender o comportamento e funcionamento de dados ao identificar grupos formados com base em similaridades. Essa técnica de aprendizado não supervisionado possibilita capturar da estrutura natural dos dados e revelar classes subjacentes presentes nas informações (TAN et al., 2006; OYEWOLE e THOPIL, 2023). Alguns algoritmos de clusterização podem funcionar de formas variadas, entre eles, o K-Means, um algoritmo multidimensional não supervisionado, o qual trabalha particionando elementos em diferentes agrupamentos (IKOTUN et al., 2023).

No caso da segmentação de clientes, busca-se dividir os clientes em diferentes grupos, em que tais grupos possuam uma relação prioritária de preferência do ponto de vista da empresa. Por exemplo, o *marketing* da empresa pode estar interessado em segmentar os clientes em quatro grupos: platina, ouro, prata e bronze (KUMAR, N., 2025). A partir dessa divisão, o *marketing* planejará diferentes campanhas como objetivo de fidelização, acompanhamento ou criação de valor. Contudo, os métodos de clusterização tradicionais, como o método de partição K-Means, podem não garantir a otimização desses grupos, uma vez que esses métodos focam na criação de grupos homogêneos e não na priorização dos mesmos. Como forma de contornar essa limitação, esse trabalho propõe o uso de um algoritmo de clusterização

multicritério ordinal que tem seu foco na relação ordinal dos *clusters* (DE SMET e GUZMÁN, 2004).

Para Bashir et al. (2023), umas das dificuldades que se tem numa classificação supervisionada são quando os grupos não são conhecidos *a priori*. Para a resolução desse impasse, autores como Boujelben (2017) e Meyer e Olteanu (2013) elaboraram três tipos de seleção: relacional, não relacional ou ordenado, respectivamente. Esse último ainda contêm abordagens de total ou parcialmente ordenados (BOUJELBEN, 2017). Eles também completam que o *cluster* ordenado ajuda na formação de *clusters* ordenados de segmentação e formação de vínculos de prioridade entre subconjuntos de cada *cluster*, sendo bem vantajoso na classificação. Baseado nessa problemática, esse trabalho tem como objetivo o uso de uma abordagem de clusterização ordinal para segmentação de clientes utilizando um método TOPSIS desenvolvido no trabalho de Silva et al. (2024).

1.1 OBJETIVOS

1.1.1 Objetivo geral

Desenvolver um novo modelo para segmentação de clientes com base na análise RFM, através de um método de clusterização multicritério ordinal, fundamentada no método TOPSIS.

1.1.2 Objetivos específicos

- Pré-processamento da base de dados e formação da tabela RFM;
- Implementar computacionalmente o modelo proposto ao conjunto de segmentação RFM;
- Análise comparativa do modelo proposto com o método de clusterização K-Means;
- Análise de performance através de estatísticas descritivas e gráficas dos dois modelos.

1.2 JUSTIFICATIVA

O uso da segmentação RFM nas empresas é uma ferramenta essencial devido a seu frequente uso entre os setores de *marketing*, na identificação de clientes que são importantes e podem alavancar o negócio com estratégias direcionadas, através da sua última compra, frequência de compra e gastos transacionais. Com essa ferramenta, se pode saber quem são os clientes potenciais para que a área de *marketing* construa campanhas de produtos direcionados, ter melhor distribuição dos recursos com o direcionamento para o público certo e, por fim, elaborar novos projetos de reengajamento daqueles consumidores que tem pouca recência de compra, frequência e valor monetário. O trabalho de alcançar melhor engajamento social é facilitado pelo conhecimento mais profundo do seu público e assim construir estratégias mais assertivas (CHRISTY et al., 2021).

Para diversificar ainda mais esse campo de negócios, ultimamente a segmentação RFM tem sido combinada com outros métodos da Mineração de Dados, entre eles a clusterização, visto que, além de identificar e pontuar os clientes, também os seleciona a grupos específicos. O agrupamento em si é diferente da classificação, visto que este último é um aprendizado supervisionado. O agrupamento não possui rótulos e é utilizado para identificar aspectos de similaridades existentes em clientes corporativos (IBRAHIM e TYASNURITA, 2022).

O trabalho de Güçdemir e Selim (2015), por exemplo, orientou uma abordagem de segmentação de clientes utilizando o modelo RFM, que integrava algoritmos de clusterização e decisão multicritério sobre dados de uma empresa. O estudo incluía o ranqueamento dos *clusters*, o qual foi conduzido através da média ponderada dos centroides e, em seguida, por uma análise *a posteriori* dos agrupamentos. No entanto, não trabalhou com o conceito de dados unidimensionais, nem com a proposta de um novo modelo de clusterização multicritério ordinal, havendo apenas usado o método *fuzzy* AHP para determinar a importância dos segmentos. Adicionalmente, focou no desempenho de diferentes algoritmos de clusterização e como poderia facilitar na melhora de estratégias CRM.

Com as poucas pesquisas voltadas à clusterização ordinal, à elaboração de métodos para diferentes problemas reais e atuais e ao desenvolvimento de extensões do mesmo método, torna-se evidente a escassez de trabalhos direcionados à essa área. Em primeiro lugar, destaca-se algumas limitações ao usar o K-Means tradicional, sem a utilização de algum método multicritério: todas as alternativas são agrupadas sem considerar nenhum critério de pesos ou de importância, o algoritmo é sensível a

diferentes escalas das variáveis, em termos de magnitude, e não considera as preferências do decisor. Além disso, o K-Means não diferencia a performance dos grupos de clientes formados diretamente na clusterização, carecendo de uma análise *a posteriori*.

Nesse sentido, o uso de uma abordagem de clusterização ordinal já garantiria previamente a ordenação entre os *clusters*, facilitando a classificação *a posteriori* do analista. O método avalia as ligações prioritárias com base na preferência do decisor, o que é muito importante quando se trata de desempenho alto ou baixo em um conjunto de dados (MELO et al., 2024).

Com a produção de artigos dentro de áreas como Pesquisa Operacional e Gerenciamento, além da sua utilidade nesses setores, a sua pouca produção torna ainda um tanto desafiador no que diz respeito à sua diversidade e aprofundamento, especialmente na área de segmentação de clientes. Dessa forma, o presente trabalho visa contribuir para a literatura ao criar um novo modelo, usar o TOPSIS-Ckmeans em dados de segmentação RFM, e verificar sua performance com um algoritmo tradicional na literatura, o K-Means, julgando o desempenho em termos de eficiência e otimização, implementação, limitação, dentre outros, destacando outras formas eficientes de segmentar clientes em um contexto real e ampliar essa área que ainda se encontra incipiente no meio acadêmico.

1.3 CONTRIBUIÇÕES COM OS OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL DA AGENDA 2030

Este trabalho oferece algumas contribuições que apoiam os Objetivos Estratégicos de Desenvolvimento Sustentável (ODS), aos quais foram estabelecidas pela ONU, Organização das Nações Unidas, para até o ano de 2030. O presente estudo traz algumas soluções que podem auxiliar tanto a indústria e a economia, quanto o mercado, reforçando alguns dos objetivos e metas descritas na Agenda 2030. Destaca-se algumas ODS contempladas (IPEA, 2018):

- A ODS 8 descreve a promoção do crescimento econômico inclusivo e sustentável, o emprego pleno e produtivo e trabalho descente. Nesse sentido, este estudo busca contribuir com o crescimento econômico ao elaborar uma metodologia inserida no campo de Gestão da Produção e Pesquisa Operacional. O objetivo geral do trabalho é construir um modelo de clusterização multicritério ordinal que oferece a

otimização de processos de decisão e amplia o estudo dentro da área de segmentação de clientes, além de poder incentivar o desenvolvimento econômico e, por fim, a melhoria de condições de trabalho.

- A ODS 9 traz a construção de infraestruturas resilientes, promove a industrialização inclusiva e sustentável, e a inovação. Nessa perspectiva, o presente estudo oferece inovação ao ter um direcionamento estratégico de identificação e retenção de melhores perfis de consumo de maneira mais eficiente. Além disso, a metodologia fortalece a otimização de processos de tomada de decisão e o sistema de competividade do mercado, ajudando no desenvolvimento sustentável e na inovação.
- A ODS 12 assegura padrões de produção e de consumo sustentáveis. A pesquisa, por sua vez, pode contribuir para otimização e distribuição de recursos produtivos, diminuição de custos operacionais e computacionais de áreas que envolvem processos de produção e estratégias de *marketing*. Junto a isso, pode ajudar no desenvolvimento de estratégias de gestão sustentável, em termos de orientação de campanhas e planos de ação, o que pode favorecer o consumo consciente. Por fim, permite minimizar também os desperdícios de recursos em modelos de produção e, consequentemente, os impactos sobre o meio ambiente.

Dessa forma, o presente estudo apoia a construção de soluções que tragam impactos positivos, sejam elas de médio à longo prazo, na gestão da produção e, consequentemente ao mercado, indústria, meio ambiente e sociedade. Com essa metodologia, dentro do campo de decisão multicritério e segmentação de clientes, pode atuar de maneira positiva no desenvolvimento de um mercado consciente e mais sustentável.

1.4 ESTRUTURA DO TRABALHO

O trabalho se divide em seis capítulos, cada um com suas subseções e tópicos. O primeiro capítulo aborda a introdução que inclui a apresentação das ideias iniciais e o contexto da pesquisa, incluindo os objetivos que se dividem em objetivos gerais e específicos, os quais servem de norte na composição deste trabalho e, por fim, a justificativa, que é o motivo pelo qual este trabalho está sendo produzido.

O segundo capítulo apresenta o referencial teórico com a apresentação dos métodos e abordagens que foram utilizadas no trabalho, em subtópicos, como as

principais referências temáticas que envolvem a Mineração de Dados, a segmentação de clientes, a análise RFM, a clusterização, que traz os principais pontos desses algoritmos e, especialmente, o K-Means, o subtópico de decisão multicritério que traz uma breve introdução dos seus métodos, que estabelece uma conexão com os métodos de clusterização multicritério ordinal, culminando no modelo principal do trabalho.

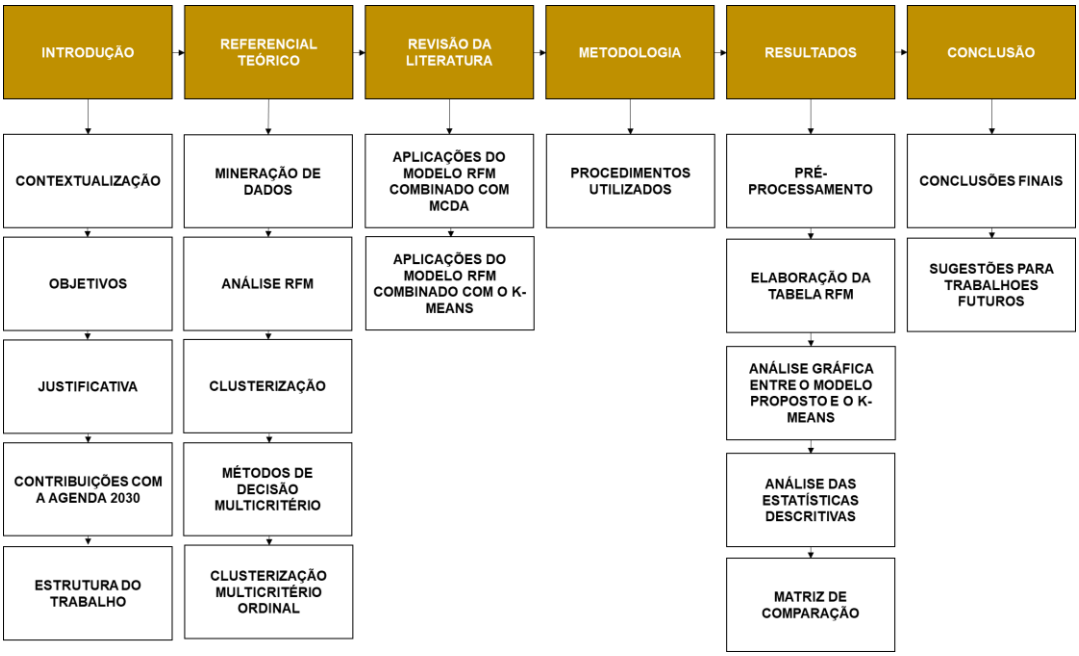
O terceiro capítulo traz a revisão de literatura onde se encontra os principais trabalhos que estão na área de interesse, que embasa o uso de RFM em vários contextos e situações, com trabalhos que tem suas particularidades e forma de uso, importantes para o embasamento dessa pesquisa.

O quarto capítulo traz a metodologia, onde se encontra os principais pontos, desde os métodos utilizados às informações dos dados que foram usados na pesquisa, a abordagem que foi utilizada, alguns trabalhos que serviram para nortear a utilização de cada metodologia e o tipo de procedimento para extração de dados importantes.

O quinto capítulo traz e discute os resultados obtidos ao longo do estudo, desde o tratamento dos dados coletados à comparação dos modelos analisados no trabalho.

O sexto capítulo apresenta as conclusões do trabalho, trazendo novos *insights* relacionados ao tema de pesquisa e sugestões para trabalhos futuros.

Figura 1 - Estrutura do trabalho



Fonte: a autora (2025)

2 REFERENCIAL TEÓRICO

Este capítulo irá tratar os fundamentos teóricos que estruturam este trabalho. Serão abordados os principais conceitos necessários para a realização desse estudo, como Mineração de Dados, segmentação de clientes, enfatizando o método RFM, a clusterização e os métodos de decisão multicritério, os quais servirão de base para os próximos capítulos.

2.1 MINERAÇÃO DE DADOS E SEGMENTAÇÃO DE CLIENTES

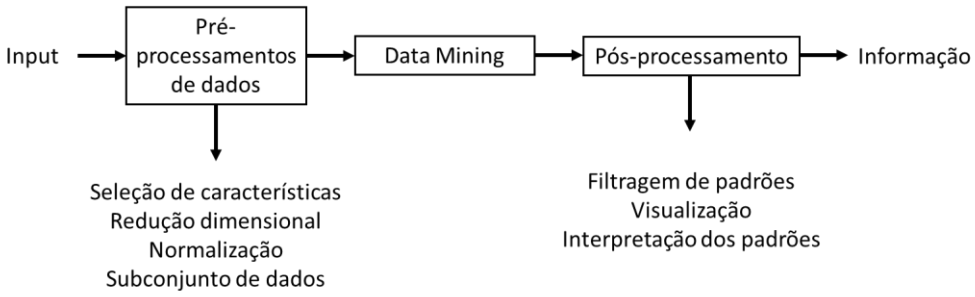
Segundo Tan et al. (2006), Mineração de Dados nada mais é do que o processo de descobrir informação úteis de repositórios de dados. Para FAYYAD et al. (1996), se aplica algum algoritmo para extrair padrões de grandes *datasets*. Essa técnica vasculha enormes bancos de dados e registros, transformando as informações brutas em elementos apropriados para interpretação de resultados. Por conta da necessidade de superar dificuldades no que diz respeito às técnicas tradicionais de análise de dados, alguns pontos foram determinantes para criação da mineração de dados (TAN et al., 2006):

- **Escalabilidade:** devido ao crescimento tecnológico e aumento das informações, a utilização de dimensões maiores de armazenamento de dados tem sido cada vez mais recorrente. *Gigabytes, terabytes* e outras escalas começaram a se popularizar, como por exemplo, tem-se informações biológicas e genéticas, que se utilizam de grandes espaços para alocação de atributos. Nisso, precisa-se de maior capacidade computacional para processar e obter análises detalhadas;
- **Heterogeneidade e complexidade de dados:** com o crescimento de objetos não tradicionais, variados e complexos, tal como dados não estruturados, textos, entre outros, técnicas novas unidas a mineração de dados têm sido bastante usuais em vários setores profissionais. Essas técnicas facilitam descobrir padrões, conexões que vão além de informações tradicionais e mais simples.
- **Propriedade e distribuição do dados:** os dados que são importantes para análise nem sempre estão armazenados em apenas um local, mas são distribuídos em diferentes espaços ou “endereços”. Por isso, é importante o desenvolvimento de técnicas que consolidem os resultados da mineração de dados provenientes de várias fontes. Essa questão está diretamente relacionada à

segurança de dados, especialmente na busca de atributos, onde a segurança deve ser mantida após a análise.

- **Análise não tradicional:** por conta das limitações da estatística tradicional, devido às altas demandas de tempo e custo em relação ao desempenho em experimentos de alta complexidade e ao alto esforço de processamento, tornou-se uma possibilidade inviável do uso de métodos comuns. Tan et al., (2006) ressaltam a necessidade de automatizar o grande reduto de informações, com o uso de mineração de dados para gerar e avaliar hipóteses de forma ampla e eficiente, até com formulações que estão além dos métodos estatísticos tradicionais.

Figura 2 - Processo de obter conhecimento em base de dados



Fonte: Tan et al. (2006). Adaptado.

A transformação de informações cruas é essencial para a interpretação e um bom resultado. Indo por essa linha de raciocínio dos autores, esse processo pode ser estruturado em três etapas importantes: Pré-processamento, Mineração dos dados e Pós-processamento (TAN et al., 2006).

A Mineração de Dados tem apresentado uma significativa evolução na forma de ser obter informações relevantes, com o crescimento da tecnologia ao longo dos tempos. Ele encontra dados que tenham funcionalidade em grandes repositórios de informações, como registros de compra no varejo, passagens, aquisição de equipamentos, identificando padrões e detectando tendências. Conforme Fayyad et al., (1996), essa prática é amplamente usada pelos estatísticos e analistas, como um meio de se extrair conhecimento a partir do processamento de dados, processo este denominado *Knowledge Discovery in Databases* (KDD).

O pré-processamento de dados é a etapa inicial e fundamental para garantir uma qualidade das informações antes das técnicas de mineração. Nele, irá conter alguns filtros para organização. O primeiro, a seleção de características, visa

identificar e manter apenas aquelas variáveis relevantes para a análise. Posteriormente, no segundo filtro, passam por uma redução dimensional, para que se diminua a complexidade do conjunto e não se perca informação prestigiosa com a limpeza de ruído e remoção de duplicados. O terceiro filtro irá normalizar os dados, ajudando nas comparações e melhorando o desempenho para os algoritmos. Após a normalização, os dados são segmentados para diferentes análises. Essa é a parte mais trabalhosa e desgastante no processo de encontrar conhecimento em conjuntos (TAN et al., 2006).

Na Mineração de Dados, os dados serão extraídos, desde os seus padrões até o conhecimento útil deles. Essa etapa propõe a aplicação de técnicas estatísticas, para a detecção de tendências e anomalias. Já no pós-processamento, irá envolver identificação de informações, descarte de valores irrelevantes e, por fim, a contextualização os padrões achados.

Esses tipos de informações são bastantes utilizadas na segmentação de clientes, prática que é essencial para a compreensão de diferentes perfis de consumo. Por meio dessas ferramentas disponíveis da Mineração de Dados, é possível identificar grupos de consumidores e verificar quais são as preferências de cada um. Em empresas, por exemplo, a segmentação pode ocorrer através de um sistema projetado em detectar clientes, como em um modelo de negócios estratégico claro e um mercado específico, onde a empresa agrupa os clientes de acordo com seus atributos, comportamentos, necessidades, preferências, psicologia do consumidor ou outras características para que, em seguida, diversas estratégias de *marketing* para cada grupo de clientes possam ser usadas (LING e WEILING, 2025).

Nesse cenário, o uso de ferramentas em gerir o relacionamento com o cliente tornou-se um ponto indispensável para os negócios. O CRM, Gestão de Relacionamento com o Cliente, é uma das abordagens que integra pessoas, negócios e tecnologia a fim de se entender as necessidades dos consumidores e torná-los ainda mais satisfeitos. Cada vez mais as empresas estão entendendo o valor de estabelecer relacionamentos próximos e duradouros com os clientes para aumentar a retenção. Nisso, o CRM tem uma certa semelhança com a segmentação de clientes, pois é uma das partes básicas que gere a Gestão de Relacionamento com o Cliente. O sistema ajuda as empresas a registrarem as informações dos consumidores, incluindo suas metas e necessidades. Junto a isso, alguns outros benefícios são considerados, como melhorias nos serviços prestados aos clientes e no conhecimento, aumento de

serviços personalizados, segmentação e até a customização do *marketing* (MOHAMMADHOSSEIN e ZAKARIA, 2012; ZHOU et al., 2011).

Quatro dimensões do CRM são listadas: **identificação do cliente**, também chamado de **aquisição de clientes**, é o modo onde a empresa busca seus clientes-alvo e os segmenta com base na organização e estratégia de *marketing*. Isso é significativo, pois enquanto a segmentação tradicional trabalha com a subdivisão de uma base de clientes em grupos menores, essa dimensão organiza características subjacentes dos clientes com a busca de segmentos lucrativos dos mesmos; **atração de clientes**, em que, após identificar o segmento de clientes apropriado, o próximo passo é a organização para atrair o público com a oferta de produtos ou serviços alocando recursos para atender as suas necessidades; **retenção de clientes** é uma parte vital do CRM, pois envolve a satisfação e a busca de fidelização do público-alvo. Finalmente, o **desenvolvimento do cliente**, que consiste no valor de vida útil do cliente (CLV), *upselling* e *crosselling*, e análise de cesta de compras (HOSSENI e TAROKH, 2011).

A partir da adoção da ferramenta CRM, quando se considera melhorias em termos de retenção e valor da clientela, o CLV (*Customer Lifetime Value*), torna-se possível estimar o valor de vida útil do cliente ao longo de um período de tempo. Para que os gestores possam usar o CLV, é preciso considerar três fatores: o valor atual dos clientes, normalmente obtido através dos dados transacionais, valor potencial e rotatividade de clientes. Também é utilizado dados sociodemográficos a fim de entender os segmentos lucrativos da clientela (HOSSENI e TAROKH 2011).

2.1.1 Análise RFM (Recência, Frequência e Valor Monetário)

O que muitos consumidores desejam dos varejos, de modo geral são produtos únicos, serviços especiais, informação, atenção e reconhecimento. Para alcançar os desejos do seu público-alvo, varejistas e empresas de serviços procuram alternativas de aprimorar o atendimento com os seus clientes atuais e abarcar novos consumidores. No entanto, percebeu-se que, embora buscassem várias alternativas de alcance, as empresas entenderam que não bastava apenas construir uma ligação com o seu público, mas também, saber quem são os consumidores de alto valor da sua marca, os potenciais a até aqueles que estão em risco (HANDOJO et al., 2023).

O uso da segmentação de clientes já é antiga, sendo trabalhada desde as décadas de 1970 e 1980, especificamente há mais de 60 anos (HUGHES, 2012). Essa abordagem está inserida na área de *Database Marketing*, utilizada por vários profissionais dos setores de *marketing* e varejo, os quais buscavam por melhores estratégias de negócios, aproximação com o seu público-alvo e identificar os mais lucrativos.

A análise de RFM (*Recency, Frequency, Monetary*) tem sido muito abordado na literatura com o intuito de segmentar clientes, com base na sua atividade de compra na instituição e valores investidos em produtos. Assim, seria possível encontrar quais são os consumidores fiéis, os promissores, os que precisam de atenção e aqueles que estão em risco de perda. O método foi criado por Jan Roelf Bult e Tom Wansbeek (1995) com esse desígnio de segmentar clientes colocando-os em agrupamentos.

Conforme Hughes (2012), o método RFM ou RFV, em português, é um método que categoriza os registros em um banco de dados de clientes para que possa saber quais são os clientes mais recentes, os mais frequentes no ambiente comercial e os que gastam uma grande quantidade de dinheiro em produtos de varejo ou de outros setores. Todas essas informações são retiradas de bancos de informações gravadas de compras.

Identificar os que são mais propensos a serem os mais valiosos, pela atividade e maiores valores gastos na instituição e localizar os clientes que tem pouca participação nas compras, constitui um dos principais objetivos de muitas empresas. Várias vezes o setor de *marketing* cria estratégias, programas de gratificação e estudos para estender a permanência dos clientes valiosos, assim como estratégias de resgate dos potenciais clientes, buscando entender os motivos pelos quais se tem pouca interação com a organização, a fim de fornecer um tratamento personalizado de maneira individual a cada um deles (DOĞAN et al., 2018). Assim, as empresas tentam achar soluções para aumentar o capital e ser a melhor escolha dos seus clientes-alvo. (CHEN et al., 2012).

Além disso, a Análise RFM é citada em trabalhos de uso combinado com outros métodos, como algoritmos de clusterização, especialmente o K-Means (AKANDE et al., 2024), com a intenção de otimizar o sistema de análise de dados para extrair padrões de consumo e informações valiosas de *datasets* muito extensos (LING et al., 2024). O suporte desses algoritmos facilitam a distribuição de *clusters* que podem

indicar o melhor agrupamento de clientes para a empresa, numa análise a *posteriori* de seus resultados, e assim, facilitar o processo de estratégias e tomada de decisão.

Cada letra da sigla RFM representa uma variável diferente relacionada ao comportamento do cliente. O parâmetro M (Valor Monetário) se refere ao gasto total do cliente em produtos durante um período de tempo e, quanto maior esse montante, mais valor a organização dará a aquele cliente. A variável R (Recência) representa o tempo desde a última transação do cliente no ambiente de vendas e, quanto menor os valores dessa variável, mais recentes e melhores são os registros. Quanto ao F (Frequência), representa o número total de transações que o cliente realizou na empresa. Quanto maior o valor da frequência, maior é a visita do cliente à organização (ANITHA e PATIL, 2022).

Uma maneira de segmentar os clientes utilizando essa abordagem RFM é dividir os clientes em quartis ou intervalos em cada uma das variáveis e pontuar esse cliente em uma escala de 1 a 5. Essa classificação permite que clientes sejam segmentados em diferentes perfis de consumo. Com isso, os clientes são alocados em categorias pré-nomeadas, de acordo com a pontuação em cada um dos três elementos R, F e M (COUTINHO, 2022; CHRISTY et al., 2021). Os clientes que receberam a numeração de 5 respondem muito melhor que os que receberam 4 e assim sucessivamente, com o 3, 2 e 1 (HUGHES, 2012).

Quando esse processo de codificação é finalizada nos três parâmetros da análise, cada cliente passa a possuir, em seus registros de banco de dados, três dígitos únicos que indicam o seu desempenho, variando, por exemplo, entre 555, 541, 544, etc. Dessa forma, é classificado cada um dos consumidores de acordo com esses indicadores (HUGHES, 2012).

Essa mesma técnica pode ser trabalhada classificando o cliente com rótulos, de acordo com sua atividade de compra na organização. Algumas companhias segmentam os clientes como *premium,* platina*,* ouro*,* bronze e entre outros (DOĞAN, 2018).

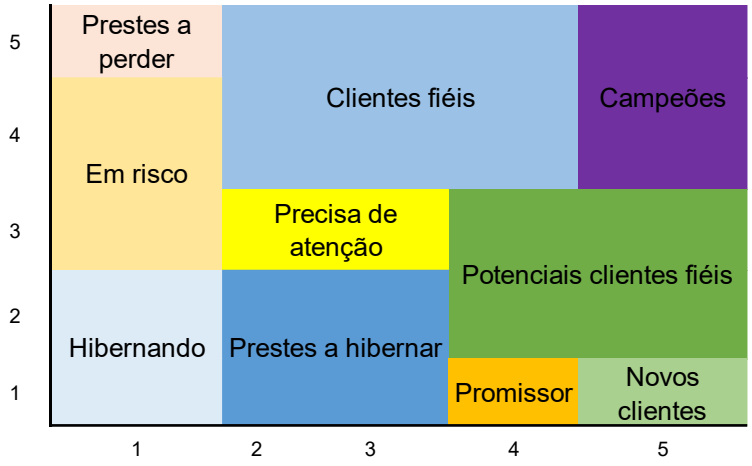
Figura 3 - Pontuações de acordo com a faixa de valores dos indicadores

	Faixas		
Score	Recency	Monetary	Frequency
1	30 - 60	100 - 199	1 - 2
2	21-30	200 - 399	3 - 5
3	15 - 20	400 - 599	6 - 9
4	8 - 14	600 - 799	10 - 14
5	1 - 7	800 - 1000	15 - 20

Fonte: Felix (2022).

De acordo com o exemplo da Figura 3, a recência que estiver entre 1-7 é computado com o *score* 5, indicando as compras mais recentes do cliente, assim como, se a frequência estiver com 1-2 o *score* será de 1, sugerindo quase uma inatividade do cliente na empresa. Em relação ao valor monetário, se estiver entre 200 e 399, o *score* será de 2, mostrando que o cliente não gasta muito com produtos daquela empresa.

Figura 4 - Exemplo de rotulação segmentada de clientes



Fonte: Felix (2022). Adaptado.

Esse modelo pode ser ainda implementado com outras técnicas para aprimoramento da segmentação, como o CLV, falado na seção anterior, que estima o valor total que um cliente pode criar ao longo dos períodos de consumo com aquela empresa, além disso, mede as variações de comportamento do cliente que podem interferir nos ganhos empresariais (YOSEPH et al., 2019; DOMINGUEZ e RUANO, 2024). Além disso, esses tipos de técnicas podem ser complementadas com outras abordagens como métodos de clusterização e análise de séries temporais, o que

amplia a gama de possibilidades de segmentação e até com a previsão comportamental dos consumidores (ABBASIMEHR e SHABANI, 2021).

2.1.2 Extensões e variações do modelo RFM

Além das abordagens tradicionais utilizando as variáveis RFM para a segmentação, também existem diferentes extensões, como pode ser visto na tabela 1. Nessa tabela, é possível visualizar modelos de segmentação do cliente que utilizam combinações de diferentes métodos com o modelo RFM.

Tabela 1 - Variações do Modelo RFM e combinações

Autores	Títulos dos trabalhos	Modelos utilizados	Abordagens
Sundari et al. (2024)	Customer segmentation based on Recency, Frequency, Monetary, Variety and Duration (RFMVD)	RFMVD (*Recency, Frequency, Monetary, Variety, Duration*)	Usaram um método extenso do RFM, adicionando duas mais variáveis combinado com o K-Means em dados de transações
Qi et al. (2023)	F-RFM-Miner: an efficient algorithm for mining fuzzy patterns using the recency-frequency-monetary model	F-RFM-Miner (*Fuzzy RFM Miner*)	Utilizaram o algoritmo F-RFM-Miner para obter padrões *fuzzy*-RFU-*tree* de forma mais eficiente que outros algoritmos
Ozkan e Kocakoc (2021)	A customer segmentation model proposal for retailers: RFM-V	RFM-V (o parâmetro V representa a variabilidade das compras dos clientes)	Formaram um novo modelo de RFM, propondo também uma nova matriz, Matriz Profundidade Cliente-Modelo, com um parâmetro V adicionado na segmentação RFM
Faran et al. (2023)	Combination of RFM's (Recency Frequency Monetary) Method and Agglomerative Ward's	Combinação de RFM e o método Agglomerative Ward	Dados de transações são particionados e agrupados em dois *clusters* com o método Ward. A pesquisa substituiu a procedimento de pontuação

	Method for Donors Segmentation		tradicional do RFM pela análise de *cluster* baseada em RFM
Husnah e Novita (2022)	Clustering of Customer Lifetime Value with Length Recency Frequency and Monetary Model Using Fuzzy C-Means Algorithm	*Fuzzy* C-Means e LFRM (*Length*, *Recency, Frequency, Monetary)*	Determinaram a segmentação de clientes com base no valor de vida útil do cliente. O processamento utilizou o LFM e o Fuzzy C-Means para agrupamento
Gata et al. (2019)	Implementation of Decision Tree Algorithm in Customer Recency, Frequency, Monetary, and Cost Profiling: a Case Study of Plastic Packing Industry	RFM-C (parâmetro C é o custo) e Árvore da Decisão	Modelaram a forma de classificação do perfil do cliente usando o algoritmo C4,5 e o Randon Forest. A intenção era descobrir quais dos dois algoritmos gerariam o melhor modelo de classificação de clientes de acordo com o comportamento de compra dos clientes

Fonte: da autora (2025)

2.2 TÉCNICAS DE CLUSTERIZAÇÃO

Clusterização ou análise de *clusters* é o entendimento na divisão de grupos significativos que compartilham as mesmas características (TAN et al., 2006), ou seja, a partição de elementos em classes, de modo que as informações que estão agrupadas em um mesmo *cluster* tem alta similaridade entre si, porém, mas se diferenciam de modo significativo quando comparado com outros elementos de outros agrupamentos. Tem sido usado para três principais objetivos (JAIN, 2010):

- **Estruturas ocultas**: para obter um melhor entendimento sobre os dados, formular hipóteses, detectar anomalias, identificar características importantes.
- **Agrupamento natural**: para identificar o grau de similaridade entre formas ou organismos.
- **Compressão**: como um método para organizar os dados e resumi-los por meio de protótipos de *clusters*.

É uma técnica não supervisionada, encontrada dentro das áreas como de Aprendizado de Máquina e Análise de Dados. O seu uso é encontrado em vários contextos: nos negócios, na psicologia e medicina, na engenharia, dentre outros. Ao longo das décadas, vários autores foram desenvolvendo algoritmos dos mais variados, alguns sendo mais sensíveis e outros sendo mais complexos.

Em primeiro lugar, os algoritmos de clusterização tem uma classificação de três tipos distintos: algoritmos combinatórios (*Combinatorial algorithms*), modelagem por mistura (*Mixture modeling*) e buscadores de modos (*Mode seeking/seekers*). O primeiro, **algoritmos combinatórios**, um dos conjuntos de algoritmos de clusterização mais populares, trabalham principalmente com dados observados que não tenha uma referência direta a algum modelo probabilístico fundamental. A **modelagem por mistura** assume que os dados são uma amostra independente e identicamente distribuída (i.i.d) de alguma população descrita por uma função de densidade de probabilidade. Por fim, **os buscadores de modos**, ou também chamados de "***bump hunters***"*,* que levam em consideração uma perspectiva não paramétrica, que identificam esses picos diferentes da função densidade de probabilidade. As observações mais relevantes ou "mais próximas" a cada modo de perspectiva é o que define os *clusters* individuais (HASTIE et al., 2009).

Existem dois tipos de aprendizado de máquina, aprendizado supervisionado, em que já se tem o estabelecimento de rótulos, baseados em informações pré-estabelecidas; e aprendizado não supervisionado. É nesse grupo que os algoritmos de clusterização são encontrados. É conhecido dessa forma pois as informações dos rótulo dessas classes não são conhecidos ou presentes (HAN et al., 2012).

Para utilizar com a clusterização em Mineração de Dados, Han et al. (2012) trazem alguns requisitos e tipos de algoritmos com suas respectivas características. Em termos de requisito pode-se destacar:

- **Escalabilidade:** sabe-se que muitos algoritmos direcionados a análise de *clusters* podem lidar muito bem com conjuntos pequenos. Neste caso, o tempo de execução é aceitável, a exigência de recursos computacionais é menor e os dados podem ser visualizados e melhor entendidos. Da mesma forma, pode ser eficiente em banco de dados extremamente extensos como um banco que contém dados genéticos de um único ser humano, exigindo um esforço maior de processamento. Quando se trata de um agrupamento de uma pequena amostra desses conjuntos, percebe-se que a pesquisa se direciona, intencionalmente, a um caminho enviesado, atingindo

resultados altamente tendenciosos, além dos *clusters* serem mal definidos. Para se evitar esse tipo de ação, é importante a utilização de algoritmos que lidem mais com altas escalabilidades de informações, de modo que não se perca desempenho ou precisão;

- **Capacidade de lidar com diferentes tipos de atributos:** essa disposição refere-se a se adequar a vários tipos de informações, sejam eles nominais, binários, ordinais ou um conjunto desses tipos de dados. Com o crescimento de vários bancos de dados com diferentes tipos de critérios, tem-se a necessidade de técnicas de clusterização para, por exemplo, gráficos, imagens e outros tipos de dados complexos;
- **Interpretabilidade e usabilidade:** o entendimento sobre os agrupamentos, para os usuários, deve ser claro de fácil entendimento, além disso, úteis. Todo o resultado que tiver deve estar alinhado com os objetivos propostos. Quando os agrupamentos não refletem nesses princípios, não tem valor prático.
- **Capacidade de lidar com dados ruidosos:** a maioria dos conjuntos de informações contêm dados ruidosos, ou seja, que tem valores discrepantes, os chamados *outliers*. Isso de certo modo, dificulta a leitura pois se tornam imprecisas ou erradas, por conta dos mecanismos de detecção e devido a interferências de elementos que circundam os dados. Os algoritmos que realizam o processo da clusterização são bastante sensíveis a essas informações ruidosas, como o K-médias, também chamado de K-Means, produzindo um resultado de baixa qualidade. Sendo assim, a necessidade de se ter métodos de análise de *cluster* robustos a esse tipo de problemática são importantes.

Dentre os diversos métodos de clusterização na literatura, é possível reconhecer as várias classificações desse campo. Para sintetizar essa vasta área, Han et al. (2012) resumiram a alguns grupos de métodos: métodos de partição, métodos hierárquicos, métodos baseados em densidade e métodos baseados em grade.

Os métodos de partição, em síntese, são baseados na distância. Dado um número k de partições, o método cria subdivisões com os elementos disponíveis. Em seguida, utiliza um método para realocação iterativa a fim de melhorar o particionamento movendo os elementos de um agrupamento a outro. Essas movimentações só param quando esses elementos do mesmo *cluster* estejam próximos entre si, enquanto os outros pontos de outros *clusters* estejam distantes ou

diferentes. Esses métodos podem ser um pouco complicados no que diz respeito a otimização global, pois exige uma enumeração de todas as partições possíveis. Neste caso, várias outras aplicações aplicam métodos heurísticos que aperfeiçoam a qualidade dos grupos em alcançar otimalidade, como K-médias ou K-medoids.

Os métodos hierárquicos são outros grupos de métodos de clusterização que cria uma espécie de “árvore”, também chamada de dendrograma, para mostrar em como os dados podem ser organizados, nesta condição em diferentes níveis. Podem ser classificados como aglomerativo ou divisivo. O método aglomerativo, também chamado de *botton-up*, começa com cada elemento formando um *cluster* separado e individual. Consequentemente, esses grupos começam a obter outros objetos próximos, unindo-os em *clusters* maiores.

O método divisivo, conhecido também como *top-down*, se caracteriza em começar com os elementos em um *cluster,* e a cada iteração, esse agrupamento é dividido em *clusters* menores, até chegar numa condição em que todos os pontos estejam fixos em seus respectivos grupos. Além dessas divisões, os métodos hierárquicos se baseiam em distância, densidade e continuidade. Uma característica importante é que, devido a sua rigidez que facilitam e diminuem os custos computacionais, no término de uma etapa, não poderá ser possível desfazê-la. Apesar de não corrigir decisões equivocadas, há métodos que estão sendo aprimorados para aperfeiçoar a qualidade da clusterização.

Em relação aos métodos baseados em densidade, o representante mais conhecido destes métodos é o método DBSCAN. Sua ideia geral é desenvolver um determinado *cluster* enquanto a densidade, o que os autores definem de “números de pontos de dados”, superar um limite. A densidade dos elementos é apurada de acordo com uma região em volta de cada ponto. A vizinhança de uma região específica precisa conter um número mínimo de elementos. Esses métodos são eficientes no que diz respeito a filtragem de *outliers* e valores diferentes, descobrindo *clusters* de outros formatos. Cabe destacar que esses métodos operam com *clusters* exclusivos e não são adequados para ocasiões que consideram agrupamentos difusos (*fuzzy*).

Finalmente, os métodos baseados em grade, que medem o espaço do elemento em um espaço finito de objetos, formando uma estrutura de grade. Um ponto positivo que se deve tirar desses métodos é o tempo de processamento, independentemente do número de dados, mas dependendo do número de células em cada uma das dimensões que compõem a grade. O agrupamento é realizado

diretamente nessa grade e não nos elementos individualmente. Outra característica interessante é que esse método pode ser combinado com outras abordagens, quando exigido.

Na tabela abaixo, apresenta os principais algoritmos para cada tipo de método de clusterização:

Tabela 2 - Listas dos tipos de métodos de clusterização e seus respectivos algoritmos

Métodos de clusterização	**Algoritmos**
Métodos de partição	K-Means, K-modes (variante do K-Means), K-medoids como CLARA, PAM, CLARANS
Métodos hierárquicos	BIRCH, Chameleon, AGNES, DIANA, algoritmo hierárquico probabilístico
Métodos baseados na densidade	DBSCAN, OPTICS, DENCLUE, Mean Shift
Métodos baseados em grade	STING, CLIQUE

Fonte: Han et al. (2012). Adaptado

Além desses diversos métodos de agrupamento, existem ainda outros conjuntos de abordagens, como os métodos baseados em modelos, em redes neurais, em lógica *fuzzy*, em kernel, métodos grafos e computação evolucionária. Contudo, os quatro apresentados na Tabela 2 são os mais utilizados, especialmente os métodos particionais e hierárquicos (CASSIANO, 2014).

2.2.1 K-Means e variações

O algoritmo foi proposto oficialmente em 1956, por Hugo Steinhaus, estabelecendo parâmetros para o algoritmo, com a finalidade que ele pudesse particionar um conjunto de informações de maneira eficaz. Posteriormente, outros autores foram desenvolvendo o algoritmo até o seu estágio final, desenvolvido por Hartigan e Wong (1979).

Supondo que existe um conjunto de dados D, contendo n objetos em um determinado espaço Euclidiano, o K-Means distribui esses elementos de D em k *clusters*, contendo os objetos $C_i = C_1, C_2, \dots, C_k$, onde $C_i \subset D$ e $C_i \cap C_j = \emptyset$, para $(1 \leq i, j \leq k)$. A função objetivo nessa relação é utilizada para avaliar a qualidade da

partição de forma que os objetos dentro de um *cluster* sejam similares entre si, porém diferentes quando comparados com outros *clusters*. Uma partição baseada em centroides usa o centroide de um *cluster* para representar esse agrupamento (HAN et al., 2012).

Os centroides são colocados aletoriamente para que então, as distâncias deles sejam calculadas, assimilando cada elemento a um centro mais próximo. Com isso, o algoritmo classifica os centros utilizando a distância dos pontos, refazendo o processo repetidas vezes até os centroides não se movimentarem mais. O centroide pode ser definido de duas maneiras, pela média ou medoide dos pontos. Para diferenciar um objeto $p \subset C_i$ e c_i, que é um centroide do *cluster*, é calculada a sua diferença em relação ao centroide $dist\,(p, c_i)$, onde $dist(x, y)$ representa a distância Euclidiana entre dois pontos x e y. A qualidade do *cluster* C_i pode ser dada através da variação *intracluster (within-cluster)*, que representa a soma dos erros quadrados entre todos os objetos que estão em C_i e o centroide c_i. O cálculo é definido como (Ibidem et al., 2012):

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \qquad (1)$$

Em que:

- E representa a soma dos erros quadráticos de todos os objetos do conjunto;
- p representa o ponto em um espaço que representa um dado elemento;
- c_i é o centroide do *cluster* C_i;

As principais vantagens do algoritmo incluem, primeiramente, simplicidade, facilidade de implementação e versatilidade, em que quase todos os aspectos do método (inicialização, função de distância, critério de término, etc.) podem ser modificados. Isso é confirmado com várias extensões em diversos trabalhos ao longo dos vários anos de utilização do algoritmo (CELEBI et al., 2013).

Por outro lado, o algoritmo também apresenta algumas desvantagens: precisa-se selecionar os centroides inicialmente para cada uns dos *clusters* e, além disso, especificar a quantidade de *clusters* k desejada para obter os resultados requeridos.

Ademais, o algoritmo é bastante sensível a anomalias (*outliers*), dados estes que fogem da escala. A respeito da inicialização, o algoritmo também é demasiadamente sensível. Os efeitos adversos na inicialização inadequada incluem *clusters* vazios, lenta convergência e uma alta chance de ficar preso em um péssimo mínimos locais (CELEBI, 2011). Por isso, alguns autores criaram algumas variações do algoritmo com a intenção de minimizar alguns dos problemas, a saber, o K-medoids, que é mais robusto a dados com *outliers* e em aspecto de tempo de execução e o Bisecting K-means em termos de grandes base de dados, sendo mais eficiente em evitar a formação de *clusters* vazios (ARORA et al., 2016; MAHMUD et al., 2018).

Sistematizando o processo (MAJILYA et al., 2024):

Entrada:

- k é o número de *clusters*;
- D é o conjunto de dados que contém n objetos.

Saída:

- Um grupo de k *clusters* onde o centro de cada agrupamento é retratado pelo valor médio dos objetos dentro do *cluster*.

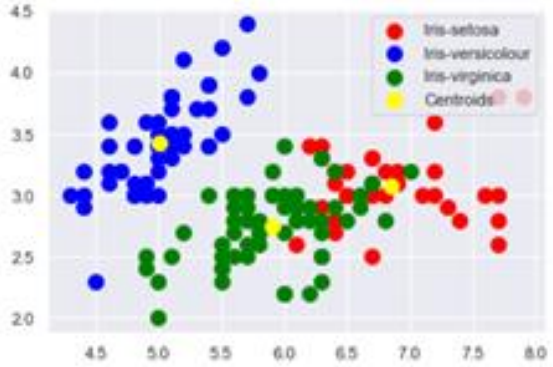
Passos:

- Escolher aletoriamente k objetos do conjunto D como centros iniciais do *cluster*;

Repetir

- Calcular a distância de todos os objetos de todos os k-centroides
- Atribuir cada objeto ao seu centroide mais próximo
- Atualizar os centros dos *cluster* calculando a média de todos os objetos atribuídos a ele, **até que os centroides dos *clusters* não mudem**.

Figura 5 - Exemplo de clusterização com K-Means utilizando o *dataset* Iris



Fonte: Bandgar (2021)

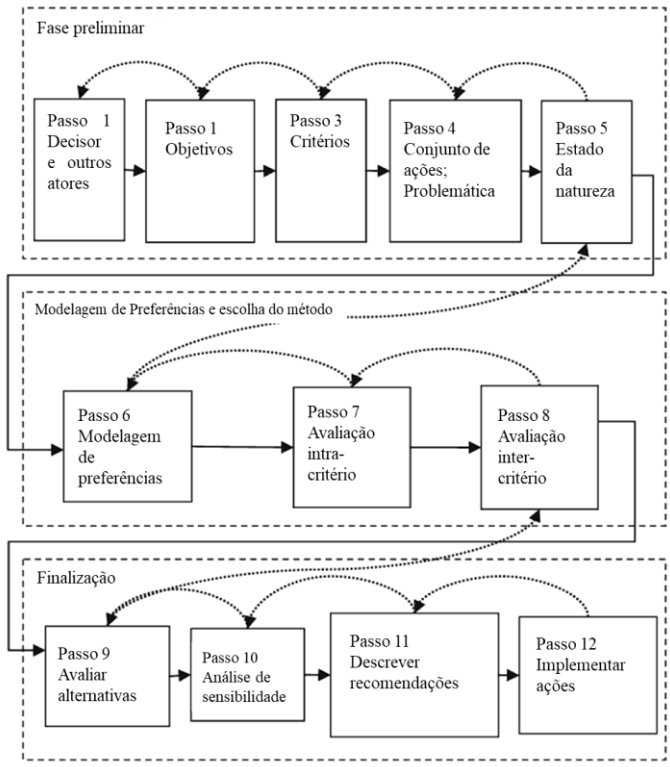
Além do algoritmo ter essa sensibilidade, a quantidade de *clusters* tem que ser escolhida, mas não de forma aleatória. É preciso utilizar outros métodos para que os resultados não se modifiquem com a quantidade de grupos. Para saber o número ideal de agrupamentos k, é necessário aplicar no modelo alguns métodos, entre eles os mais comuns, o método do Cotovelo (*Elbow method*) e o método da Silhueta (*Silhouette Score*), para definir a quantidade ótima de *clusters* (JUANITA e CAHYONO, 2024).

2.3 MÉTODOS DE DECISÃO MULTICRITÉRIO

Os primeiros usos estruturados dos métodos de decisão multicritério foram por volta da década de 1960 em diante com a elaboração do método ELECTRE por Roy (1968) e, posteriormente, o método AHP por Saaty (1980). Porém, antes da elaboração mais robusta dos métodos, Benjamin Franklin já tinha feito abordagens de problemas de decisão e como essas dificuldades poderiam ser sanadas através de uma visualização de prós e contras, onde aplicava uma análise ponderada das alternativas, isso em sua carta de 1772 (FRANKLIN, 1956). Posteriormente, Churchman et al. (1957) trataram em seu livro uma abordagem que envolve ponderação de critérios e fatores multiobjetivos. Embora ainda não existisse o termo MCDM (*Multiple-Criteria Decision Making*), eles consideraram a ideia de atribuição de pesos em problemas decisórios, dando início a um marco importante nos métodos de apoio à decisão.

De modo geral, um problema de decisão pode ocorrer quando há pelo menos duas ou mais alternativas para escolha e vários objetivos relevantes, como, por exemplo, pode acontecer em uma seleção de fornecedores. Nesses casos, com esses objetivos estabelecidos, torna-se possível direcionar o processo de escolher uma alternativa mais adequada, pois eles associam a variáveis que podem guiar o decisor (ALMEIDA, 2013).

Figura 6 - Etapas de resolução de um problema MCDM



Fonte: Almeida et al. (2015). Adaptado

Segundo Almeida (2013) todo método multicritério baseia-se em três elementos: o conjunto de alternativas, de critérios e matriz de decisão, que alinha as avaliações das alternativas em relação aos critérios estabelecidos. A primeira fase para a resolução de um problema está em caracterizar quais são os atores que fazem parte do processo decisório, quais são os objetivos e os critérios que irão definir o rumo da problemática. Uma vez estabelecidos os elementos principais, a próxima etapa será na modelagem de preferências, que conduz à análise de desempenho, resolução do problema e, por fim, à implementação da decisão.

Os métodos foram projetados para indicar uma alternativa mais adequada, classificar alternativas em um número reduzido de categorias e ranqueá-las em uma ordem de preferência. A maioria dos métodos MCDM lidam com alternativas discretas, as quais são descritas por um conjunto de critérios. Alguns problemas são encontrados no processo de decisão multicritério. Neste caso, a escolha do procedimento de agregação para resolver o problema de decisão. No entanto ao longo do tempo, os analistas da decisão, que trabalham com múltiplos critérios, trouxeram

e ainda trazem vários procedimentos de agregação, como modelos formais e metodologias de avaliação (MARDANI et al., 2015).

Segundo Roy (1996) existem três problemáticas que são encontradas no processo de decisão: a problemática de escolha, de ordenação e classificação. A problemática de escolha, a mais tradicional, traz o problema em termos de escolher a melhor alternativa dentre várias outras consideráveis em um conjunto de alternativas. Já a problemática de ordenação se refere a colocar um conjunto de ações em categorias que as definem como, por exemplo, verdadeiro, bom, potencialmente ruim ou falso. As alternativas são organizadas em uma ordem de preferência, da melhor entre elas à pior. Por último, a problemática de classificação diz respeito a comparar as alternativas para que elas sejam agrupadas em classes predefinidas que podem ser ordenadas de acordo com o desempenho. Adicionalmente, a classificação permite uma ordenação parcial ou completa das alternativas. Dependendo do tipo de problemática, é atribuído um método multicritério que seja coerente com o problema de decisão.

Esses métodos podem ser classificados em duas categorias de racionalidades do decisor: compensatórios e não compensatórios. Os métodos compensatórios se descrevem como métodos que compensam um dado critério de uma alternativa de baixo desempenho utilizando um outro critério de desempenho maior, já os não compensatórios não é permitida essa compensação. Um critério com baixo desempenho não poderia ser “ajudado” por um outro de valor maior, ou seja, não há *trade-offs* entre eles. Cada critério é tratado de forma independente da sua pontuação (ALMEIDA, 2013).

A maioria dos estudos na literatura são voltados para áreas importantes, com aplicações voltadas para Pesquisa Operacional e áreas da Computação, Energia, Meio Ambiente e Sustentabilidade, e Sistemas de Manufatura. Grande parte dessas aplicações se utilizam do método AHP (*Analytic Hierarchy Process*), seguido por métodos de decisão híbridos, métodos de agregação e o método TOPSIS. Na literatura, destaca-se outros métodos amplamente discutidos, como PROMETHEE, ELECTRE, DEMATEL, VIKOR, entre tantos outros (MARDANI et al., 2015).

Conforme Almeida et al. (2015), os métodos são classificados em três categorias distintas: métodos de critério único de síntese, sobreclassificação e métodos interativos. Os métodos de critério único de síntese se baseia em um processo de combinação analítica de todos os critérios para produzir uma avaliação

global. Por essa razão, são considerados métodos que possuem um único critério que resume todos os critérios originais. Dentro desse tipo, há os métodos aditivos que são base para vários métodos determinísticos aditivos, tal qual o AHP, SMARTS, MACBETH, o MAUT, que também está incluído nesse grupo, dentre outros. Eles utilizam a estrutura de preferências (P,I).

Os métodos de sobreclassificação, pelo contrário, não se utilizam de um critério único de síntese, mas uma pré-ordenação parcial das alternativas. Muitos desses métodos produzem uma recomendação final sem precisar de uma ordenação total de todas as instâncias. Dentro desse conjunto, há os métodos da família ELECTRE e a família PROMETHEE.

Com relação aos métodos interativos, eles estão associados a contextos que contenham critérios discretos quanto também a critérios contínuos, apesar do principal caso de aplicação desse conjunto de métodos envolva os Problemas de Programação Multiobjetivo ou MOLP. Esses métodos envolvem interação direta com os tomadores de decisão durante esses processos. Incluem métodos de desagregação que consistem em absorver informações diretamente do decisor sobre a avaliação global de algumas alternativas, para inferir os parâmetros de um modelo de agregação, o qual será usado para classificar as alternativas restantes.

Ainda existem outros métodos que fazem parte dos métodos MDCM, que são o TOPSIS, que é baseado no Ponto Ideal e Nadir, modelos aditivos com veto, método lexicográfico, método de Borda, método de Condorcet, dentre vários outros. Outrossim, há autores que desenvolvem modelos e extensões de métodos de decisão multicritério, combinados com outras abordagens, a fim de estudar diferentes problemáticas sobre um conjunto de dados, ou ainda, questões relacionadas tanto ao próprio método de decisão, como por exemplo problemas relacionados à reversão de ordem, a exemplo do caso do PROMETHEE e do TOPSIS (ALMEIDA, 2013).

2.3.1 O método TOPSIS

O TOPSIS (*Technique for Order Performance by Similarity to Ideal Solution*) é um método multicritério que é baseado no conceito de que a alternativa ideal escolhida deve ter a menor distância da solução ideal e a maior distância da solução anti-ideal ou pior solução (HWANG e YOON, 1981; ALMEIDA, 2013). O método permite *trade-off*s entre os critérios, no qual um critério que tem um péssimo desempenho pode ser

compensado com um critério de alto desempenho, fazendo parte dos métodos compensatórios. Existem outras variações do método, como o uso de lógica *fuzzy* (SUN e WANG, 2025; CHEN et al., 2025), DBSCAN modificado (LI e ZHANG, 2021); *Fuzzy* C-means (SWINDIARTO et al., 2018), etc.

Apresenta uma grande diversidade de aplicações devido a sua facilidade nas etapas de desenvolvimento, o que faz ser um dos métodos que tem flexibilidade na formação de variações e extensões, principalmente na ordenação de alternativas a sensibilidade nos pesos dos critérios, que podem ser ajustados de acordo com a importância. Contudo, não consegue lidar com decisões sob incertezas (JUNIOR e CARPINETTI, 2013; ROCHA et al., 2024). Por ser um métodos globalmente usado, ele é uma abordagem que traz confiabilidade nos resultados e no processo (ROCHA et al., 2024). Além disso, com uma abordagem robusta e intuitiva, é bastante aceito e implementado em problemas reais que incluem negócios (SHARMA, H. et al., 2024), psicologia (TABATABAEI, 2024), análise de solos (GÜRBÜZ, 2025), indústria têxtil (SITHI et al., 2025), entre outros ramos.

O método aponta outras vantagens: possui uma lógica sólida que representa a razão da escolha humana; gera um valor escalar que considera as melhores e as piores alternativa simultaneamente; possui uma simples elaboração que pode ser facilmente implementado em alguma planilha e permite que as medidas de desempenho de todas as alternativas em relação aos atributos possam ser visualizadas em um poliedro, pelo menos em duas dimensões (SHIH et al., 2007).

As desvantagens incluem o fato de que a ponderação se trata de um processo difícil e incerto; em alguns casos não é possível determinar os dados incertos com precisão devido à possibilidade de julgamentos humanos vazios, especialmente quando as informações apresentadas não são suficientes; outro ponto importante é que o método tem um problema chamado reversão de ordem. Esse fenômeno acontece por causa das mudanças no ranqueamento das alternativas pela adição de uma alternativa, critério ou a remoção de um deles. Por fim, o método utiliza a Distância Euclidiana sem estar considerando a correlação dos atributos, o que pode impactar os resultados na sobreposição de informações (MADANCHIAN e TAHERDOOST, 2023).

O TOPSIS segue 6 etapas (PANDEY e KOMAL & DINCER, 2023):

Passo 1: Deve-se criar uma matriz A de avaliação, consistindo de alternativas $\{A_1, A_2, \dots, A_m\}$ e critérios $\{C_1, C_2, \dots, C_n\}$, posteriormente aplicar a normalização na matriz de avaliação:

$$A = [a_{ij}]_{m \times n} \tag{2}$$

Em que:

- a_{ij} é a informação extraída da alternativa A_i do critério C_j da matriz de avaliação;
- $i = 1,2, \dots, m$ é o número de alternativas;
- $j = 1,2, \dots, n$ é o número de critérios.

Passo 2: Calcular a matriz de decisão já normalizada e ponderada $V = [v_{ij}]$, onde $v_{ij} = w_j r_{ij}$. Cada critério tem o vetor de peso $\{w_1, w_2, \dots, w_n\}$ que representam os níveis de importância e satisfaz a condição $\{w_j \in [0,1],\ j = 1, 2, \dots, n$ e $\sum_{j=1}^{n} w_j = 1$, os quais são definidos pelo decisor. A matriz normalizada é denotada por $R = [r_{ij}]_{m \times n}$, onde:

$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} a_{ij}^2}}, \qquad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n; \tag{3}$$

Em que:

- w_j representa o peso dos critérios j.
- r_{ij} representa o valor normalizado da alternativa i no critério j;
- v_{ij} representa o valor da alternativa i no critério j normalizado e ponderado;

Passo 3: Determinar a melhor (V^+) e a pior alternativa ideal (V^-):

$$V^+ = \left[\left(\max v_{ij}; j \in J\right), \left(\min v_{ij}; j \in J'\right), i = 1, 2, \dots, m\right] = \{v_1^+, v_2^+, \dots, v_n^+\} \tag{4}$$

$$V^- = \left[\left(\min v_{ij}; j \in J\right), \left(\max v_{ij}; j \in J'\right), i = 1, 2, \dots, m\right] = \{v_1^-, v_2^-, \dots, v_n^-\} \tag{5}$$

Em que:

- V^+ e V^- são os melhores e menos favoráveis escolhas, respectivamente, enquanto J e J' são os conjuntos de critérios de tipo benefício e custo, respectivamente.

Passo 4: Calcular as distâncias euclidianas do PIS, que representa a melhor alternativa ideal e do NIS, que representa a pior alternativa ideal, para se obter as pontuações finais de cada alternativa:

$$S_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2}, \qquad i = 1, 2, \dots, m \tag{6}$$

$$S_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2}, \qquad i = 1, 2, \dots, m \tag{7}$$

Em que:

- S_i^+ é a distância euclidiana em relação à solução ideal positiva;
- S_i^- é a distância euclidiana em relação à solução ideal negativa;

Passo 5: Calcular a similaridade de acordo com a fórmula abaixo para cada alternativa a fim de se obter os *scores*:

$$CC_i = \frac{S_i^-}{S_i^+ + S_i^-}, \qquad i = 1, 2, \dots, m \tag{8}$$

Em que:

- $CC_i \in [0,1]$ representa o índice de similaridade da alternativa i em relação à solução ideal. Quanto mais próximo de 1, melhor a alternativa;

Passo 6: Por fim, ordenar todas as alternativas de acordo com os *scores* do melhor para o pior desempenho.

2.4 CLUSTERIZAÇÃO MULTICRITÉRIO ORDINAL

O primeiro estudo sobre os *clusters* alinhados a uma perspectiva multicritério foram De Smet e Gilbart (2001) com a utilização de um problema real de avaliação de risco país, onde empregaram um modelo de otimização para se obter agrupamentos ordenados, com o uso do método PROMETHEE sobre dados de vinte países. Onze critérios foram colocados baseados em riscos econômicos, financeiros e políticos (MELO et al., 2024; DE SMET e GUZMÁN, 2004; DE SMET, 2014). Em seguida, o estudo com De Smet e Guzmán (2004) foi estendido para tratar de problemas que envolviam o agrupamento de elementos em *clusters*. Foram os primeiros a tratar da Clusterização Multicritério Ordinal, pois em seu trabalho, desenvolveram a utilização de uma extensão com multicritério e clusterização para o aprofundamento do estudo de categorias multicritério. Os centroides são selecionados aletoriamente e as alternativas são alocadas aos *clusters* que tem correlação mais próxima desse elemento da vez. Com isso, são particionados esses elementos em classes. Os autores consideram que as alternativas só são semelhantes caso sejam preferíveis, indiferentes e incomparáveis mais ou menos às mesmas alternativas (DE SMET e GUZMÁN, 2004).

Outras variações do modelo foram aplicados posteriormente por outros autores, em diversas áreas, industrial, agrícola, empreendedora, medicina, com diversos tipos de métodos de decisão multicritério, como o TOPSIS (OMURBEK et al., 2021; SWINDIARTO et al., 2018; ERNAWATI et al., 2021); SMARTS (MARSILI e BÖDEFELD, 2021); ELECTRE (FERNANDEZ et al., 2010; ROCHA e DIAS, 2013), PROMETHEE (ROSENFELD e DE SMET, 2019; LIU et al., 2020) e vários outros catalogados na literatura, unidos a algoritmos diversos de clusterização como C-means, K-Means, K-medoids, P2*CLUSTER* e etc. (MELO et al., 2024), entre outras diversas abordagens.

Na pesquisa feita por Melo et al. (2024), o TOPSIS é o segundo mais usado entre a clusterização ordinal, com a primeira aplicação em 2018, com mais oito trabalhos até o ano de 2024. Nessa revisão, foi descoberto que, a vantagem de se usar clusterização multicritério permite a formação de *clusters* ordenados e ligações prioritárias entre subconjuntos, trazendo um adição importante ao processo de ordenação.

Em termos gerais, existem dois problemas de análise dos dados. Quando os grupos não são conhecidos *a priori*, é preciso identificá-los e extrair informações a partir das características contidas neles. Diante disso, trata-se de um problema de clusterização e os grupos são chamados de *clusters*. Por outro lado, uma vez que os conjuntos são definidos *a priori*, o problema de atribuir um elemento a um deles é denominado problema de classificação e os grupos são chamados de classes (NEMERY e DE SMET, 2005).

Outros trabalhos de De Smet foram publicados, incluindo principalmente o PROMETHEE combinado com clusterização, entre eles o artigo de De Smet et al. (2012), onde construíram um modelo de clusterização ordinal para buscar a partição ordenada em um banco de dados do Índice de Desenvolvimento Humano (IDH). Eles abordaram um problema de reagrupamento com base em categorias ordenadas de acordo com os níveis de preferência. Utilizaram PROMETHEE e um algoritmo de partições de K-ordenados proposto e em seu resultados obtiveram consistência entre a partição ordenada e o *ranking* IDH.

Segundo Melo et al. (2024) a pesquisa de clusterização ordinal só começou a ser mais explorada nos períodos de 2013, 2014 e 2021. Os trabalhos de De Smet e suas extensões com o método PROMETHEE abriram caminhos para a pesquisa mais detalhada do tema. Contudo, essa área ainda carece de informações mais precisas e variadas. Outro autor que trabalhou com a temática foi Boujelben (2017) com a temática da análise de princípios do PROMETHEE em *clusters* ordenados. O objetivo principal do seu trabalho era utilizar essa extensão PROMETHEE e um algoritmo de clusterização que considerasse cada critério de cada tipo de forma independente e aplicar medidas, a fim de analisar cada partição nos diferentes critérios. Sua pesquisa se baseou no método de De Smet et al. (2012).

Em termos de revisões de literatura, a sua proporção é quase inexistente, com diversas lacunas de uso e aplicações de modelos. Amor et al. (2023) elaboraram uma revisão bibliométrica e mapeamento sobre o métodos de ordenação, classificação e de *cluster* dos métodos MCDA, destacando as principais tendências dessas áreas no meio acadêmico. Nesse aprofundamento sobre os métodos, deram um pequeno enfoque em métodos de clusterização multicritério ordinal, destacando alguns dos principais autores e metodologias. Embora trate dessa nova área, sua abordagem foi bastante rasa, porém teve muita importância para estudos futuros de foco em revisões e experimentos, dando apoio para pesquisas brasileiras.

2.4.1 Diferenças entre a classificação ordinal multicritério e clusterização ordinal

Segundo Perny (1998) existem duas categorias principais em problemas de decisão. O primeiro refere-se a problemas relacionados à preferências, em que se deseja escolher as melhores alternativas ou ranquear todas as alternativas em ordem de preferência. A segunda categoria traz problemas relacionados à similaridade, ao qual deseja-se particionar as alternativas, de modo que eles sejam segmentados em *clusters* homogêneos, o que é definido como classificação parcial, ou em categorias pré-definidas, que é nomeado como classificação total. Em termos de performance, a desvantagem ao se usar esse tipo de método para a criação de classes é que pode levar a vários níveis de inconsistências e requer que os números de objetos em cada classe sejam fornecidos (FERNANDEZ et al., 2010; NEMERY e DE SMET, 2005).

Acerca da clusterização, os grupos não são conhecidos previamente e os objetos dentro de um grupo são semelhantes entre si, enquanto diferenciam-se se comparados com objetos de outros grupos (FERNANDEZ et al., 2010). Na clusterização ordinal, as alternativas são inicialmente ranqueadas e em seguida são agrupadas com base nas preferências desse *ranking*. Os *clusters* são ordenados conforme seus níveis de preferência. Isso permite com que a clusterização ordinal seja essencial para definir os perfis de classes para uma classificação *a posteriori* (SILVA et al., 2024).

Logo, apesar das suas concepções gerarem um pouco de confusão, existem diferenças marcantes que definem cada um dos problemas envolvendo classificação e clusterização. No geral, o que se diferencia a classificação multicritério da clusterização ordinal é o seu tipo de aprendizado. Enquanto a classificação multicritério é definida como técnica supervisionada em que as categorias são definidas *a priori*, a clusterização usa aprendizagem não supervisionada, em que os grupos são definidos *a posteriori*. Ao contrário da classificação, a clusterização não se preocupa com categorias pré-estabelecidas. (DE SMET e GUZMÁN, 2004; FERNANDEZ et al., 2010).

2.4.2 O modelo TOPSIS-Ckmeans

O método utilizado baseia-se no artigo de pesquisa de Silva et al. (2024) onde foi proposto um método de clusterização ordinal utilizando o método TOPSIS. Utilizaram um procedimento de clusterização unidimensional onde os *clusters* são obtidos a partir de um processo que utiliza programação dinâmica para extrair as partições. Dado a natureza do problema, duas situações se destacam:

- Os *clusters* formados preservam a ordenação obtida pelo método TOPSIS, ou seja, as alternativas pertencentes a um *cluster* com maior preferência são preferíveis (dominam) as alternativas pertencentes ao *cluster* de menor preferência.
- Dada a natureza do problema (unidimensional), o método gera um ótimo global para o problema da minimização intra-*cluster*.

O TOPSIS-Ckmeans é a combinação de um método multicritério TOPSIS e uma variante do K-Means e tem por objetivo ordenar dados e transformá-los em agrupamentos com dados unidimensionais. O diferencial desse método é que os valores do CC_i, que representa o índice de proximidade de cada alternativa do método são utilizados para se obter o agrupamento unidimensional, ou seja, as alternativas são segmentadas em *clusters* a partir do índice CC_i, de forma que a soma dos quadrados das distâncias sejam os menores intragrupo. O método de clusterização ordinal desenvolvido é definido pelos seguintes elementos (SILVA et al., 2024):

- O vetor de valores CC é organizado em ordem crescente, onde cada valor representa uma alternativa com valor ótimo, gerado a partir dos índices de proximidade do TOPSIS $CC = \{CC_1, CC_2, \dots CC_m\}$;
- O conjunto de *clusters* k são ordenados, onde $Clus = \{Cl_1, Cl_2, \dots, Cl_k\}$. O *cluster* Cl_l é pior que o *cluster* Cl_{l+1}, ou seja, o *cluster* Cl_{l+1} é preferível ao *cluster* Cl_l;
- $M = \{m_1, m_2, \dots, m_k\}$ simboliza o conjunto das médias que servem como representantes de cada *cluster* e funcionam de forma semelhante a centroides das alternativas pertencentes aos dos *clusters.*

O método segue as mesmas etapas do TOPSIS tradicional, porém tem-se a adição de mais dois passos para obter os *clusters*. O vetor do coeficiente de proximidade é gerado a partir do índice de proximidade, sendo utilizado o procedimento contido em Wang e Song (2011) para a geração dos *clusters* unidimensionais (SILVA et al, 2024).

- **1ª etapa:** carregar matriz de decisão D ;

- **2ª etapa:** definir pesos para cada atributo e o tipo de critério;
- **3ª etapa:** normalizar matriz D;
- **4ª etapa:** calcular as soluções ideais PIS e NIS;
- **5ª etapa:** calcular a distância euclidiana;
- **6ª etapa:** obter o coeficiente de proximidade CC_i e ranquear os valores em ordem crescente;
- **7ª etapa:** uso do índice CC_i para originar os vetores de proximidade;
- **8ª etapa:** formação dos agrupamentos unidimensionais com o procedimento descrito em Wang e Song (2011).

O algoritmo de clusterização Ckmeans.1d.dp, desenvolvido em Wang e Song (2011), é uma variação do método K-Means. A diferença desses dois é que o K-Means considera dados multidimensionais, enquanto o algoritmo proposto considera apenas dados em uma dimensão. Por considerar unidimensional, esse método garante que a solução obtida é um ótimo global. Nesse presente estudo, assim como no trabalho de Silva et al. (2024), foi utilizado o algoritmo Ckmeans.1d.dp para a formação dos *clusters* ordenados.

O algoritmo chamado de Ckmeans.1d.dp, apresentou valores ótimos com eficiência e precisão em comparação com o K-Means, ultrapassando o algoritmo em níveis de otimalidade em grupos maiores de *clusters*, repetibilidade e tempo de execução (WANG e SONG, 2011). Por esses motivos foi usado em um trabalho de Song e Zhong (2020) para mapear padrões de desregulação ao longo de uma cadeia de cromossomos e em Silva et al. (2024) combinado com o TOPSIS para geração de *clusters* ordenados.

Além disso, os resultados não variam a cada execução como o K-Means, pois não depende de várias eventuais inicializações, permitindo trazer resultados mais eficientes e estáveis. Combinando com o TOPSIS, garante que essas alternativas possam ser ranqueadas, a fim de definir os melhores resultados com base no seu índice de proximidade CC_i e serem agrupados com uma consistência maior que a do K-Means (SILVA et al., 2024).

De maneira geral, o uso de algoritmos de clusterização unidimensional não são comuns na literatura. Há uma certa dificuldade quanto à pesquisa nessa vertente. Alguns autores utilizam dados unidimensionais com algum método ou combinação de métodos e algoritmos de clusterização, mas até onde este trabalho alcançou, não foi

encontrado nenhum método multicritério combinado com clusterização unidimensional. Sabe-se que Poon, Liu e Zhang (2018), Liu, Poon e Zhang (2015), Costa (2020) e Grønlund et al. (2017) trabalharam com esse campo de pesquisa de clusterização unidimensional, no entanto nada foi encontrado artigos que tenha relação com o TOPSIS, variantes do mesmo método ou outros métodos bastante conhecidos como PROMETHEE, ELECTRE e etc.

O TOPSIS por ser um método que tenha mais facilidade de uso, é um dos métodos mais usados por pesquisadores por sua facilidade de implementação, com tomadas de decisão menos subjetivas e uma análise mais robusta. A utilização dele se destaca também na clusterização ordinal principalmente combinado com K-Means, o qual os *clusters* que são gerados nesse tipo de abordagem passam a ter uma relação ordinal em termos de preferências, do mesmo modo que as classes na classificação multicritério, no entanto, a diferença entre a classificação multicritério é que o *clusters* não são conhecidos *a priori* (SILVA et al., 2024), necessitando de menos esforço cognitivo para geração das classes por parte do decisor.

No geral, com a existência de pesos nos critérios, devido ao uso de um método de decisão multicritério, essa nova abordagem tem a flexibilidade de aplicar valores de peso diferentes em cada atributo de uma base ou banco de dados, o que pode ajudar quando se trata de decisores que queiram aplicar importância relativa em critérios de interesse, considerando-os mais importantes que outros. Dessa forma, permite gerar agrupamentos mais significativos de acordo com os anseios do decisor. Esse tipo de metodologia é mais comum em organizações que dão mais importância a determinados critérios que outros, de acordo com o alinhamento de algum projeto, visto que pode influenciar significativamente os resultados no processo de decisão (ODU, 2019).

3 REVISÃO DE LITERATURA

Essa seção irá discorrer sobre os principais trabalhos relacionados ao modelo de RFM combinados com métodos de decisão multicritério e o algoritmo de K-Means, levando em consideração o que cada autor tratou em suas pesquisas, em termos de abordagens e objetivos.

3.1 APLICAÇÕES DO MODELO RFM COMBINADO COM MCDA

Essa subseção explora os trabalhos que integram a análise RFM combinada com decisão multicritério, tratando dos seus principais pontos e da metodologia usada, descrevendo como cada método foi utilizado. Há uma variedade de aspectos que foram considerados em termos de abordagens na segmentação de clientes, alguns com metodologias inovadoras, outros com estudos de caso que incluíam ferramentas de análise de clientes.

Zaheri et al. (2012) abordaram um modelo que estuda a fidelização do cliente com base no RFM a fim de priorizar o consumidor através do TOPSIS e nas propriedades de fidelidade. Para calcular o coeficiente de importância relativa, utilizaram a matriz de comparação de pares que é inspirada no método AHP. O resultados mostraram que 30% dos clientes tinham valor monetário menor que o valor monetário médio total, e que os clientes com maior valor monetário médio não tinham uma prioridade muito maior, em um determinado período de tempo, comparando com outros clientes.

Barrera et al. (2024a) desenvolveram um método de matriz de classificação multicritério, de fácil interpretação para classificar alternativas de uma cadeia de suprimentos. Primeiramente, houve uma busca global que pré-classificou as alternativas por meio dos fluxos líquidos do PROMETHEE II. Em seguida, foi feito duas buscas locais que exploram propriedades discriminantes dos sinais de fluxo líquido, a fim de melhorar a qualidade dessas atribuições. As dimensões do RFM e Colaboração do Cliente (CC) foram utilizadas. O modelo foi validado com dados reais pela segmentação RFM de vários clientes e comparado com um outro método alternativo. Foi revelado que o método se provou robusto à variação de parâmetros e facilidade na interpretação dos resultados.

Barrera et al. (2024b) introduziram um modelo sistemático multicritério, que integra o AHP, o PROMETHEE II e o algoritmo GLNF, para apoiar tomada de decisão relacionada à segmentação de clientes em um contexto B2B. O modelo baseado no comportamento transacional do cliente é estendido, ou seja, além dos critérios RFM, há a presença de outras variáveis. Um indicador de qualidade, nomeado de SILS, foi implementado e validado em um banco de dados reais de uma multinacional. Os resultados do sistema proposto são comparados com um modelo baseado no K-Means. O sistema foi validado através de dados de clientes, onde foram agrupados em quatro grupos, ordenados de acordo com as suas características preferenciais. Os resultados indicaram que o grupo C apresentou clientes mais valiosos.

Aggarwal e Yadav (2020) aplicaram a segmentação de clientes baseado nas variáveis de compra: Recência, Frequência e Valor Monetário. Em seguida, foi dado pesos às variáveis através da técnica *Fuzzy* AHP. O estudo segmentou os dados em oito *clusters*, posteriormente, foram classificados com base nos valores de Vida Útil do Cliente (CLV). Ao final, os dados foram validados através da ANOVA, indicando que os *clusters* são significativamente diferentes um dos outros a um nível de significância de 5%.

Rezaeinia et al. (2012) focaram em dados de um banco comercial, estes foram processados e segmentados através do método RFM ponderado composto. A fim de calcular os pesos de cada variável de compra, foi usado o método AHP e o K-Means para segmentar cada cliente. Para verificar a consistência dos resultados do método multicritério, fizeram uso da técnica CRT para analisar o nível de confiança sobre as aplicações. Os resultados indicaram confiabilidade, permitindo uma validação positiva sobre o estudo.

Liu e Shih (2005) desenvolveram uma metodologia de recomendação de produtos que combinou técnicas de tomada de decisão multicritério e mineração de dados. Os pesos relativos das variáveis RFM foram calculados através do método AHP para avaliação do valor de vida útil (CLV) ou da fidelidade de cada cliente. Eles se utilizaram de técnicas de agrupamento para agrupar os clientes de acordo com o valor RFM ponderado. Foi demonstrado que a metodologia pode gerar recomendações de maior qualidade, mas não a todo tipo de cliente. Além disso, melhora em consumidores mais fiéis no que diz respeito à recomendação de mais itens.

Güçdemir e Selim (2015) foram outros pesquisadores que incluíram decisão multicritério e segmentação RFM. Em seu trabalho, propuseram uma abordagem de segmentação em que as variáveis são identificadas. Em seguida, quatro algoritmos dos tipos aglomerativo hierárquico e particional, foram testados para verificar qual o melhor método para agrupamento. O estudo usou o AHP para determinar a importância dos segmentos. Foram utilizadas variáveis de Recência, Frequência e Valor Monetário, onde os *clusters* foram segmentados com rótulos, acordo com a classificação do seu desempenho, numa análise *a posteriori* através dos cálculo das médias ponderadas dos centroides. Os dados do trabalho revelaram que a abordagem é efetivamente útil na prática de segmentação de clientes empresariais.

Mozafari et al. (2016) identificaram os usuários de uma biblioteca com base no Valor de Vida Útil e no modelo RFM. A análise da série temporal das frequências de visitas foram realizadas através de redes neurais artificiais e os usuários foram classificados com base na vida útil de cada um. O método AHP foi usado para calcular os pesos de cada variável de Recência, Frequência e Valor Monetário. Nessa pesquisa os conjunto de dados foi dividido em cinco *clusters*. Os autores descobriram que 3% dos usuários eram considerados de alto valor.

Azadnia et al. (2011) propuseram um estudo de caso de um modelo de avaliação de valor do cliente integrado com decisão multicritério e métodos de agrupamento *fuzzy*. O *Fuzzy* AHP foi utilizado para calcular os pesos das variáveis de segmentação RFM. Em seguida, o *Fuzzy* C-means foi aplicado para segmentar os clientes. Finalmente, agregaram o método TOPSIS para ranquear o valor de vida útil do cliente (CLV). O estudo conseguiu desenvolver um novo modelo de agrupamentos de clientes com base no CLV. Além disso, a pesquisa pode ajudar os gestores a tomar decisões melhores quanto as suas estratégias em segmentos de mercado.

Mahdiraji et al. (2019) sugeriram a utilização de uma abordagem orientada ao cliente, com dados de um banco, em que as variáveis de agrupamento foram computados com o base nos parâmetros RFM. Em seguida, o clientes foram agrupados pelo K-Means, totalizando seis *clusters*, sendo classificados por opiniões de especialistas, a fim de ponderar e classificar os *clusters* sob determinadas circunstâncias e classificados pelo BWM-COPRAS. Dois *clusters* apresentaram bons desempenhos, ademais foram selecionados e desenvolvido estratégias para os agrupamentos.

Por fim, há várias metodologias aplicadas a segmentação RFM combinadas com decisão multicritério, onde percebe-se que essa área há muitas aplicações para o campo empresarial. A maior parte desses trabalhos focam muito no método AHP para cálculo dos pesos sobre os critérios de Recência, Frequência e Valor Monetário. Há de se concluir que essas combinações de decisão multicritério e segmentação RFM relacionadas principalmente no identificação de clientes potenciais são muito usadas e importantes na área que envolve dados de clientes.

3.2 APLICAÇÕES DO MODELO RFM COMBINADO COM O K-MEANS

Nessa subseção, será explanado os principais trabalhos que versaram sobre segmentação RFM e o algoritmo K-Means, em que vários autores aplicaram algumas abordagens para classificar clientes. É discutido os processos dos métodos sobre dados de clientes, evidenciando alguns contextos importantes e sua abundância na área.

Chen et al. (2012) descreveram em seu artigo o objetivo de utilizar o método RFM mesclado ao K-Means para identificar os clientes potenciais e ajudar a empresa de varejo a entender seus compradores, conduzindo-os para um *marketing* centrado e efetivo. Eles elaboraram um estudo construindo um modelo através de um banco de dados de uma empresa de varejo, focados apenas nas vendas direcionadas no Reino Unido. Com isso, elaborou os cinco *clusters* em um gráfico tridimensional, classificou os tipos de consumidores existentes e os segmentou. Somado a isso, analisou e descreveu os perfis de clientes na segmentação.

Majilya et al. (2024) estudaram metodologias de agrupamento de clientes com a intenção de garantir às empresas vantagens competitivas. Para tal, desenvolveram um modelo que se utiliza do K-Means e RFM que segmenta cada clientes em determinados grupos. Eles pretendiam entender o comportamento de compra dos clientes a fim de identificar aqueles clientes de alto valor, de médio alcance e grupos de risco, com base nas pontuações de Recência, Frequência e Valor Monetário. Descobriram quatro grupos distintos de clientes, o que ajuda na elaboração de estratégias centradas à cada característica de clientes.

Ling e Weiling (2025) investigaram vários algoritmos de clusterização sobre dados de segmentação de clientes para investigar a precisão de agrupamentos e obter *insights* mais profundos sobre o comportamento do cliente. O intuito do artigo é

aprimorar a segmentação de clientes no campo de e-marketing. A análise partiu de dados de RFM, com o processo de refinamento do material que melhora a integridade e facilita a identificação dos segmentos. O K-means++ demonstrou um desempenho melhor no quesito precisão e em outras condições em relação aos outros algoritmos.

Patibandla et al. (2025) apresentam uma abordagem que integra o K-Means para segmentação de clientes e junto com a Análise de Componentes Principais (PCA) para a diminuição dimensional dos dados. Foi utilizado o RFM para reduzir a complexidade das informações, facilitar a tomada de decisões no setor de varejo e melhorar a estratégias de marketing que se utiliza desses dados. Destacaram que a junção de PCA e K-Means é eficaz na segmentação de cliente em ambientes dinâmicos varejistas.

Sharma, V. et al. (2024) elaboraram uma abordagem inovadora para identificar clientes-alvo e otimizar geração de receita das empresas. Eles categorizaram os dados dos clientes em agrupamentos diferentes de acordo com as características únicas, utilizando algoritmo K-Means, junto com a aplicação do RFM para classificar e agrupar os consumidores de acordo com suas transações. Além disso, duas outras ferramentas são utilizadas, o método Elbow para selecionar o número de *clusters* e a pontuação de Davies-Bouldin para determinar a contagem ideal de *clusters*. A fim de aplicar clientes aos grupos adequados, eles utilizaram um valor médio como um indicador primário. O método proposto dividiu os clientes em cinco grupos com base nos seus valores de RFM. O resultado mostrou que o *cluster* 3 apresentou 21% dos consumidores como os melhores clientes em termos de compras recentes e gastos maiores.

Akande et al. (2024) trabalharam com o uso de segmentação RFM baseada na técnica de clusterização K-Means com o objetivo de segmentar dados comportamentais dos clientes em várias categorias, como Prata, Bronze, Ouro, Platina ou Ruim, com a intenção de fortalecer a oferta de produtos, concentrar melhor sua comunicação de marketing e aumentar a fidelidade do cliente por meio dessa estratégia. A abordagem mostrou-se notável precisão e exatidão na análise dos resultados.

Solichin e Wibowo (2022) construíram uma abordagem de clusterização com uso de K-Means baseado nas variáveis de Recência, Frequência e Valor Monetário, não só para obter o comportamento de compra dos clientes, mas também foram combinados com o parâmetro de Rastreamento de Eventos do Usuário (UET) em

dados de transações de compra. Posteriormente essas informações foram segmentadas em categorias de Platina, Ouro e Prata.

Anitha e Patil (2022) aplicaram a inteligência de negócios para identificar os clientes potenciais, com dados de transações e varejo utilizando a clusterização. O estudo baseou-se no modelo RFM e princípios de segmentação de clientes usando o K-Means. Os resultados foram validados usando a abordagem do Coeficiente de Silhueta com diferentes números de *clusters* e apresentaram uma solução ótima às variáveis de Recência de Vendas, Frequência de Vendas e Valor Monetário de Vendas, com base na pontuação da Silhueta.

Brahmana et al. (2020) analisaram vários registros de transações e os converteu em dados de Recência, Frequência e Valor Monetário para identificar clientes potenciais. Para esse estudo aplicaram conceitos de mineração de dados e o CRM, combinando o RFM com algoritmos K-Means, K-Medoids e DBSCAN. A validação dos *clusters* foi feita usando os índices de Davies-Boundin e de Silhueta. Os resultados informaram que o K-Means apresentou melhor nível de validade em comparação aos outros algoritmos de clusterização, outrossim os algoritmos trouxeram clientes de classes *Golden* e *Dormant*.

Monalisa e Kurnia (2019) trouxeram dois o de algoritmos de clusterização em dados de RFM. O objetivo principal é identificar *outliers* em dados de comportamento de clientes. O comportamento foi determinado através do modelo de RFM, com o uso de K-means e DBSCAN na clusterização dos dados de clientes. Os resultados mostraram que o *cluster* 1 apresentou 100% de similaridade, no entanto a similaridade total entre os *outliers* foi de 67%, o que indicam que o comportamento dos clientes são de baixa frequência de compras, mas com alta recência e valor monetário.

Percebe-se que, por fim, há uma ampla variedade no uso da segmentação RFM combinado ao algoritmo K-Means. Enquanto alguns estudos focam em na aplicação clássica do modelo RFM alinhada ao agrupamento, outros buscam em diversificar as possibilidades por meio de ferramentas de *marketing* ou comparação com outros algoritmos de clusterização. No geral, o K-means é bastante adotado nesse contexto, pois representa as primeiras abordagens em tratar com dados transacionais de clientes, possibilitando a identificação de clientes potenciais por meio de agrupamentos envolvam variáveis de Recência, Frequência e Valor Monetário, destacando sua importância de uso na obtenção de conhecimento na formação de estratégias empresariais.

4 METODOLOGIA

A metodologia deste trabalho foi estruturada em algumas etapas descritas a seguir. Inicialmente, a extensão do TOPSIS-Ckmeans é descrita voltada à segmentação de clientes, com base em estudos anteriores que fundamentam sua aplicabilidade. Posteriormente, o estudo descreve o procedimento de processamento dos dados, a formação da tabela RFM e o processo de clusterização. Além disso, a análise exploratória dos resultados do TOPSIS-Ckmeans e do K-Means é realizada, avaliando a performance de ambos os métodos. Por fim, são discutidas as conclusões deste trabalho e as recomendações para pesquisas futuras. Trata-se de um estudo de caso, adotando procedimentos metodológicos de natureza quantitativa, com o propósito de apresentar contribuições sobre o modelo proposto.

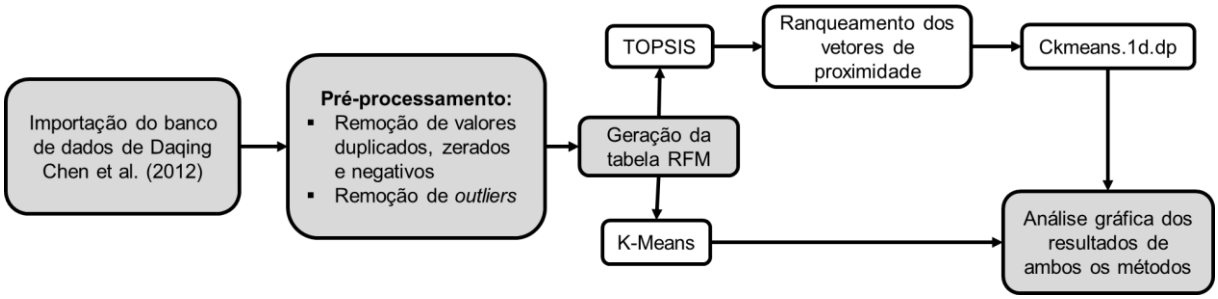
O referencial teórico se tratou de embasamento teórico, definições das ferramentas e métodos da Mineração de Dados que envolvem a segmentação de clientes, a análise RFM e a importância do seu uso na seleção de dados importantes. Junto a isso, focou na clusterização em geral com os tipos de algoritmos existentes na literatura, tratou de métodos de decisão multicritério e, por fim, descreveu a Clusterização Multicritério Ordinal, suas diferenças com os métodos de classificação e abordagens da metodologia.

Em seguida, a revisão de literatura reuniu os principais estudos relacionados à aplicação da abordagem RFM com o uso de métodos de decisão multicritério e segmentação RFM combinada com o algoritmo K-Means. Os trabalhos selecionados foram apresentados conforme suas respectivas abordagens em diferentes contextos de aplicação, juntamente com a descrição de resultados obtidos, contribuindo para o fortalecimento teórico do uso dessas duas metodologias na segmentação de clientes.

Com isso, o presente trabalho consistiu em criar um nova abordagem de clusterização multicritério ordinal em dados de segmentação RFM. Para análise da performance do método desenvolvido, foram utilizados dados realísticos da literatura.

A aplicação da metodologia do trabalho se divide em quatro etapas, baseados no pré-processamento de Chen et al. (2012), descrito abaixo:

Figura 7 - Etapas da pesquisa



Fonte: a autora (2025)

Nessa primeira etapa, os dados foram obtidos do repositório público UCI Machine Learning Repository, cedidos por Chen (2015). Os registros de cada compra abrangem o período entre 1º de dezembro de 2010 e 9 de dezembro de 2011, totalizando 541.909 linhas referentes a transações realizadas por cliente de um determinado E-*commerce* localizado no Reino Unido. O conjunto de dados inclui valores faltantes e dados ruidosos e contêm 4.372 registros de código distintos de compradores. Dentre esses registros, 25.900 correspondem a transações únicas (*InvoiceNo*), enquanto 406.829 instâncias possuem códigos postais válidos. O arquivo está no formato compatível com Microsoft Excel e contém 8 colunas: *InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID e Country.* Para essa pesquisa, foi utilizado o Python 3.11.9, executado em um notebook equipado com processador AMD RYZEN 5 5000U de 2.10GHz, 8gb de memória RAM e sistema operacional Windows 11.

Na segunda etapa, com a intenção de transformar os dados brutos em informação útil, foi realizado o pré-processamento e filtragem de *outliers* para evitar mudanças significativas nos resultados finais. Junto a isso, o estudo também se concentrou exclusivamente em dados com códigos postais do Reino Unido, considerando apenas as transações realizadas no período de janeiro de 2011 a dezembro de 2011.

Após essa etapa de filtragem, os dados foram transformados em informações relevantes para a construção das variáveis de segmentação de clientes: Recência, Frequência e Valor Monetário. A variável de Recência foi calculada com base na data da última compra registrada por cliente. A Frequência foi obtida através do total de transações realizadas por cada cliente, dentro do período de um ano. Por fim, a variável de Valor Monetário foi calculada por meio da soma de todos os valores gastos

por cliente durante o período analisado, resultando em um montante total. Nisso, o conjunto de dados foi separado para cada método, a fim de que cada um aplicasse as suas metodologias, desde o procedimento de normalização ao agrupamento das instâncias. Antes de passarem pela clusterização, os métodos foram submetidos ao método de Elbow, a fim de verificar quantos *clusters* seriam determinados para a segmentação.

Na quinta etapa, para ilustrar e graficamente as tabelas e os resultados obtidos, foi empregado o Jupyter Lab para visualizar os procedimentos. Os pacotes aplicados no estudo foram o `Pandas`, a fim de visualizar os registros de transações e a biblioteca `Numpy` para tratamento de valores numéricos e funções matemáticas. Além disso, os pacotes `Matplotlib.pyplot` e `Seaborn` foram utilizados para análises gráficas. Adicionalmente, foram empregados a função `Scipy.stats` e o procedimento de `StandardScaler()`, da biblioteca `Sklearn.preprocessing`, no processo de padronização das escalas da tabela RFM, especialmente vinculado à clusterização do método K-Means. Por fim, o pacote `Sklearn.cluster` para chamar o algoritmo KMeans no processo de clusterização, dentre outras funções.

Com relação ao procedimento do TOPSIS-Ckmeans, foi elaborado no JupyterLab uma classe do método TOPSIS tradicional que carregasse a matriz de decisão para ranqueamento das alternativas, com base no vetor de proximidade. Após isso, o Ckmeans.1d.dp foi importado através de um módulo `%pip install ckmeans-1d-dp` para agrupar os vetores com base em sua importância.

Por fim, os dois métodos, o TOPSIS-Ckmeans e o K-Means, foram comparados utilizando o agrupamento de K=3, que gerou os *clusters* 0, 1 e 2. A avaliação foi conduzida através das estatísticas descritivas, com a intenção de analisar isoladamente o desempenho de cada agrupamento e de cada variável de Recência, Frequência e Valor Monetário. Para complementar a análise, os gráficos de *boxplot* foram empregados para verificar o comportamento das médias e medianas de cada variável. Por fim, a matriz de comparação foi adotada para examinar a correspondências e divergências de cada *cluster* formado, evidenciando quais perfis foram agrupados de forma semelhante e quais foram segmentados em *clusters* diferentes, considerando cada um dos métodos.

5 RESULTADOS

Neste capítulo serão discutidos os resultados do procedimento proposto na metodologia. Para medir o desempenho de cada um dos dois métodos, foram feitas análises da geração de estatísticas descritivas e da geração de gráficos, considerando os diferentes *clusters* formados pelo método TOPSIS-Ckmeans em contraste com o método K-Means.

Vale ressaltar que as estatísticas descritivas, seleção de dados relevantes, pré-processamento dos mesmos, visualização em gráficos dinâmicos e *boxplots* são alguns recursos importantes para entender as informações captadas. O pré-processamento faz com o que melhore o desempenho da mineração de dados, no que diz respeito ao custo, tempo e qualidade das informações (TAN et al., 2006). Além das estatísticas descritivas e análises gráficas, uma matriz de comparação foi empregada com o objetivo de contrastar as diferenças dos agrupamentos em relação aos dois métodos analisados, possibilitando um entendimento mais acurado acerca da correspondências entre os *clusters* de cada método.

5.1 PRÉ-PROCESSAMENTO

A primeira e a segunda etapa desde trabalho consistem na extração dos dados a partir do repositório compartilhado por Chen (2015) e, em seguida, a filtragem do material, com o objetivo de identificar informações úteis e relevantes. Para evitar que anormalidades intervissem nos resultados do estudo, entre elas valores e código postais vazios, dados negativos, inconsistências que podem contribuir em dados com pouca precisão (HAN et al., 2012), foi aplicado o filtro de limpeza no conjunto de dados. Essa segunda etapa foi essencial para que se assegurasse uma maior robustez nos resultados, especialmente para os agrupamentos. O conjunto de transações original era composto por oito colunas e 541.909 instâncias.

Tabela 3 - Definição das variáveis

Variável	Descrição	Tipo
InvoiceNo	Pedidos de clientes por cartão postal	Numérico
StockCode	Código de cada produto vendido	Nominal
Description	Nome de cada produto comprado	Nominal

Quantity	Quantidade de vezes que um mesmo produto foi comprado	Numérico
InvoiceDate	Data em que determinados produtos foram comprados pelo consumidor naquele varejo	Numérico
Unitprice	Valor unitário de cada produto	Numérico
CustomerID	Código que diferencia um cliente de outro na compra	Numérico
Country	País por onde aquela compra foi feita	Nominal

Fonte: a autora (2025).

A Tabela 3 destaca a definição de cada coluna (variável) do conjunto de dados de varejo. A maioria das variáveis são representadas por dados numéricos, colunas estas mais importantes para identificar *insights* valiosos dos padrões de comportamento de compra e a formação dos parâmetros R, F e M.

Tabela 4 - Banco de dados brutos

Invoice No	Stock Code	Description	Quantity	Invoice Date	Unit Price	Customer ID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17850.0	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17850.0	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17850.0	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17850.0	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09/12/2011 12:50	0.85	12680.0	France
581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09/12/2011 12:50	2.10	12680.0	France
581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09/12/2011 12:50	4.15	12680.0	France

581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09/12/2011 12:50	4.15	12680.0	France
581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09/12/2011 12:50	4.95	12680.0	France

Fonte: a autora (2025)

A Tabela 4 representa as primeiras e últimas linhas de transações de vendas *online* realizadas por clientes de uma empresa de varejo. Cada item vendido é representado por um InvoiceNo, de maneira que cada linha do mesmo código correspondem a produtos de uma mesma fatura, vinculadas a um cliente específico. Por exemplo, na primeira linha, no dia 01/12/2010, o cliente 17850.0, localizado no Reino Unido, comprou 6 unidades do item “WHITE HANGING HEART T-LIGHT HOLDER”, cada um custando 2,55. Mas também, na mesma fatura ele comprou outras unidades de itens diferentes, registrados por um mesmo InvoiceNo.

Tabela 5 - Estatísticas descritivas dos dados brutos

	Quantity	**InvoiceDate**	**UnitPrice**	**CustomerID**
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	01/12/2010 08:26	-11062.060000	12346.000000
25%	1.000000	28/03/2011 11:34	1.250000	13953.000000
50%	3.000000	19/07/2011 17:17	2.080000	15152.000000
75%	10.000000	19/10/2011 11:27	4.130000	16791.000000
max	80995.000000	09/12/2011 12:50	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

Fonte: a autora (2025)

Na Tabela 5, podem ser visualizadas as estatísticas descritivas para cada uma das variáveis numéricas antes do pré-processamento dos dados. Percebe-se que a coluna Quantity apresenta valores mínimos de -80.995 assim com a coluna e Unitprice com valor negativo de -11.062. Verificando o desvio padrão das duas colunas, entende-se que há uma variabilidade muito grande de preços e de quantidade. É importante visualizar esse ponto, pois esses valores negativos são ruídos que afeta os resultados da clusterização (HAN et al., 2012). Duas ações podem ser feitas para

minimizar o impacto da variabilidade: substituir esses valores negativos ou totalmente removê-los. Nesse caso, optou-se pela última alternativa, uma vez que o número de valores inconsistentes é ínfimo em relação a quantidade total de dados.

Primeiramente, os dados foram filtrados para que contivessem apenas instâncias válidas na coluna CustomerID e estivessem relacionadas exclusivamente ao caixa postal do Reino Unido, contabilizando no total, 361.878 instâncias. A fim de que se pudesse obter os valores da variável de Valor Monetário na terceira etapa, foi preciso criar uma variável auxiliadora, Amount, que é adquirida através do cálculo do produto entre as colunas de UnitPrice e Quantity. Esse cálculo resulta no valor total gasto de cada linha de venda. A partir disso, os valores do Amount foram agregados por cliente, permitindo calcular o montante total gasto por cada CustomerID. Durante o processo de preparação do conjunto de transações, foi aplicado o procedimento de exclusão dos valores negativos, inconsistentes e duplicados contidos nos registros. Em adição para a análise, foram considerados apenas os dados a partir de janeiro de 2011 a dezembro do mesmo ano, totalizando um total de 325.745 instâncias.

Para gerar a variável de Recência, a coluna de InvoiceDate foi utilizada para encontrar o tempo desde a data da última compra, considerando a respectiva caixa postal de cada cliente. Os valores foram divididos por 30 para serem convertidos em meses, considerando que 0 representa uma compra concluída há aproximadamente um mês e 11 representa uma transação finalizada há 12 meses. Com relação à variável de Frequência, foi calculada a partir do número de transações por cliente dentro do período de um ano. Para isso, utilizou-se a coluna de InvoiceNo, agrupando os registros de compras por CustomerID e contabilizando a quantidade de transações de cada cliente. Com base nessas informações, a tabela de segmentação RFM foi gerada, na qual cada linha representa um cliente e seus respectivos valores nas três variáveis R, F e M.

Tabela 6 - Tabela RFM gerada a partir do conjunto de registros de transações dos clientes

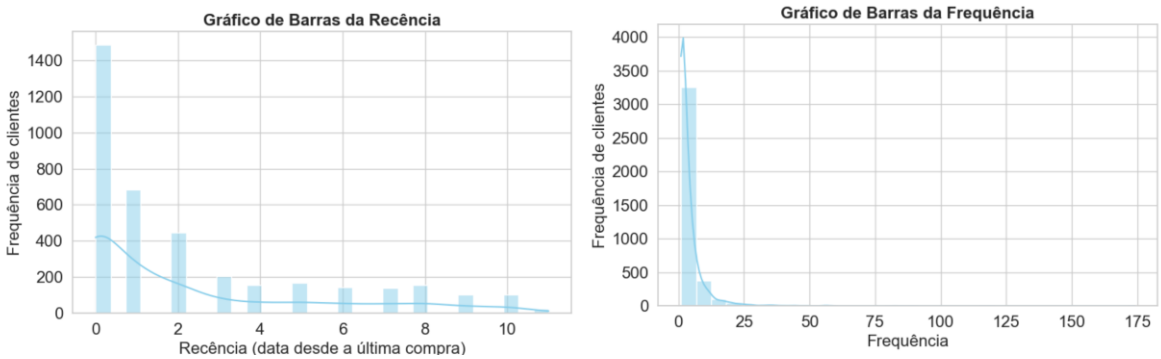
CustomerID	Recência	Frequência	Valor Monetário
12346.0	10	1	77183.60
12747.0	0	9	3489.74
12748.0	0	174	28868.19
12749.0	0	5	4090.88
12820.0	0	4	942.34
...	...	...	...

Fonte: a autora (2025)

A Tabela 6 representa as cinco primeiras instâncias da Recência, Frequência e Valor Monetário. Cada código de cliente contém a quantidade de meses desde última transação registrada, a quantidade total de vezes que realizou transações e o montante total realizados dentro do período de 1 ano.

A quantidade de instâncias ao final desse processo de refinamento, com unificação de todas as compras por cliente foi de 3.813 linhas de dados. A tabela indica que, por exemplo, na primeira linha, o valor de 10 para o cliente de código 12346.0 mostra que esse cliente realizou a primeira compra na empresa de varejo há mais ou menos 11 meses, gastando em um total de 77.183.60 em um único dia. Ao observar o cliente de número 12748.0, é possível avaliar que ele realizou 174 transações, gastando um montante total de 28.868.19. Junto a isso, a sua última compra foi realizada há 1 mês, em um período de 1 ano. Quando mais próximo de zero, mais recente o cliente comprou na empresa. Nesse sentido, a Recência se apresenta como um critério de minimização, enquanto a Frequência e o Valor Monetário são considerados critérios de maximização.

Figuras 8a e 8b - Gráfico de barras da Recência e da Frequência antes da eliminação de *outliers*



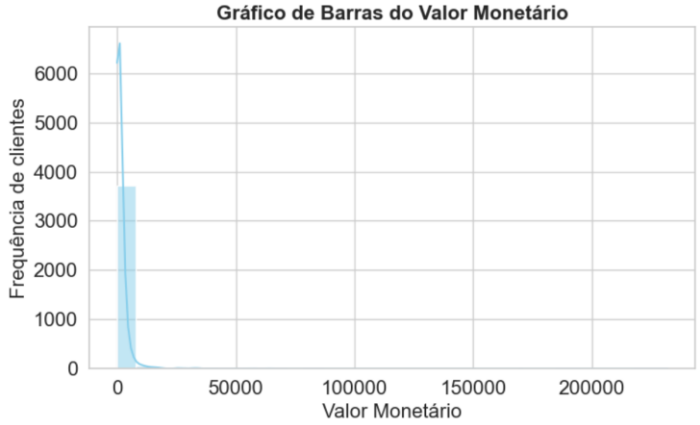
Fonte: a autora (2025)

Os Gráficos de Barras 8a e 8b apresentam a distribuição dos dados nos critérios de Recência e Frequência, respectivamente, onde cada variável foi visualizada antes da utilização de qualquer procedimento de filtragem dos *outliers*. No caso da Recência, observa-se um valor máximo de 11 meses, com uma concentração significativa de clientes, mais de 1.400 registros de recências iguais a zero. Isso representa que concretizaram a última compra no mês de dezembro de 2011. A média

geral para essa variável permaneceu em 2,35 meses desde a última compra, o que sugere que a média de clientes realizaram compras há mais de 3 meses.

Na Frequência, o número máximo de transações concluídas por um cliente foi 174 vezes, dentro do período analisado. A maioria dos consumidores apresentaram um frequência de compra abaixo de 25. De maneira mais específica, observou-se que 1.363 clientes efetuaram apenas uma transação no ambiente comercial, 717 frequentaram apenas 2 vezes, 458 clientes somente 3 vezes e 347 apenas 4 vezes. A média da variável de Frequência de cada consumidor é de 4,02 vezes em um ano.

Figura 9 - Gráfico de barras do Valor Monetário antes da eliminação dos *outliers*



Fonte: a autora (2025)

Em relação à Figura 9 do gráfico de barras do Valor Monetário, observou-se um gasto máximo de 231.822,69 por um cliente, enquanto o menor valor registrado foi de 3,75. No que diz a respeito da frequência de códigos postais por cliente, mais de 3.000 consumidores gastaram menos em produtos, com uma média de 1.780,36 por cliente. De acordo com Chen et al. (2012), esses dados atípicos trazem prejuízos quanto a aplicação dos métodos de agrupamento e interferem na análise dos resultados finais. Para evitar distorções nas análises de clusterização, foi preciso realizar um procedimento para eliminação dos *outliers*.

Previamente, foi definido um limite superior para filtrar os dados que estivessem acima dessa restrição. Para isso, utilizou-se o procedimento de Chen et al. (2012), onde o valor de corte foi definido para 1% do total das instâncias com valores superiores. Desta forma, 181 instâncias foram removidas dessa etapa de filtragem. Nessa pós-filtração, as instâncias foram reduzidas a 3.632 valores por código de cada cliente.

Tabela 7 - Estatísticas descritivas do RFM após a eliminação dos *outliers*

	Recência	**Frequência**	**Valor Monetário**
Contagem	3.632	3.632	3.632
Média	2,11	3,59	1.259,79
Mediana	1	2	646,79
Desvio Padrão	2,63	3,72	1.743,41
Mínimo	0	1	3,75
Máximo	9	26	16.362,90
Curtose	0,30440	7,99280	16,89999
Assimetria	1,22538	2,54474	3,51663

Fonte: a autora (2025)

A Tabela 7 mostra as estatísticas descritivas da matriz de dados RFM após a eliminação dos *outliers* ao qual descrevem aspectos importantes ao visualizar o comportamento dos dados. Na Recência, o valor mínimo de 0 sugere que a transação mais recente dentro do período analisado foi realizada há um mês, já o valor máximo foi de 9 e representa a transação mais antiga registrada. Na Frequência, o mínimo de transações cadastradas no conjunto de dados foi de uma única vez, enquanto o valor máximo que um perfil de cliente realizou foi de 26 transações. Por fim, na variável de Valor Monetário, o montante mínimo encontrado nos registros foi de 3,75, enquanto o montante total máximo chega a 16.362,90.

As medianas descrevem que 50% dos clientes realizaram a última compra há aproximadamente um mês, o que sugerem que uma base dos clientes são consideravelmente ativos. Na Frequência, a mediana apresenta que metade dos consumidores realizaram duas transações dentro de um ano. Com relação ao Valor Monetário, a mediana é de 646,79, revelando que 50% dos consumidores gastaram até esse montante. Enquanto isso, a média dessa variável é 1.259,79, sugerindo que clientes tem gastos significativamente maiores que a mediana.

Em relação ao formato de distribuição da curva, a Recência apresenta uma curtose de 0,30440 e uma assimetria de 1,22538. Isso indica que, ao observar essas medidas, a curtose é positiva, o que caracteriza uma curva levemente leptocúrtica em relação à média, ou seja, mais alongada e concentrada, refletindo em dados mais localizados. Quanto à assimetria, os valores indicam que a distribuição é assimétrica à direita, isto é, a maioria dos dados estão localizados à esquerda da média e uma cauda mais alongada à direita. Essa observação sugere que uma parcela considerável de clientes realizou transações recentes, enquanto poucos consumidores estão há mais tempo sem comprar algum produto da empresa.

Na variável Frequência, a curtose apresenta um valor bastante positivo (7,99280) o que indica uma distribuição fortemente leptocúrtica. A assimetria de 2,54474, por sua vez, também revela uma distribuição positiva, revelando uma cauda à direita e dados reunidos à esquerda, o que significa que uma boa parte dos clientes compram poucas vezes, enquanto uma minoria realizam transações com mais frequência.

De mesmo modo, o Valor Monetário exibe uma curtose de 16,89999 e assimetria de 3,51663, ambas positivas, trazendo uma distribuição mais centralizada e assimétrica à direita. Esses dados indicam a presença de poucos clientes que apresentam valores elevados e são destacados como um grupo de alto valor, enquanto uma grande parcela de outros clientes realizam gastos moderados ou baixos.

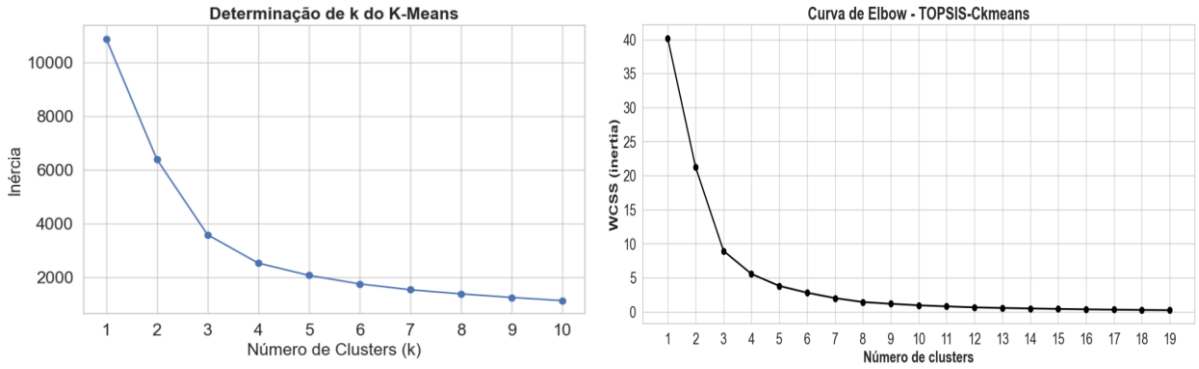
No TOPSIS-Ckmeans, as variáveis de RFM foram avaliadas com os pesos igualmente distribuídos $\{0{,}333, 0{,}333, 0{,}333\}$ para que cada parâmetro fosse avaliada de maneira equilibrada. Esse tipo de abordagem evita com que as variáveis tenham uma influência desproporcional sobre os resultados. Quanto ao tipo de critério adotado, a abordagem foi de otimização na Recência e minimização na Frequência e Valor Monetário.

5.1.1 Implementação e interpretação do K-Means e TOPSIS-Ckmeans

Ainda no terceiro passo da pesquisa, consistiu-se em analisar os resultados do K-Means quanto do TOPSIS-Ckmeans paralelamente. Essa etapa é importante para avaliar a coerência de cada *cluster* diante das variáveis de Recência, Frequência e Valor Monetário.

Para definição dos *clusters*, foi utilizado o método gráfico de Elbow. Essa ferramenta é empregada para definir quantos *clusters* serão apropriados no processo de agrupamento, sendo o número de *clusters* determinados através da posição da “dobra” ou “cotovelo” (HUMAIRA e RASYDAH, 2020). Analisando os dois gráficos abaixo, percebe-se que tanto para o K-means quanto para TOPSIS-Ckmeans, a partir do valor de K=3, não se obtém tanta vantagem em termos de separabilidade dos *clusters* ao se aumentar o valor de k. Logo, para esse trabalho, os dois métodos foram comparados, considerando o valor de K=3.

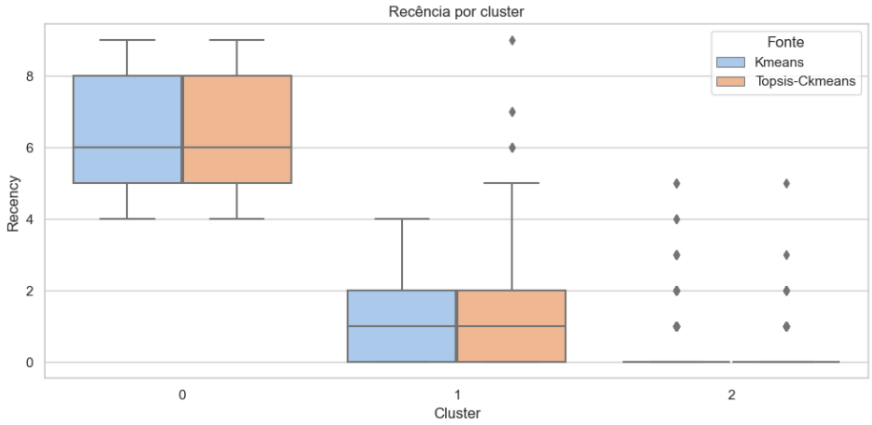
Figuras 10a e 10b - Gráfico de Elbow para determinar a quantidade de *clusters* do K-Means/TOPSIS-Ckmeans



Fonte: a autora (2025)

A fim de comparar os resultados pelos dois métodos, foram utilizados gráficos de *boxplot* para verificar os resultados das três variáveis isoladamente, considerando cada um dos *clusters*. Além disso, foram comparadas estatísticas descritivas geradas a partir de cada *cluster*, analisando cada um dos métodos.

Figura 11 - *Boxplot* da Recência por *cluster* pós-filtragem K=3



Fonte: a autora (2025)

A Figura 11, representa grupos decrescentes, devido a natureza de minimização da Recência. Pode ser observado que, através dos gráficos de *boxplot*, os valores de Recência tendem a diminuir, considerando ambos os métodos. Há aparentemente uma similaridade em termos de formato do *boxplot* no *cluster* 0 nas duas metodologias. Ou seja, mesma estrutura de *boxplot* em limite superior, limite inferior e medianas. Ao visualizar as medianas dos *clusters 1 e 2*, o TOPSIS-Ckmeans apresentou também uma performance similar, ao agrupar perfis de clientes com

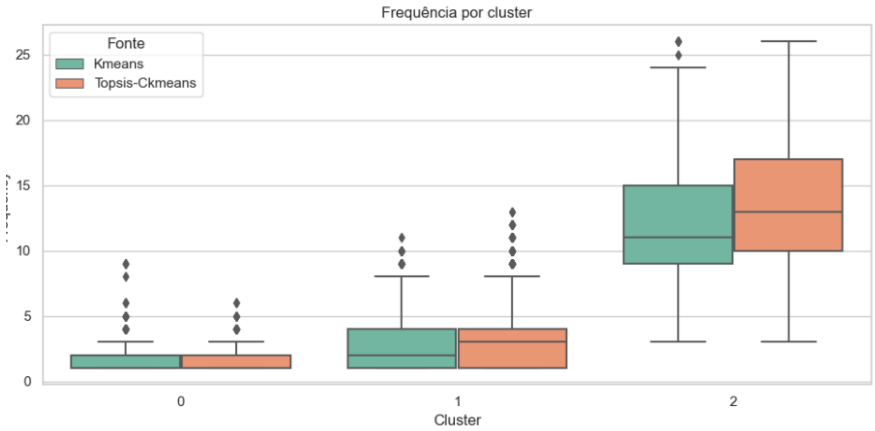
recências baixas, intermediárias e ótimas. Vale ressaltar que, apesar das semelhanças nesse critério com o K-Means, uma das vantagens ao usar o TOPSIS-Ckmeans nesse atributo é a sua distribuição mais criteriosa, conseguindo separar clientes que podem ser considerados menos, médio ou mais importantes de maneira mais heterogênea, o que pode ajudar o decisor, por exemplo, a encontrar os agrupamentos com uma melhor separação.

À medida que os gráficos tem uma tendência a avançar para os *clusters* 1, onde estão localizados os clientes de recências intermediárias, depreende-se que o método TOPSIS-Ckmeans apresenta valores ligeiramente maiores em relação ao limite superior que o K-Means, alcançando outros perfis de clientes. Em relação ao K-Means, o método apresentou uma homogeneidade maior no *cluster* 1. Nos *clusters* 2, o TOPSIS-Ckmeans manteve uma performance melhor em relação à uma menor concentração de *outliers*.

Apesar das semelhanças nas medianas dos três agrupamentos, em termos gerais, o TOPSIS-Ckmeans apresentou uma leve performance positiva, ainda que discreta, especialmente no *cluster* 2 em valores de recência, comparando com o método tradicional. Nessa visualização, TOPSIS-Ckmeans demonstrou maior capacidade de agrupar perfis de clientes com recências mais baixas no *cluster* 2, com uma distribuição mais criteriosa e uma menor presença de *outliers*.

No *cluster* 0, ele apresentou semelhanças com o K-Means, contudo, o TOPSIS-Ckmeans, considerando o *cluster* 0 como o pior agrupamento, esboçou uma performance um pouco melhor de segmentação, em termos de outras medidas de tendência central, como a média, que é explicada detalhadamente na análise das estatísticas descritivas. No *cluster* 1, os dados apresentaram uma leve diferença pois o modelo proposto reuniu clientes com valores que condiziam com um agrupamento de clientes nem muito bons nem muito ruins. Vale observar que, as diferenças irão se destacar ainda mais à medida que o intervalo de recência for ainda maior.

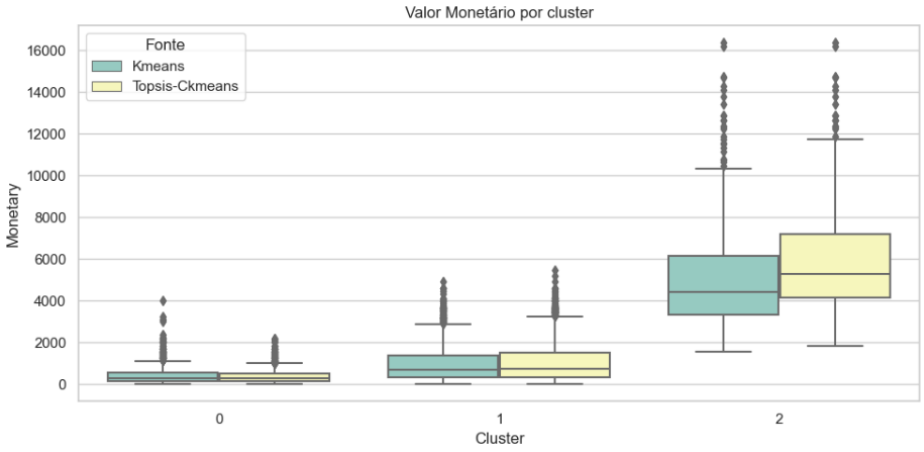
Figura 12 - *Boxplot* da Frequência pós-filtragem K=3



Fonte: a autora (2025)

Ao comparar a Figura 12, pode ser observado, através dos gráficos, que os valores de Frequência tendem a crescer quando a análise avança do *cluster* 0 para o *cluster* 2, considerando as duas abordagens. Comparando-os, percebe-se que o método TOPSIS-Ckmeans possui valores superiores de medianas, em relação aos *clusters* 1 e 2. A situação se altera quando é considerado o *cluster* 0, onde o modelo proposto apresenta valores de mediana semelhantes ao K-Means, diferenciando-se apenas nos valores atípicos, já que o TOPSIS-Ckmeans contém menor quantidade de *outliers*. Nesse sentido, o modelo apresentou melhor distribuição de medianas nos *cluster* 0, 1 e 2, especialmente o agrupamento 2, pois agrupou perfis de clientes de melhores frequências que o K-Means e maior seleção de perfis de consumo de alto valor.

Figura 13 - *Boxplot* da Valor Monetário pós-filtragem K=3



Fonte: a autora (2025)

Na Figura 13, encontra-se os gráficos de *boxplot* dos valores monetários dos três agrupamentos. Assim como no caso do gráfico anterior, nota-se que os *clusters* tem uma tendência de crescimento, em relação à mediana, à medida em que se avança do *cluster* 0 para o *cluster* 2, tanto no TOPSIS-Ckmeans, quanto no K-Means. No *cluster* 0, O TOPSIS-Ckmeans, apresentou um mediana menor, cerca de 305,76, quando comparado com o método K-means, cerca de 309,54, o que indica que, considerando o pior agrupamento, o modelo proposto conseguiu segmentar melhor os clientes de baixo montante com mais destaque. A análise muda quando se considera o *cluster* 1 e 2, onde o método TOPSIS-Ckmeans apresenta valores de mediana intermediários e superiores com maior performance, respectivamente, ao comparar com o K-Means.

Ao avaliar o TOPSIS-Ckmeans nessa variável, evidencia-se que o modelo proposto é mais eficiente em detectar e distribuir perfis de consumo com baixo, médio e alto valor monetário. Com relação ao K-Means, o método assume uma postura de favorecer medianas relativamente menores, o que resulta na absorção de perfis que tenham valores monetários ligeiramente mais baixos nos três agrupamentos.

Em termos gerais, todos os *clusters*, considerando a Recência, Frequência e Valor Monetário, o modelo proposto trouxe performances positivas, especialmente em saber separar com definição cada tipo de agrupamento. Destaca-se ainda o *cluster* 2 como o melhor perfil de consumo.

5.1.2 Comparação das estatísticas descritivas

Nessa subseção será explanado uma comparação entre as estatísticas descritivas de cada método, considerando as clusterizações obtidas a partir de cada método. A análise estatística dos *clusters* é importante para entender o perfil de cliente que cada método segmentou, analisando as médias, desvios e outros indicadores, destacando as principais características dos segmentos. As estatísticas descritivas são utilizadas para resumir dados de maneira organizada, a fim de verificar a relação entre variáveis em uma amostra ou uma população (KAUR et al., 2018).

Tabela 8 - Estatísticas descritivas do *cluster* 0

K-Means	TOPSIS-Ckmeans

	Recência	Frequência	Valor Monetário	Recência	Frequência	Valor Monetário
Contagem	847	847	847	818	818	818
Média	6,36	1,57	438,95	6,42	1,48	394,11
Desvio Padrão	1,64	1,01	436,06	1,63	0,80	335,78
Mínimo	4	1	3,75	4	1	3,75
Máximo	9	9	4.036,96	9	6	2.188,50
Curtose	-1,22703	13,34055	12,39936	-1,21574	4,36546	5,05508

Fonte: a autora (2025)

Na Tabela 8, o *cluster* 0 destaca-se o pior agrupamento em perfis de consumo com alta recência, baixa frequência e baixo montante. Comparando as recências de ambos os métodos, percebe-se que o TOPSIS-Ckmeans apresenta leves diferenças entre as médias, com 6,42, comparando com 6,36 do K-Means. Nas frequências, o TOPSIS-Ckmeans indicou 1,48 enquanto o K-Means teve uma performance melhor, com 1,57. Em termos de Valor Monetário, o modelo proposto apresentou uma média de 394,11 e no método tradicional, 438,95. Isso representa que, considerando o *cluster* 0 o pior agrupamento em segmentar perfis de clientes, o TOPSIS-Ckmeans conseguiu destacar melhor os grupos de perfis de clientes com a pior performance possível nas três variáveis. Em contrapartida, o K-Means selecionou clientes levemente melhores, o que permite constatar que ele não conseguiu segmentar com a mesma performance que o TOPSIS-Ckmeans, em termos de baixo desempenho.

Isso pode ser observado também nos valores máximos. O TOPSIS-Ckmeans apresentou Frequência de 6 e Valor Monetário de 2.188,80. Enquanto isso, o K-Means agrupou máximas de 9 de frequência e 4.036,96 de montante total. Essas diferenças, indicam que, em termos de variabilidade, o TOPSIS-Ckmeans foi mais restritivo, segmentando os piores perfis de compra. Ao contrário das recências, o valor máximo são iguais em ambos os métodos, isto é, iguais a 9.

Em relação à curtose, no primeiro momento, o K-Means apresentou uma curva de -1,22703, 13,34055, 12,39936 na Recência, Frequência e Valor monetário, respectivamente. Esses dados informam que, a Recência apresentou uma curtose negativa (platicúrtica), direcionando uma curva mais achatada, mostrando que os dados estão mais distribuídos. Contudo, ao olhar a Frequência e o Valor Monetário, as informação são de curtoses mais positivas, indicando que os dados estão muito mais alongados, onde os clientes estão mais concentrados em torno da média e pode indicar que há uma maior presença de valores extremos.

Com relação ao TOPSIS-Ckmeans, os dados tem uma variação mais expressiva na Frequência e no Valor Monetário, mas sem maior destaque tanto quanto o K-Means nesses dois atributos, com -1,21574 (Recência), 4,36546 (Frequência), 5,05508 (Valor Monetário). No geral, ambos apresentaram uma recência semelhante, contudo, os outros critérios, principalmente do TOPSIS-Ckmeans apresentam uma curva um pouco menos acentuada, mas ainda bastante positivas, trazendo dados agrupados em torno da média, em uma curtose leptocúrtica.

Tabela 9 - Estatísticas descritivas do *cluster* 1

	K-Means			**TOPSIS-Ckmeans**		
	Recência	**Frequência**	**Valor Monetário**	**Recência**	**Frequência**	**Valor Monetário**
Contagem	2.423	2.423	2.423	2.562	2.562	2.562
Média	0,89	3,00	956,28	0,92	3,28	1.056,79
Desvio Padrão	1,01	1,99	813,07	1,09	2,36	921,87
Mínimo	0	1	6,2	0	1	6,2
Máximo	4	11	4.919	9	13	5.472
Curtose	-0,31829	0,53563	1,69498	2,18883	1,19166	1,37308

Fonte: a autora (2025)

Na Tabela 9, podem ser comparadas as estatísticas descritivas dos *clusters* 1 gerados pelo método TOPSIS-CKmeans e o método K-means. Os *clusters* 1 se tratam de agrupamentos com valores intermediários, ou seja, dados que tem uma performance um pouco melhor com relação aos *clusters* 0. Ao analisar a média do TOPSIS-Ckmeans, o método apresentou médias maiores nas variáveis de Frequência, com 3,28 e Valor Monetário de 1.056,79. Contudo, ao comparar a média da Recência, considerando que, quanto mais recente a última compra, melhor, esse princípio se inverte. A média do TOPSIS-CKmeans apresenta ser ligeiramente maior, com 0,92, em relação ao K-Means que permanece com uma Recência de 0,89.

A diferença na variável de Recência também pode ser observada ao comparar os valores máximos de ambos os métodos. Enquanto o K-Means apresentou uma Recência menor de perfis de consumo, igual a 4, o TOPSIS-Ckmeans alcançou dados de até 9. Considerando ser o agrupamento com valores intermediários, essa mudança de performance do TOPSIS-Ckmeans em relação à Recência sugere que o método destacou mais perfis de consumo nem muito bons, nem muito ruins, apresentando um desempenho ainda superior se comparado com o *cluster* 0.

A curtose nesse agrupamento, no TOPSIS-Ckmeans, apresentou valores positivos, o que indica que os dados apresentaram maior concentração em torno da média, com curvas mais alongadas 2,18883 (Recência), 1,19166 (Frequência), 1,37308 (Valor Monetário), enquanto o K-Means apresentou valores de -0,31829 (Recência) com uma curtose negativa e, por sua vez, mais achatada; 0,53563 (Frequência) e 1,69498 (Valor Monetário), ambos com curtoses positivas. A curtose do TOPSIS-Ckmeans trouxe curvas que indicam que os dados são mais consistentes. Em relação ao K-Means, principalmente na Recência, traz uma curtose mais achatada e com maior variação de dados.

Tabela 10 - Estatísticas descritivas do *cluster* 2

	K-Means			**TOPSIS-Ckmeans**		
	Recência	**Frequência**	**Valor Monetário**	**Recência**	**Frequência**	**Valor Monetário**
Contagem	362	362	362	252	252	252
Média	0,27	12,28	5.211,87	0,17	13,57	6.133,72
Desvio Padrão	0,73	4,78	2.798,09	0,55	5,11	2.866,80
Mínimo	0	3	1.539,49	0	3	1.829,60
Máximo	5	26	16.362,90	5	26	16.362,90
Curtose	14,31324	0,28033	2,29494	26,41782	-0,30721	1,38363

Fonte: a autora (2025)

Por fim, a Tabela 10 aponta os resultados das estatísticas descritivas dos *clusters* 2, que revelam ser os melhores agrupamentos, com clientes mais valiosos. É possível destacar algumas variações levemente expressivas entre os dois métodos no perfil de clientes classificados. No K-Means, a média de Recência apresenta 0,27, na Frequência, 12,28 e no Valor Monetário, 5.211,87. No TOPSIS-Ckmeans a média de Recência é de 0,17, na Frequência indica 13,57 e Valor Monetário apresenta média de 6,133,72.

Pode ser observado que as médias do TOSPIS-Ckmeans apresentam destaque ao serem maiores que as médias do K-Means. Comparando ambos os métodos, o TOPSIS-Ckmeans conseguiu segmentar os melhores clientes com mais eficiência, garantindo uma performance superior em todas as três variáveis. Considerando os valores mínimos, as diferenças na seleção de perfis de clientes tem variação em somente no Valor Monetário, o que pode ser observado que o TOPSIS-Ckmeans trouxe o mínimo de 1.829,60, enquanto o K-Means agrupou o montante mínimo de 1.539,49. Considerando que o *cluster* 2 é o melhor grupo dentre os três

clusters analisados, o montante menor do K-Means garante que ele não conseguiu agrupar os melhores perfis de consumo em comparação com o TOPSIS-Ckmeans, limitando-se a valores monetários inferiores.

Por fim, nas curtoses da Recência de ambos os métodos, K-Means com 14,31324 e TOPSIS-Ckmeans com 26,41782, nota-se que o modelo proposto apresenta uma curva bastante alongada, com dados mais concentrados em relação ao método tradicional, adequando clientes com menor variação de valores em torno da média. Em relação às Frequências de ambos (K-Means com valores de 0,28033 e TOPSIS-Ckmeans com -0,30721) já se percebe uma alteração. O TOPSIS-Ckmeans apresentou uma curtose negativa (platocústica), trazendo maior variabilidade sobre os dados dos clientes, diferentemente do K-Means ao trazer dados menos dispersos. Com relação ao critério de Valor Monetário, as variações são menores (K-Means com curtose em 2,29494 e TOPSIS-Ckmeans em 1,38363), com uma curtose positiva, leptocústica.

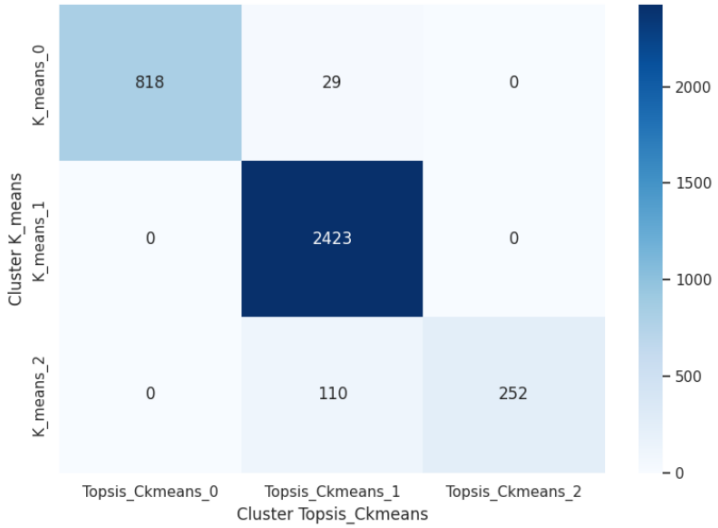
De modo geral, destaca-se então que o TOPSIS-Ckmeans apresentou desempenhos superiores nos *clusters* 0, 1 e 2. Isso mostra que o método tem maior sensibilidade em segmentar perfis de clientes, conseguindo identificar grupos com características que se diferenciam melhor entre si. Essas diferenças de médias evidencia-se mais no *cluster* 2, com performances ainda melhores que os *clusters* anteriores.

Baseado nas características de ambos nos *clusters* de ambos os métodos, é possível definir quais são os clientes que são valiosos e os que não são, cada grupo pode ser categorizado em Ouro, Prata e Bronze, baseado no trabalho de Kumar N. (2025). A primeira categoria define os clientes muito recentes, com alta frequência e com alto valor de compra, sendo aqueles de maior interesse pelas organizações. A segunda categoria são aqueles que, apesar de não consumirem muito, são recentes e mantém uma frequência razoável. A terceira categoria se destaca os clientes que não compram há um bom tempo, tem baixa frequência e gastam pouco em produtos, respectivamente. Logo, aplicando as categorias aos agrupamentos, o *cluster* 2 destaca-se os grupos de clientes Ouro, o *cluster* 1 os clientes Prata e o *cluster* 0 os clientes Bronze, o que permite a empresa elaborar estratégias diferentes para cada grupo.

5.1.3 Matriz de comparação

Para poder analisar o grau de correspondência entre os *clusters* dos dois métodos, aqueles obtidos pelo K-Means e os agrupamentos gerados pelo TOPSIS-Ckmeans, a matriz de comparação foi necessária para entender se o modelo teve um grau de consistência considerável e enxergar se apresentou diferença na classificação das alternativas nos dois métodos. Na matriz em questão, é feita uma análise sobre quantas observações permaneceram em um mesmo agrupamento em ambos os métodos e quantas foram colocadas em diferentes agrupamentos.

Figura 14 - Matriz de comparação dos dois métodos



Fonte: a autora (2025)

Na Figura 14, os *clusters* 0, 1 e 2 foram nomeados de *K_means_0, K_means_1 e K_means_2*, concomitantemente, para se referir aos agrupamentos do método K-Means e *Topsis_Ckmeans_0, Topsis_Ckmeans_1, Topsis_Ckmeans_2*, para se referir aos agrupamentos 0, 1 e 2 do método multicritério ordinal, respectivamente. Todas as 3.632 instâncias são apresentadas distribuídas nos três *clusters*.

Na matriz de comparação, observa-se que há uma consistência significativa entre os agrupamentos, especialmente na visualização do *cluster* 1, nomeado como *K_means_1*. Nele, 2.423 observações foram classificadas identicamente tanto pelo método tradicional, quanto pelo TOPSIS-Ckmeans. Esse resultado indica uma alta estabilidade e concordância entre os dois métodos, pois reconheceram o mesmo perfil, com nenhum outro dado alocado em outro agrupamento.

No *cluster K_means_0*, apresentou uma correspondência de 818 observações que coincidem ao mesmo *cluster* 0 do TOPSIS-Ckmeans, nomeado como *Topsis_Ckmeans_0*. No entanto, 29 observações foram alocadas no *Topsis_Ckmeans_1*, o que indica uma possível sobreposição de alguns perfis dos dois agrupamentos. Em outras palavras, o modelo reconheceu esses dados como pertencentes do *cluster* 1 do TOPSIS-Ckmeans, destacando sua natureza distinta de agrupar com base no *ranking* das alternativas.

Finalmente, o *cluster K_means_2*, o agrupamento mostrou maior diferença. 110 observações foram identificadas e alocadas no *cluster* 1 do TOPSIS-Ckmeans. Esse padrão encontrado sugere que esse grupo tinha dados menos homogêneos e que, possivelmente, a maneira como método TOPSIS-Ckmeans classifica os elementos pode ter contribuído para uma distribuição mais sensível, capaz de segmentar e distinguir com mais afinidade os perfis que anteriormente estavam mais agregados no K-Means. Ainda assim, 252 observações no *K_means_2* foram identificadas corretamente no *Topsis_Ckmeans_2*, o que demonstra uma certa estabilidade nas alocações.

Consequentemente, a matriz de comparação evidencia, que apesar dos métodos compartilharem semelhanças quanto a sua distribuição dos dados nos *clusters* 0 e 1, a diferença se acentua muito mais na distribuição das instâncias no *cluster* 2. Embora os métodos corroborem em grande parte dos *clusters*, a escolha do método pode interferir significativamente na distribuição dos elementos no agrupamentos. Assim, pode-se dizer que o TOPSIS-Ckmeans abordou uma performance mais analítica, especialmente no *cluster* 2, havendo uma reorganização maior, em comparação ao K-Means, ao avaliar parte das observações do *cluster* 2 no *cluster* 1. Ou seja, 110 instâncias foram distribuídas no *cluster* 1 em vez do *cluster* 2 o que indica uma diferença substancial no método de classificação entre um método e outro.

6 CONCLUSÃO

O presente estudo trouxe dados importantes quanto a sua análise de *clusters* e permitiu que estes fossem avaliados com base em um modelo multicritério ordinal. Houve algumas observações adicionais quanto ao modelo, mas de modo geral, trouxe sugestões para uma análise mais diversificada.

O trabalho propôs um novo modelo segmentação de clientes, utilizando um método de clusterização multicritério ordinal, TOPSIS-Ckmeans, sobre dados de segmentação RFM, comparando o método proposto com o K-Means, a fim de entender como os dados serão comparados e segmentados, analisando suas performances em termos de estatísticas descritivas e análise gráfica. Os resultados mostraram que apesar dos métodos compartilharem certas semelhanças entre si, o TOPSIS-Ckmeans mostrou destaque nas médias nos três *clusters*.

Descobriu-se que o TOPSIS-Ckmeans teve melhor robustez quanto ao *cluster* 2 em todos os parâmetros avaliados, refletindo em médias mais acentuadas. O modelo proposto, no *cluster* 2, trouxe uma seleção de clientes com valores monetários mais altos, recências menores e frequências maiores. Em relação aos *clusters* 0 e 1, a sua performance foi melhor, no que diz respeito a agrupar o pior *cluster* possível e o *cluster* de valores nem muito ruins nem muito bons, respectivamente, o que destaca a sua capacidade de conseguir separar os perfis de clientes com mais eficiência que o K-Means. Além disso, a matriz de comparação mostrou uma diferença considerável no *cluster* 2 o que pode significar uma perda de valor.

Esse estudo trouxe *insights* significativos com esse novo modelo, pois pode beneficiar a área de clusterização multicritério ordinal e a segmentação de clientes. Além disso, pode trazer estratégias empresariais para minimizar o custo de operações que se esforçam para retirar conhecimento em dados de clientela, em termos de tempo e implementação. Em adição, o novo modelo proposto garante segmentar o clientes com melhor distribuição e robustez. De certo, este modelo gera uma solução ótima global e termos de clusterização, enquanto o uso do K-Means pode ter um pouco de dificuldade com grandes bancos de dados, no cálculo dos centroides e segmentação de instâncias a grupos de maneira mais homogênea.

Além disso, o estudo contribuiu em aspectos tal como a criação de um novo modelo de clusterização multicritério ordinal com a priorização das preferências do decisor, que envolvesse a segmentação de clientes à ampliação do uso da análise

RFM em diferentes contextos. Por fim, ajudou em novas abordagens de ranqueamento de alternativas e agrupamento com um ótimo global utilizando o TOPSIS.

Em termos de decisão, o método é interativo e flexível que considera a escolha do decisor, especialmente na aplicação dos pesos em cada variável e na seleção do tipo de critério. Na utilização do vetor de proximidade, que é utilizado na ranqueamento, garante uma facilidade para implementação do Ckmeans.1d.dp a dados extremamente grandes em menos tempo, além de garantir um ótimo global.

Como recomendação para trabalhos futuros, sugere-se a inclusão de um modelo de decisão multicritério baseado em um método TOPSIS sem o problema da reversão de ordem, combinando com algoritmos de agrupamentos para verificar a robustez em termos de *clusters* ordenados, com testes de remoção ou adição de alternativas. Além disso, sugere-se a necessidade de expandir o TOPSIS-Ckmeans que lide com situações sob incerteza, reforçando a necessidade de ampliar o método no conceito de lógica *fuzzy* (SILVA et al., 2024; MADANCHIAN e TAHERDOOST, 2023).

Uma segunda sugestão consiste na mudança de pesos dos critérios de Recência, Frequência e Valor Monetário. Nem sempre as três variáveis são consideradas com a mesma a relevância para todos os setores, seja da indústria ou mercado de varejo. Para algumas empresas, uma variável é significativamente mais importante que as outras e vice-versa, e o RFM tradicional não adota o conceito de pesos de preferências do decisor para se adequar a cada situação (YEH et al., 2009). O TOPSIS-Ckmeans traz essa possibilidade de considerar vetores de peso, podendo se adequar a diferentes ocasiões conforme a necessidade.

Outra sugestão seria a combinação de outros métodos de decisão multicritério com o Ckmeans.1d.dp no aprofundamento do tema de *clusters* unidimensionais, além da utilização do TOPSIS-Ckmeans em versões expandidas do modelo RFM, que incorporam novas variáveis de comportamento do cliente, a fim de verificar outras análises de perfis de consumo.

Além disso, recomenda-se a implementação de testes de desempenho que comparem o modelo com outros algoritmos, a saber, uma variação do k-medoids, o CLARA, em termos de grande base de dados e processamento (BRITO et al., 2011) e o Bisecting K-means, no que diz respeito à eficiência de agrupamentos (ROHILLA et al., 2019). Para lidar com dados categóricos, sugere-se a criação de um método de

clusterização multicritério ordinal combinado com o K-modes, a fim de abarcar variáveis nominais que não são consideradas adequadamente.

Em suma, propõe-se o desenvolvimento de versões híbridas do método multicritério ordinal com esses mesmos algoritmos de clusterização, integrados a métodos de decisão multicritério, como o TOPSIS, PROMETHEE, ELECTRE, SMARTS, entre outros. Com essas abordagens, permitiriam a ampliação de métodos de decisão combinados com métodos de agrupamentos de diferentes tipos, na tentativa de solucionar outros problemas relacionados à robustez, alocação de *clusters*, dentre outros.

REFERÊNCIAS

ABBASIMEHR, H.; SHABANI, M. A new framework for predicting customer behavior in terms of RFM by considering the temporal aspect based on time series techniques. **Journal of Ambient Intelligence and Humanized Computing**, v. 12, p. 515-531, 2021. https://doi.org/10.1007/s12652-020-02015-w.

ABCOMM. Números do e-commerce brasileiro. **Portal Associação Brasileira de Comércio Eletrônico (ABComm)**. Disponível em: <https://dados.abcomm.org/numeros-do-ecommerce-brasileiro>. Acesso em: 4 ago. 2025.

AGGARWAL, A. G.; YADAV, S. Customer Segmentation Using Fuzzy-AHP and RFM Model. In: **IEEE International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)**, 8., 2020, Delhi. New Delhi: IEEE, 2020.

AKANDE, O. N.; AKANDE, H. B.; ASANI, E. O.; DAUTARE, B. T. Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy. **International Conference on Science, Engineering and Business for Driving Sustainable Development Goals**, SEB4SDG, 2024.

ALMEIDA, A. T. **Processo de decisão nas organizações: construindo modelos de decisão multicritério.** São Paulo: Atlas, 2013.

ALMEIDA, A. T.; CAVALCANTE, C. A. V.; ALENCAR, M. H.; FERREIRA, R. J. P. ALMEIDA-FILHO, A. T.; GARCEZ, T. V. **Multicriteria and multiobjective models for risk, reliability and maintenance decision analysis**. Cham, Switzerland: Springer International Publishing, 2015.

AMOR, S. B.; BELAID, F.; BENKRAIEM, R.; RAMDANI, B.; GUESMI, K. Multi-criteria classification, sorting, and clustering: a bibliometric review and research agenda. **Annals of Operations Research**, v. 325, p. 1-23, 2023

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, 2022.

ARORA, P.; DEEPALI; SHIPRA, V. Analysis of k-means and k-medoids algorithm for big data. **Procedia Computer Science**, v. 78, p. 507-512, 2016.

AZADNIA, A. H.; SAMAN, M. Z. M.; WONG, K. Y.; HEMDI, A. R. Integration model of Fuzzy C means clustering algorithm and TOPSIS Method for Customer Lifetime Value Assessment. *In*: **2011 IEEE International Conference on Industrial Engineering and Engineering Management**. IEEE, p. 16-20, 2011.

BANDGAR, S. **CLUSTERING ON IRIS DATASET IN PYTHON USING K-Means - Analytics Vidhya - Medium**. Disponível em: <https://medium.com/analytics-vidhya/*cluster*ing-on-iris-dataset-in-python-using-k-means-4735b181affe>. Acesso em: 29 jun. 2025.

BARRERA, F.; SEGURA, M.; MAROTO, C. A Multicriteria Customer Classification Method in Supply Chain Management. **Mathematics**, v. 12, n. 3427, 2024a.

BARRERA, F.; SEGURA, M.; MAROTO, C. Multiple criteria decision support system for customer segmentation using a sorting outranking method. **Expert Systems with Applications**, v. 238, p. 122310, 2024b.

BASHIR. M. A.; MUHIUDDIN, G.; RASHID, T.; SARDAR. M. S. Multicriteria Ordered the Profile Clustering Algorithm Based on PROMETHEE and Fuzzy c-Means. **Mathematical Problems in Engineering**, v. 2023, p. 13, 2023.

BERALDO, F,. Marketplaces em 2025 no Brasil: Qual o Cenário e Vale a Pena Entrar? **Ciclo**. Disponível em: <https://cicloecommerce.com.br/marketplaces-em-2025-no-brasil-qual-o-cenario-e-vale-a-pena-entrar/>. Acesso em: 4 ago. 2025.

BOUJELBEN, M. A. A unicriterion analysis based on the PROMETHEE principles for multicriteria ordered clustering. **Omega**, v. 69, p. 126–140, 2017.

BRAHMANA, R. W. S.; MOHAMMED, F. A.; CHAIRUANG, K. Customer segmentation based on RFM model using K-means, K-medoids, and DBSCAN methods. **Lontar Komput. J. Ilm. Teknol. Inf**, v. 11, n. 1, p. 32, 2020.

BRITO, W. M.; SEMAAN, G. S.; BRITO, J. A. M. Um Algoritmo Genético para o Problema dos K-Medoides. *In:* 0th Brazilian Congress on Computational Intelligence (CBIC'2011), **Anais do 10. Congresso Brasileiro de Inteligência Computacional**, Fortaleza, 2011.

BULT, J. R.; WANSBEEK, T. Optimal Selection for Direct Mail. **Marketing Science**, v. 14, n. 4, p. 378-394, 1995.

CASSIANO, K. M. Análise de Séries Temporais usando Análise Espectral Singular (SSA) e clusterização de suas componentes baseada em densidade. 2014. **Tese** (Doutorado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, p. 155, 2014. Disponível em: http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=24787@1. Acesso em: 12 jun. 2025.

CELEBI, M. E. Improving the performance of k-means for color quantization. **Image and Vision Computing**, v. 29, n. 4, p. 260-271, 2011.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. **Expert systems with applications**, v. 40, n. 1, p. 200-210, 2013.

CHEN, D. **Online Retail [Dataset].** UCI Machine Learning Repository, 2015. Disponível em: <https://archive.ics.uci.edu/dataset/352/online+retail>. Acesso em: 20 set. 2024.

CHEN, D.; SAIN, S. L.; GUO, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. **Journal of Database Marketing &Customer Strategy Management**, v.19, n. 3, p. 197-208, 2012.

Disponível em: <https://link.springer.com/article/10.1057/dbm.2012.17>. Acesso em: 04 set. 2024.

CHEN, Y.; RAUF, A.; SHAFIQUE, A.; TCHIER, F.; ASLAM, A.; TOLA, K. A. Physicochemical profiling and ranking of parkinson's disease drugs through QSPR and Fuzzy TOPSIS analysis. **Scientific Reports**, v. 15, n. 1, p. 1-15, 2025.

CHRISTY, A. J.; UMAMAKESWARI, A.; PRIYATHARSINI, L.; NEYAA, A. RFM ranking – An effective approach to customer segmentation. **Journal of King Saud University - Computer and Information Sciences,** v. 33, n. 10, p. 1251–1257, dez. 2021.

CHURCHMAN, C. W.; ACKOFF, R. A.; ARNOFF, E. L. **Introduction to Operations Research**. Nova York: Wiley, 1957.

COSTA, L. F. Toward Generalized Clustering through an One-Dimensional Approach. **ArXiv**, p. 1-8, 2020. Disponível em: https://arxiv.org/abs/2001.02741.

COUTINHO, T. **Entenda o que é a Matriz RFM e aprenda como desenvolvê-la.** 8 ago. 2022. Disponível em: <https://voitto.com.br/blog/artigo/matriz-rfm>. Acesso em: 27 set. 2024.

DE SMET, Y. P2CLUST: An extension of PROMETHEE II for multicriteria ordered clustering. *In:* **2013 IEEE International Conference on Industrial Engineering and Engineering Management**, Bangkok, p. 848-851, 2014.

DE SMET, Y.; GILBART, F. A class definition method for country risk problem. 2001.

DE SMET, Y.; GUZMÁN, L. M. Towards multicriteria clustering: An extension of the k-means algorithm. **European Journal of Operational Research**, v. 158, n. 2, p. 390–398, 2004.

DE SMET,Y.; NEMERY, P.; SELVARAJ, R. An exact algorithm for the multicriteria ordered clustering problem. **Omega**, v. 40, n. 6, p. 861-869, 2012.

DOĞAN, O.; AYÇİN, E; BULUT, Z. A. Customer segmentation by using RFM model and clustering methods: a case study in retail industry. **International Journal of Contemporary Economics and Administrative Sciences**, v. 8, n. 1, p. 1-19, 2018.

DOMINGUEZ, G. A.; RUANO, F. C. Aplicação de um modelo RFM alargado em Mercados B2B: uma abordagem multivariada aplicada ao mercado do Aço. p. 1-86, 2024. Lisboa: Portugal. **Projeto de mestrado**. Disponível em: https://repositorio-aberto.up.pt/bitstream/10216/164321/2/700925.pdf

ERNAWATI; BAHARIN, S. S. K.; KASMIN, F. Spatial-temporal analysis using two-stage clustering and GIS-based MCDM to identify potential market regions. **Journal of System and Management Sciences,** v. 11, n. 4, p. 87–112, 2021.

FARAN, J.; TRIAYUDI, A.; ALDISA, R. T. Combination of RFM's (Recency Frequency Monetary) method and Agglomerative Ward's method for donors segmentation. *In*: **2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)**. IEEE, 2023. p. 962-967.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 26 feb. 2025.

FELIX, H. Clusterização**: Começando pelo básico com RFM Segmentation Model**. Disponível em: <https://medium.com/@heitorfelix/*cluster*iza%C3%A7%C3%A3o-come%C3%A7ando-pelo-b%C3%A1sico-com-rfm-segmentation-model-c850331d1a6d>, 2022. Acesso em: 5 jun. 2025.

FERNANDEZ, E.; NAVARRO, J.; BERNAL, S. Handling multicriteria preferences in cluster analysis. **European Journal of Operational Research**, v. 202, n. 3, p. 819-827, 2010.

FRANKLIN, B. **Sr. Franklin: A Selection from His Personal Letters**. Yale University Press, 1956.

GATA, W.; ISKANDAR; BASRI, H.; PUSPITAWATI, D.; HIDAYAT, S; WALIM. Implementation of Decision Tree Algorithm in Customer Recency, Frequency, Monetary, and Cost Profiling: a Case Study of Plastic Packing Industry. *In*: **IOP Conference Series: Materials Science and Engineering**. IOP Publishing, 2019. p. 022032.

GRØNLUND, A.; KASPER, G. L.; MATHIASEN, A.; NIELSEN, J. S.; SCHNEIDER, S.; SONG, M. Fast exact k-means, k-medians and Bregman divergence clustering in 1D. **arXiv preprint arXiv:1701.07204**, 2017.

GS1 BRASIL. Pesquisa GS1 Brasil aponta estratégias de crescimento. **GS1 Brasil.** Disponível em: <https://noticias.gs1br.org/pesquisa-gs1-brasil-estrategias-crescimento/>. Acesso em: 4 ago. 2025.

GÜÇDEMIR, H.; SELIM, H. Integrating multi-criteria decision making and clustering for business customer segmentation. **Industrial Management & Data Systems**, v. 115, n. 6, p. 1022-1040, 2015.

GÜRBÜZ, M. A. Identification of Multiple Extraction Methods to be Used in Extraction of Macro and Micronutrients of Neutral and Alkaline Soils by a Multi-Criteria Decision-Making Technique (TOPSIS). **Journal of Soil Science and Plant Nutrition**, v. 25, n. 2, p. 4669-4686, 2025.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques.** 3 ed. EUA, Morgan Kaufmann, 2012.

HANDOJO, A.; PUJAWAN, N.; SANTOSA, B.; SINGGIH, M. L. A multi layer recency frequency monetary method for customer priority segmentation in online transaction. **Cogent Engineering**, v. 10, n. 1, 2023.

HARTIGAN, J. A.; WONG, M. A. A K-means clustering algorithm. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 28, n. 1, p. 100-108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN J**. The elements of statistical learning: Data Mining, Inference, and Prediction.** 2 ed, Springer: New York, 2009.

HOSSENI, M. B.; TAROKH, M. J. Customer segmentation using CLV elements. **Journal of Service Science and Management**, v. 4, n. 3, p. 284-290, 2011.

HUGHES, A. M. **Strategic Database Marketing.** 4 ed. McGraw-Hill Companies, 2012.

HUMAIRA, H.; RASYIDAH, R. Determining the Appropriate Cluster Number Using Elbow Method for K-Means Algorithm. *In:* **Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA 2018)**, Padang, Indonésia, 2020.

HUSNAH, M.; NOVITA, R. Clustering of Customer Lifetime Value with Length Recency Frequency and Monetary Model Using Fuzzy C-Means Algorithm. *In*: **2022 International Conference on Informatics Electrical and Electronics (ICIEE)**. IEEE, 2022. p. 1-4.

HWANG, C. L.; YOON, K. **Multiple Attribute Decision Making: Methods and Applications.** Springer-Verlag, New York, 1981.

IBRAHIM, M. R. K.; TYASNURITA, R. LRFM Model Analysis for Customer Segmentation Using K-Means Clustering. *In*: **2022 International Conference On Electrical and Information Technology (IEIT)**, 2022. IEEE, 2022.

IKOTUN, A. M.; EZUGWU, A. E.; ABUALIGAH, L., ABUHAIJA, 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.

IPEA – INTITUTO DE PESQUISA ECONÔMICA APLICADA. **Agenda 2030: ODS-Metas nacionais dos objetivos de desenvolvimento sustentável**. Ipea, 2018.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, Seul, v. 31, n. 8, p. 651–666, jul. 2010. DOI: 10.1016/j.patrec.2009.09.011. Disponível em: https://www-scopus-com.ez16.periodicos.capes.gov.br/record/display.uri?eid=2-s2.0-77950369345&origin=resultslist&sort=cp-f&src=s&sid=a58cce55ef996d6a8aaf38e27701c43f&sot=a&sdt=a&s=TITLE%28Data+*cluster*ing%3A+50+years+beyond+K-means%29&sl=47&sessionSearchId=a58cce55ef996d6a8aaf38e27701c43f. Acesso em: 10 fev. 2025.

JUANITA, S.; CAHYONO, R. D. K-means clustering with comparison of Elbow and Silhouette methods for medicines clustering based on user reviews. **Jurnal Teknik Informatika**, v. 5, n. 1, p. 283-289, 2024.

JUNIOR, F. R. L.; CARPINETTI, L. C. R. **Comparação entre os métodos Fuzzy TOPSIS e Fuzzy AHP no apoio à tomada de decisão para seleção de fornecedores**. 2013. Dissertação (Mestrado em Processos e Gestão de Operações) - Escola de Engenharia de São Carlos, University of São Paulo, São Carlos, 2013. doi:10.11606/D.18.2013.tde-12092013-103003. Acesso em: 2025-07-01.

KAUR, P.; STOLTZFUS, J.; YELLAPU, V. Descriptive statistics. **International Journal of Academic Medicine**, v. 4, n. 1, p. 60-63, 2018.

KOTLER, P. **Administração de marketing.** 15 ed. São Paulo: Pearson Education do Brasil, 2019.

KUMAR, N. Intelligent customer segmentation: unveiling consumer patterns with machine learning. **Journal of Umm Al-Qura University for Engineering and Architecture**, p. 1-10, 2025.

KUMAR, S.; RANI, R.; PIPPAL, S. K.; AGRAWAL, R. Customer segmentation in e-commerce: K-means vs hierarchical clustering. **TELKOMNIKA Telecommunication Computing Electronics and Control**, v. 23, n. 1, p. 119-128, 2025. DOI: 10.12928/TELKOMNIKA.v23i1.26384.

LI, Z.; ZHANG, C. Decision research of plain lake remediation based on entropy-weight TOPSIS and modified DBSCAN. *In*: **International Conference on Computer Vision, Application, and Design (CVAD 2021)**. SPIE, 2021. p. 185-189.

LING, L. S.; WEILING, C. T. Enhancing Segmentation: A Comparative Study of Clustering Methods. **IEEE**, 2025.

LING, S. S.; TOO, C. W.; WONG, W. Y.; HOO, M. H. Customer Relationship Management System for Retail Stores using Unsupervised Clustering Algorithms with RFM Modeling for Customer Segmentation. **IEEE**, 2024.

LIU, A. H.; POON, L. K. M.; ZHANG, N. L. Unidimensional clustering of discrete data using latent tree models. *In*: **Proceedings of the National Conference on Artificial Intelligence**, 2015.

LIU, D.; SHIH, Y. Integrating AHP and data mining for product recommendation based on customer lifetime value. **Information & Management**, v. 42, n. 3, p. 387–400, 2005.

LIU, X.; YU, H.; WNAG, G.; GUO, L. A multi-criteria ordered clustering algorithm based on PROMETHEE. *In*: **Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020)**. 2020. p. 43-51.

MADANCHIAN, M.; TAHERDOOST, H. A comprehensive guide to the TOPSIS method for multi-criteria decision making. **Sustainable Social Development**, v. 1, n. 1, p. 2220, 2023.

MAHDIRAJI, H. A.; ZAVADSKAS, E. K.; KAZEMINIA, A.; KAMARDI, A. A. Marketing strategies evaluation based on big data analysis: a CLUSTERING-MCDM approach. **Economic research-Ekonomska istraživanja**, v. 32, n. 1, p. 2882-2892, 2019.

MAHMUD, M. R.; MAMUN, M. A.; HOSSAIN, M. A; UDDIN, M. P. Comparative analysis of K-means and bisecting K-means algorithms for brain tumor detection. In: **2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2)**. IEEE, p. 1-4, 2018.

MAJILYA, M.; MAJI, G.; GHOSH, P.; SEN, S. Online retail customer segmentation using RFM quantiles and clustering technique. In: **International Conference on Computer, Communication, Control and Information Technology (C3IT)**, 4., 2024, Kolkata. 2024. IEEE, 2024.

MARDANI, A., JUSOH, A., MD NOR, K., KHALIFAH, Z., ZAKWAN, N., & VALIPOUR, A. Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014. **Economic Research-Ekonomska Istraživanja**, v. *28*, n. 1, p. 516–571, 2015. https://doi.org/10.1080/1331677X.2015.1075139

MARSILI, F.; BÖDEFELD, J. Integrating Cluster Analysis into Multi-Criteria Decision Making for Maintenance Management of Aging Culverts. **Mathematics**, v. 9, n. 20, 2021.

MELO, D. P.; SILVA, L. G. O.; BATISTA, K. K. R.. Uso de métodos multicritério para ordenação de clusters: uma revisão de literatura. *In*: Encontro Nacional de Engenharia de Produção – ENEGEP, 2024, 44., Porto Alegre. **Anais eletrônicos.** Porto Alegre, 2024. p. 1-15. Disponível em: https://www.abepro.org.br/biblioteca/TN_WPG_411_2014_47806.pdf. Acesso em: 11 jan. 2025.

MEYER, P.; OLTEANU, A. L., Formalizing and solving the problem of clustering in MCDA. **European Journal of Operational Research,** v. 227, n. 3, p. 494–502, 2013.

MOHAMMADHOSSEIN, N.; ZAKARIA, N. H. CRM benefits for customers: literature review (2005-2012). **International Journal of Engineering Research and Applications**, v. 2, n. 6, p. 1578-1586, 2012.

MONALISA, S.; KURNIA, F. Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour. **Telkomnika (Telecommunication Computing Electronics and Control)**, v. 17, n. 1, p. 110-117, 2019.

MOZAFARI, A.; ZEIAEI, S.; MOZAFFARI, A. Segmentation of Public Library Clients Based on Their Lifetime Values and RFM Model Combining Multi-Criteria Decision Making and Data Mining Techniques. **Journal of Studies in Library and Information Science**, v. 7, n. 2, p. 19-38, 2016.

NEMERY, P.; DE SMET, Y. Multicriteria Ordered Clustering. **Université Libre de Bruxelles**, p. 1-17, 2005.

ODU, G. O. Weighting methods for multi-criteria decision making technique. **Journal of Applied Sciences and Environmental Management**, v. 23, n. 8, p. 1449-1457, 2019.

OMURBEK, N.; AKCAKAYA, O.; URMAKAKCAKAYA, E.D. Integrating cluster analysis with mcdm methods for the evaluation of local agricultural production. Croatian **Operational Research Review**, v. 12, n. 2, p. 105–117, 2021.

OYEWOLE, G. J.; THOPIL, G. A. Data clustering: application and trends**. Artif Intell Rev,** v. 56, p. 6439–6475, 2023.

OZKAN, P.; KOCAKOC, I. D. A customer segmentation model proposal for retailers: RFM-V. **University of South Florida (USF) M3 Publishing**, v. 5, n. 2021, p. 104, 2021.

PANDEY, V.; KOMAL & DINCER, H. A review on TOPSIS method and its extensions for different applications with recent development. **Soft Computing**, v. 27, n. 23, p. 18011-18039, 2023.

PATIBANDLA, K. K.; DARUVURI, R.; MANNEM, P. Enhancing online retail insights: K-means clustering and PCA for customer segmentation. *In*: **International Conference on Advancement in Computation and Computer Technologies (InCACCT)**, 3, 2025. IEEE, 2025.

PERNY, P. Multicriteria filtering methods based onconcordance and non-discordance principles. **Annals of operations Research**, v. 80, n. 0, p. 137-165, 1998.

POON, L.; LIU, A. H.; ZHANG, N. L. UC-LTM: Unidimensional clustering using latent tree models for discrete data. **International Journal of Approximate Reasoning.** v. 92, n. 1, p. 392-409, 2018.

QI, Y.; LAI, F; CHEN, G.; GAN, W. F-RFM-Miner: an efficient algorithm for mining fuzzy patterns using the recency-frequency-monetary model. **Applied Intelligence**, v. 53, n. 22, p. 27892-27911, 2023.

RAMKUMAR, G.; BHUVANESWARI, J.; VENUGOPAL, S.; KUMAR, S.; RAMASAMY, C. K.; KARTHICK, R. Enhancing customer segmentation: RFM analysis and K-Means clustering implementation. *In*: **Hybrid and Advanced Technologies**. CRC Press, 2025, p. 70-76.

REZAEINIA, S. M.; KERAMATI, A.; ALBADVI, A. An integrated AHP-RFM method to banking customer segmentation. **International Journal of Electronic Customer Relationship Management**, v. 6, n. 1, p. 37–52, 2012.

ROCHA, C. M. M.; BENITEZ, A. S.; BUELVAS, D. A. Review and Bibliographic Analysis of Metaheuristic Methods in Multicriteria Decision-Making: A 45-Year Perspective Across International, Latin American, and Colombian Contexts, **Journal of Applied Mathematics**, v. 2024, p. 1-17, 2024.

ROHILLA, V.; KUMAR, S. S.; CHAKRABORTY, S.; SINGH, S. Data clustering using bisecting k-means. *In*: **2019 international conference on computing, communication, and intelligent systems (ICCCIS)**. IEEE, 2019, p. 80-83.

ROSENFELD, J.; SMET, Y. An extension of PROMETHEE to hierarchical multicriteria clustering. **International Journal of Multicriteria Decision Making**, v. 8, n. 2, p. 133-150, 2019.

ROY, B. Classement et choix en présence de points de vue multiples (La méthode ELECTRE). **Revue Française d'Informatique et de Recherche Opérationnelle**, Paris, v. 8, n. 8, p. 57-75, 1968.

ROY, B. **Multicriteria methodology for decision aiding.** Dordrecht/Boston: Kluwer Academic Publishers, 1996.

SAATY, T. L. **The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation**. New York: McGraw-Hill, 1980.

SHARMA, H.; SINGH, J. P.; GORE, S.; SATHAWANE, N.; GOLHAR, R. S.; SWAMI, V. M. Evolution of Big Data Adoption in Industries Using TOPSIS Method. *In*: **International Conference on Sustainable Communication Networks and Application, Icscna 2024**, 5., 2024, Theni, p. 680–686, 2024.

SHARMA, V.; AGARWAL, P.; SHAIKH, H. Y.; LENKA, R. M.; MANJHI, S. K. Smart next-generation revenue growth: A methodology for partitioning customers utilizing the K-means algorithm and RFM model. *In*: **International Conference on Smart Devices (ICSD)**, 2024. IEEE, 2024.

SHIH, H. S.; SHYUR, H. J.; LEE, E. Stanley. An extension of TOPSIS for group decision making. **Mathematical and computer modelling**, v. 45, n. 7-8, p. 801-813, 2007.

SILVA, L. G. O.; BATISTA, K. K. R.; MELO, D. P.; PEREIRA, M. M. A. Análise de clusterização ordinal multicritério baseado no método TOPSIS. *In*: Encontro Nacional de Engenharia de Produção – ENEGEP, 2024, 44., Porto Alegre. **Anais eletrônicos.** Porto Alegre, 2024. p. 1-15. Disponível em: https://www.abepro.org.br/biblioteca/TN_ST_413_2030_47411.pdf. Acesso em: 12 fev. 2025.

SITHI, S. S.; ARA, M. A.; DHRUBO, A. T.; RONY, A. H.; SHABUR, M. A. Sustainable supplier selection in the textile industry using triple bottom line and SWARA-TOPSIS approaches. **Discover Sustainability**, v. 6, n. 1, 2025.

SOLICHIN, A.; WIBOWO, G. Customer segmentation based on recency frequency monetary (RFM) and user event tracking (UET) using K-means algorithm. *In*: **IEEE 8th Information Technology International Seminar (ITIS),** *2022*. Proceedings [...]. Jakarta: IEEE, 2022.

SONG, M.; ZHONG, H. Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers. **Bioinformatics**, v. 36, n. 20, p. 5027–5036, 2020.

STEINHAUS, H. Sur la division des corps matériels en parties. **Bulletin L'Académie Polonaise des Science**, v. 4, p. 801-804, 1957.

SUN, X., WANG, D. Conflict analysis of disputes in livelihood vulnerability assessment of flood using fuzzy TOPSIS method and GMCR with triangular fuzzy numbers. **Scientific Reports,** v. 15, n. 1, p. 1-18, 2025.

SUNDARI, Agus; LYDIA, Maya Silvi; MUCHTAR, Muhammad Anggia. Customer Segmentation Based on Recency, Frequency, Monetary, Variety and Duration (RFMVD). In: **2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA)**. IEEE, 2024. p. 1-5.

SWINDIARTO, V. T. P.; SARNO, R.; NOVITASARI, D. C. R. Integration of fuzzy C-means clustering and TOPSIS (FCM-TOPSIS) with silhouette analysis for multi criteria parameter data. *In*: **2018 International Seminar on Application for Technology of Information and Communication**. IEEE, 2018, p. 463-468.

TABATABAEI, S. A new model for evaluating the impact of organizational culture variables on the success of knowledge management in organizations using the TOPSIS multi-criteria algorithm: Case study. **Computers in Human Behavior Reports**, v. 14, 2024.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to data mining.** Boston: Pearson Addison Wesley, 2006.

WANG, H.; SONG, M. Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming**. The R Journal**, v. 3, n. 2, p. 29-33, 2011.

YEH, I.; YANG, K.; TING, T. Knowledge discovery on RFM model using Bernoulli sequence. **Expert Systems with applications**, v. 36, n. 3, p. 5866-5871, 2009.

YOSEPH, F., ALMALAILY, M., MALIM, N. New market segmentation methods using enhanced (RFM), CLV, modified regression and *cluster*ing methods. **International Journal of Computer Science and Information Technology**, v. 11, p. 43-60, 2019.

ZAHERI, F.; FARUGHI, H.; SOLTANPANAH, H.; ALANIAZAR, S.; NASERI, F. Using multiple criteria decision making models for ranking customers of bank network based on loyalty properties in weighted RFM model. **Management Science Letters**, v. 2, n. 2, p. 697-704, 2012.

ZHOU, X.; ZHANG, Z.; LU, Y. Review of Customer Segmentation method in CRM. *In*: **2011 International Conference on Computer Science and Service System (CSSS)**. IEEE, 2011, p. 4033-4035.

APÊNDICE A – CÓDIGO EM PYTHON DO PRÉ-PROCESSAMENTO E CLUSTERIZAÇÃO DA TABELA RFM COM BASE NO MÉTODO K-MEANS

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans


# Importando o arquivos com os dados de venda
df = pd.read_excel("Online Retail.xlsx")
df
df.describe()
# Visualizando o tipo dos dados do DataFrame
df.dtypes
# Acessando a quantidade de dados não faltantes
print(df.count())

# Filtrando os dados que contém CostumerID
df2 = df[df['CustomerID'].notnull() & (df['CustomerID'] != '')]
print(df2.count())

# Filtrando os dados que não são do Reino Unido
df3 = df2[df2["Country"]== 'United Kingdom']
print(df3.count())

# Criando uma variável adicional chamada Amount, calculando o produto
entre UnitPrice e Quantity
df['Amount'] = df['UnitPrice']*df['Quantity']
print(df)

# Quantidade de dados duplicados
qtd_duplicadas = df3.duplicated().sum()
print(qtd_duplicadas)
# Eliminando os dados duplicados
df4 = df3.drop_duplicates()
print(df4)

# Separando datas da hora de compra
df4['date'] = df4['InvoiceDate'].dt.date
df4['time'] = df4['InvoiceDate'].dt.time

# Mantendo apenas os dados de transação a partir de 2011
from datetime import date
df5 = df4[df4['date'] >=date(2011,1,1)]

# Eliminando os dados inconsistentes
df6 = df5[(df5['Quantity'] > 0) & (df5['UnitPrice'] > 0)]

# Inserindo a variável Amount no novo dataframe
```

```
df6['Amount'] = df6['UnitPrice']*df5['Quantity']

# Obtendo o Valor Monetário do RFM
monetary_series = df6.groupby('CustomerID')['Amount'].sum()
print(monetary_series)
# obtendo o valor da frequência do RFM
frequency_series = df6.groupby('CustomerID')['InvoiceNo'].nunique()
print(frequency_series)

data_base = df6['InvoiceDate'].max()
ultima_compra = df6.groupby('CustomerID')['InvoiceDate'].max()
recency_series = (data_base - ultima_compra).dt.days
print(recency_series)
recency_series.value_counts()
frequency_series.value_counts()

# Criando um Dataframe das três variáveis RFM
rfm = pd.concat([recency_series, frequency_series, monetary_series],
axis=1)
rfm.columns = ['Recency', 'Frequency', 'Monetary']
print(rfm.head())

# Dividindo a coluna Recency por 30 para gerar os valores em meses
rfm['Recency'] = rfm['Recency'] /30
rfm['Recency'] = rfm['Recency'].astype(int)
print(rfm)
rfm.describe()

# Gerando o valor X para clusterização
X = rfm.values
print(X)

# Normalizando os dados
scaler = StandardScaler()
X_normalizado = scaler.fit_transform(X)
print(X_normalizado)

# Gerando a curva do cotovelo para escolha do número de clusters
inercia = []
intervalo_k = range(1, 11)
for k in intervalo_k:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(X_normalizado)
    inercia.append(kmeans.inertia_)

# 4. Plotar a curva
plt.figure(figsize=(8, 5))
plt.plot(intervalo_k, inercia, marker='o')
plt.title('Curva do Cotovelo para Determinação de k')
plt.xlabel('Número de Clusters (k)')
plt.ylabel('Inércia')
plt.xticks(intervalo_k)
plt.grid(True)
plt.tight_layout()
```

```
plt.show()

# Adicionando a coluna cluster (k=3) na data frame rfm_normalizado
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
rotulos = kmeans.fit_predict(X_normalizado)
rfm_cluster = rfm.copy()
rfm_cluster['cluster'] = rotulos
print(rfm_cluster)

# Gerando as estatísticas por cluster
print('Estatísticas cluster 0 para kmeans:')
print(rfm_cluster[rfm_cluster['cluster']==0].describe())
print("_____________________________________________________")
print('Estatísticas cluster 1 para kmeans:')
print(rfm_cluster[rfm_cluster['cluster']==1].describe())
print("_____________________________________________________")
print('Estatísticas cluster 2 para kmeans:')
print(rfm_cluster[rfm_cluster['cluster']==2].describe())

# Gerando os Boxplots
sns.set(style="whitegrid")

# Gráfico 1: Recency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_cluster, x='cluster', y='Recency', palette='Set3')
plt.title('Boxplot de Recency por Cluster')
plt.xlabel('cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_cluster, x='cluster', y='Frequency', palette='Set2')
plt.title('Boxplot de Frequency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_cluster, x='cluster', y='Monetary', palette='Set1')
plt.title('Boxplot de Monetary por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()

# Gráfico de barras dos dados antes da remoção dos outliers
sns.set(style='whitegrid')

# Gráfico de barras da Recência
plt.figure(figsize=(8, 5))
```

```
sns.histplot(rfm_cluster['Recency'], bins=30, kde=True, color='skyblue')
plt.title('Histograma - Recência)
plt.xlabel(Recência)
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

sns.set(style='whitegrid')

# Gráfico de barras da Frequência
plt.figure(figsize=(8, 5))
sns.histplot(rfm_cluster['Frequency'], bins=30, kde=True, color='skyblue')
plt.title('Histograma - Frequência)
plt.xlabel(Frequência')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

sns.set(style='whitegrid')

# Gráfico de barras do Valor Monetário
plt.figure(figsize=(8, 5))
sns.histplot(rfm_cluster['Monetary'], bins=30, kde=True, color='skyblue')
plt.title('Histograma - monetary')
plt.xlabel('monetary')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

# Filtro para remover outliers, delimitando a 1%
limite_superior = rfm.quantile(0.99)

# Filtra os dados: remove apenas os valores acima do limite superior
rfm_filtrado = rfm[(rfm < limite_superior).all(axis=1)]

# Mostra o resultado
print("Shape original:", rfm.shape)
print("Shape após remoção de outliers superiores:", rfm_filtrado.shape)

# normalizando rfm_filtrado
scaler = StandardScaler()
X_normalizado_filtrado = scaler.fit_transform(rfm_filtrado.values)
print(X_normalizado)

# Gerando a curva do cotovelo para escolha do número de clusters
inercia = []
intervalo_k = range(1, 11)
for k in intervalo_k:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(X_normalizado_filtrado)
    inercia.append(kmeans.inertia_)

# 4. Plotar a curva de Elbow
plt.figure(figsize=(8, 5))
```

```
plt.plot(intervalo_k, inercia, marker='o')
plt.title('Determinação de k d K-Means')
plt.xlabel('Número de Clusters (k)')
plt.ylabel('Inércia')
plt.xticks(intervalo_k)
plt.grid(True)
plt.tight_layout()
plt.show()

# Adicionando a coluna cluster (k=3) na data frame rfm_normalizado
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
rotulos = kmeans.fit_predict(X_normalizado_filtrado)
rfm_filtrado['cluster'] = rotulos
print(rfm_filtrado)

rfm_filtrado.describe()

# Gerando os Boxplots# Definir o estilo do Saborn
sns.set(style="whitegrid")

# Gráfico 1: Recency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_filtrado, x='cluster', y='Recency', palette='Set3')
plt.title('Recência - K-Means')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_filtrado, x='cluster', y='Frequency', palette='Set2')
plt.title('Frequência - K-Means')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=rfm_filtrado, x='cluster', y='Monetary', palette='Set1')
plt.title('Valor Monetário - K-Means')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()

# Gráfico de barras das variáveis pós-remoção de outliers
sns.set(style='whitegrid')

plt.figure(figsize=(8, 5))
sns.histplot(rfm_filtrado['Recency'], bins=30, kde=True, color='skyblue')
plt.title('Histograma - Recency')
plt.xlabel('Recency')
```

```
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

sns.set(style='whitegrid')

plt.figure(figsize=(8, 5))
sns.histplot(rfm_filtrado['Frequency'], bins=30, kde=True,
color='skyblue')
plt.title('Histograma - frequency')
plt.xlabel('frequency')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

sns.set(style='whitegrid')

plt.figure(figsize=(8, 5))
sns.histplot(rfm_filtrado['Monetary'], bins=30, kde=True, color='skyblue')
plt.title('Histograma - monetary')
plt.xlabel('monetary')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

# Gerando as estatísticas descritivas por cluster
print('Estatísticas cluster 1 para kmeans:')
print(rfm_filtrado[rfm_filtrado['cluster']==0].describe())
print("_____________________________________________________")
print('Estatísticas cluster 2 para kmeans:')
print(rfm_filtrado[rfm_filtrado['cluster']==1].describe())
print("_____________________________________________________")
print('Estatísticas cluster 0 para kmeans:')
print(rfm_filtrado[rfm_filtrado['cluster']==2].describe())

rfm_filtrado.to_excel('rfm_resultado.xlsx', index=False)

# Separando os clusters 0, 1 e 2
c1 = rfm_filtrado[rfm_filtrado['cluster'] == 0]
c2 = rfm_filtrado[rfm_filtrado['cluster'] == 1]
c3 = rfm_filtrado[rfm_filtrado['cluster'] == 2]

# Criar subplots separados Recência, Frequência e Valor Monetário
fig2, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(14, 6))

sns.boxplot(data=rfm_filtrado, x='cluster', y='Recency', ax=ax1)
ax1.set_title('Recência - K-Means')
ax1.set_xlabel('cluster')
ax1.set_ylabel('Frequency')

sns.boxplot(data=rfm_filtrado, x='cluster', y='Frequency', ax=ax2)
ax2.set_title('Frequência - K-Means')
ax2.set_xlabel('cluster')
ax2.set_ylabel('Frequency')
```

```
sns.boxplot(data=rfm_filtrado, x='cluster', y='Monetary', ax=ax3)
ax3.set_title('Valor Monetário - K-Means')
ax3.set_xlabel('cluster')
ax3.set_ylabel('Monetary')

# Mostrar os gráficos
plt.tight_layout()
plt.show()
```

APÊNDICE B – CÓDIGO EM PYTHON DO MODELO TOPSIS-CKMEANS SOBRE A TABELA RFM

```
# Importando as principais bibliotecas utilizadas nos algoritmos abaixo
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math

1.    Definindo a classe TOPSIS

class Topsis:
    def __init__(self, matriz_de_decisão, pesos, tipo_criterio,
normalização, R_D, R_tipo):
        self.matriz_de_decisão = matriz_de_decisão
        self.pesos = pesos
        self.tipo_criterio = tipo_criterio
        self.normalização = normalização
        self.R_D = R_D
        self.R_tipo = R_tipo

    def mostrar_matriz(self):
        print(self.matriz_de_decisão)

    def mostrar_pesos(self):
        print(self.pesos)

    def mostrar_tipo_criterio(self):
        print(self.tipo_criterio)

    def normalizar_dados_tradicional(self):
        pesos2 = np.array(self.pesos)
        matriz = self.matriz_de_decisão.copy()
        (numero_de_linhas, numero_de_colunas) = matriz.shape
        matriz_quadrada = np.square(matriz)
        soma_coluna = np.sum(matriz_quadrada, axis=0)
        print(soma_coluna)
        print(type(soma_coluna))
        soma_coluna_raiz = np.sqrt(soma_coluna)
        matriz_normalizada =
np.zeros((numero_de_linhas,numero_de_colunas))
        for i in range(numero_de_linhas):
            for j in range(numero_de_colunas):
                matriz_normalizada[i][j] =
(pesos2[j]*matriz[i][j])/soma_coluna_raiz[j]
        return matriz_normalizada

    def normalizar_dados_tradicional2(self):
       pesos2 = np.array(self.pesos)
       matriz = self.matriz_de_decisão.copy()
       (numero_de_linhas, numero_de_colunas) = matriz.shape
       soma_coluna_raiz = np.zeros((numero_de_colunas))
```

```
        for j in range(numero_de_colunas):
            soma_coluna_raiz[j] = np.sqrt(soma_coluna[j])
        matriz_normalizada = np.zeros((numero_de_linhas,numero_de_colunas))
        for i in range(numero_de_linhas):
            for j in range(numero_de_colunas):
                matriz_normalizada[i][j] =
(pesos2[j]*matriz[i][j])/soma_coluna_raiz[j]
        return matriz_normalizada

    def normalizar_dados_Rtopsis(self):
        pesos = np.array(self.pesos)
        matriz = self.matriz_de_decisão.copy()
        D2 = self.R_D.copy()
        R_tipo2 = self.R_tipo
        r_matriz = np.zeros(matriz.shape)
        for i in range(matriz.shape[0]):
            for j in range(matriz.shape[1]):
                if R_tipo2 == 'Max':
                    r_matriz[i,j] = (matriz[i,j]*pesos[j])/D2[j,1]
                if R_tipo2 == 'Max_Min':
                    r_matriz[i,j] = ((matriz[i,j]-
D2[j,0])*pesos[j])/(D2[j,1]-D2[j,0])
        matriz_normalizada = r_matriz
        return matriz_normalizada


    def normalizar_dados_hassanli(self):
        matriz = self.matriz_de_decisão.copy()
        tipo = self.tipo_criterio
        peso = self.pesos
        (numero_de_linhas, numero_de_colunas) = matriz.shape
        vetor = np.zeros(numero_de_colunas)
        matriz_normalizada1 = np.zeros((numero_de_linhas,
numero_de_colunas))
        for j in range(numero_de_colunas):
            vetor[j] = np.max(matriz[:,j])
        for i in range(numero_de_linhas):
            for j in range(numero_de_colunas):
                matriz_normalizada1[i,j] = (peso[j]*matriz[i,j])/vetor[j]
        return  matriz_normalizada1

    def normalizar_dados_maximo(self):
        matriz = self.matriz_de_decisão.copy()
        tipo = self.tipo_criterio
        peso = self.pesos
        (numero_de_linhas, numero_de_colunas) = matriz.shape
        maximo = np.max(matriz)
        matriz_normalizada1 = np.zeros((numero_de_linhas,
numero_de_colunas))
        for i in range(numero_de_linhas):
            for j in range(numero_de_colunas):
                if  tipo[j] == 'opt':
                    matriz_normalizada1[i,j] =
peso[j]*(matriz[i,j])/maximo
```

```
                if  tipo[j] == 'min':
                    matriz_normalizada1[i,j] = peso[j]*(1-
(matriz[i,j]/maximo))
        return  matriz_normalizada1


    def normalizar_dados(self):
        normalização = self.normalização
        if normalização == 'tradicional':
            matriz_normalizada = self.normalizar_dados_tradicional2()
        if normalização == 'hassanali':
            matriz_normalizada = self.normalizar_dados_hassanli()
        if normalização == 'R_TOPSIS':
            matriz_normalizada = self.normalizar_dados_Rtopsis()

        return matriz_normalizada

    def melhor_alternativa_ideal(self):
        matriz_normalizada = self.normalizar_dados()
        tipo_de_criterio =  self.tipo_criterio
        numero_de_linhas, numero_de_colunas = matriz_normalizada.shape
        maiores_valores = np.amax(matriz_normalizada, axis=0)
        menores_valores = np.amin(matriz_normalizada, axis=0)
        melhor_alternativa = np.zeros(numero_de_colunas)
        count = 0
        normalização = self.normalização
        pesos = self.pesos
        D = self.R_D
        for i in range(numero_de_colunas):
            if tipo_de_criterio[i] == 'opt':
                if normalização == 'R_TOPSIS':
                    melhor_alternativa[i] = pesos[i]
                else:
                    melhor_alternativa[i] = maiores_valores[i]
            if tipo_de_criterio[i] == 'min':
                if normalização == 'R_TOPSIS':
                    melhor_alternativa[i] = (D[i,0]/D[i,1])*pesos[i]
                else:
                    melhor_alternativa[i] = menores_valores[i]
        return melhor_alternativa

    def pior_alternativa_ideal(self):
        matriz_normalizada = self.normalizar_dados()
        tipo_de_criterio =  self.tipo_criterio
        numero_de_linhas, numero_de_colunas = matriz_normalizada.shape
        maiores_valores = np.amax(matriz_normalizada, axis=0)
        menores_valores = np.amin(matriz_normalizada, axis=0)
        pior_alternativa = np.zeros(numero_de_colunas)
        count = 0
        normalização = self.normalização
        pesos = self.pesos
        D = self.R_D
        for i in range(numero_de_colunas):
            if tipo_de_criterio[i] == 'opt':
```

```
                if normalização == 'R_TOPSIS':
                    pior_alternativa[i] = (D[i,0]/D[i,1])*pesos[i]
                else:
                    pior_alternativa[i] = menores_valores[i]
            if tipo_de_criterio[i] == 'min':
                if normalização == 'R_TOPSIS':
                    pior_alternativa[i] = pesos[i]
                else:
                    pior_alternativa[i] = menores_valores[i]
                pior_alternativa[i] = maiores_valores[i]
        return pior_alternativa

    def euclidian_distance(self):
        matriz_normalizada = self.normalizar_dados()
        melhor_alternativa = self.melhor_alternativa_ideal()
        pior_alternativa = self.pior_alternativa_ideal()
        numero_de_linhas, numero_de_colunas = matriz_normalizada.shape
        matriz_distancia = np.zeros((numero_de_linhas, 2))
        for i in range(numero_de_linhas):
            dist_melhor = np.sqrt(np.sum(np.square(matriz_normalizada[i]-
melhor_alternativa)))
            dis_pior = np.sqrt(np.sum(np.square(matriz_normalizada[i]-
pior_alternativa)))
            matriz_distancia[i,0] = dist_melhor
            matriz_distancia[i,1] = dis_pior
        return matriz_distancia

    def desempenho_global(self):
        matriz_de_distância = self.euclidian_distance()
        numero_de_linhas = matriz_de_distância.shape[0]
        vetor_indice_desempenho = np.zeros((numero_de_linhas,2))
        for i in range(numero_de_linhas):
            vetor_indice_desempenho[i,1] =
matriz_de_distância[i,1]/(matriz_de_distância[i,0]+matriz_de_distância[i,1
])
            vetor_indice_desempenho[i,0] = i
        return vetor_indice_desempenho

    def vetor_desempenho_global(self):
      matriz_desempenho_global = self.desempenho_global()
      numero_de_linhas = matriz_desempenho_global.shape[0]
      vetor_desempenho = np.zeros(numero_de_linhas)
      for j in range(numero_de_linhas):
        vetor_desempenho[j] =  matriz_desempenho_global[j,1]
      return vetor_desempenho


    def ranking(self):
        vetor_desempenho = self.desempenho_global()
        vetor_desempenho_ordenado = sorted(vetor_desempenho, key=lambda
vetor: vetor[1] , reverse=True)
        return vetor_desempenho_ordenado
```

```
# Carregando a matriz de decisão
df_matriz_decisão = pd.read_excel('rfm_resultado.xlsx')
matriz_de_decisão = df_matriz_decisão.values

# Definição dos pesos:
pesos = [0.333, 0.333, 0.333]

# Informação para Rtopsis (benchmarketing) - D(2xn) = [valor_mínimo,
valor_máximo]
R_D = np.array([[0, 1], [0, 1], [0, 1]])

# Tipo de normalização (Max ou Max_Min)
R_tipo = 'Max'

# Definição do tipo de critério - opt: otimização/min: minimização
tipo = ['min', 'opt', 'opt']
df_matriz_decisão

# Normalização da matriz de decisão
pesos2 = np.array(pesos)
matriz = matriz_de_decisão.copy()
(numero_de_linhas, numero_de_colunas) = matriz.shape
matriz_quadrada = np.square(matriz)
print(matriz_quadrada)
soma_coluna = np.sum(matriz_quadrada, axis=0)
print("_____________________________________________________")
print(soma_coluna)
print("_____________________________________________________")
soma_coluna_raiz = np.zeros((numero_de_colunas))
for j in range(numero_de_colunas):
    soma_coluna_raiz[j] = np.sqrt(soma_coluna[j])
matriz_normalizada = np.zeros((numero_de_linhas,numero_de_colunas))
print(soma_coluna_raiz)
print("_____________________________________________________")
for i in range(numero_de_linhas):
    for j in range(numero_de_colunas):
        matriz_normalizada[i][j] =
(pesos2[j]*matriz[i][j])/soma_coluna_raiz[j]
print(matriz_normalizada)
```

2. Usando o modelo

```
# Observe a definição da normalização
topsis_tradicional = Topsis(matriz_de_decisão = matriz_de_decisão , pesos
= pesos,
                    tipo_criterio = tipo, normalização = 'tradicional', R_D
= R_D , R_tipo=R_tipo)

# Checando o carregamento dos dados
print(topsis_tradicional.mostrar_matriz())
print("_____________________________________________________________")
print(topsis_tradicional.mostrar_pesos())
print("_____________________________________________________________")
print(topsis_tradicional.mostrar_tipo_criterio())
```

```
# Dados normalizados
topsis_tradicional.normalizar_dados()

# Alternativa ideal positiva
topsis_tradicional.melhor_alternativa_ideal()

# Alternativa ideal negativa
topsis_tradicional.pior_alternativa_ideal()

# Usando o método euclidian_distance(). Esse método calcula uma matriz
distância cujos valores da primeira coluna representam a distância a
melhor alternativa ideal e os valores da segunda coluna representam a
distância das alternativas a pior alternativa ideal.
topsis_tradicional.euclidian_distance()

# Usando o método desempenho_global(). Calcula uma matriz contendo os
índices de desempenho global das alternativas.
topsis_tradicional.desempenho_global()

# Usando o método desempenho_global(). Calcula um vetor contendo os
índices de desempenho global das alternativas.
topsis_tradicional.vetor_desempenho_global()

# Ranking das alternativas
topsis_tradicional.ranking()
```

3. Para definição do cluster, foi usado o pacote ckmeans_1d_dp

```
# Importando os módulo para a definição dos clusters
%pip install ckmeans-1d-d

from ckmeans_1d_dp import ckmeans

# Calculando o cluster unidimensional.
# k define o número de clusters

clus = ckmeans(topsis_tradicional.vetor_desempenho_global(), k=3)

print(clus.cluster)
print(clus.withinss)
print(clus.tot_withinss)
print(clus.centers)
clus.cluster

for item in clus.cluster:
    print(item)

df_matriz_decisão['Cluster'] = clus.cluster
df_matriz_decisão

c1 = df_matriz_decisão[df_matriz_decisão['Cluster'] == 0]
c2 = df_matriz_decisão[df_matriz_decisão['Cluster'] == 1]
c3 = df_matriz_decisão[df_matriz_decisão['Cluster'] == 2]
```

```
def media_harominica(centroides):
  distancias = np.zeros(centroides.shape[0]-1)

  dist = 0

  for j in range(1,centroides.shape[0]):
      distancias[j-1] = centroides[j] - centroides[j-1]

  maxima_distancia = centroides[centroides.shape[0]-1] - centroides[0]
  minima_distancia = np.min(distancias)

  m1 = maxima_distancia
  minima_padrao = 1/(centroides.shape[0]-1)
  m2 = minima_distancia/minima_padrao
  media_harmonica = (2*m1*m2)/(m1+m2)
  return media_harmonica

print(media_harominica(clus.centers))

clus = ckmeans(topsis_tradicional.vetor_desempenho_global(), k=3)
print("Retorna os clusters:")
print(clus.cluster)
print("_____________________________________")
print("Retorna a média de cada cluster:")
print(clus.centers)
print("_____________________________________")
print("Retorna a soma dos quadrados de cada cluster:")
print(clus.withinss)
print("_____________________________________")
print("Retorna o número de elementos de cada cluster:")
print(clus.size)
print("_____________________________________")
print("Retorna a soma dos quadrados totais entre os elementos e a média da
amostra:")
print(clus.totss)
print("_____________________________________")
print("Retorna a soma dos quadrados totais entre os elementos e a média
dos cluster:")
print(clus.tot_withinss)
print("_____________________________________")
print("Retorna a soma dos quadrados totais entre as médias do clusters e a
média da amostra :")
print(clus.betweenss)
print("_____________________________________")

# Construindo a curva de Elbow
l=[]
for k in range(1, 20):
    clus = ckmeans(topsis_tradicional.vetor_desempenho_global(), k=k)
    l.append(clus.tot_withinss)

plt.figure(figsize=[15,6])
plt.title("Curva de Elbow - TOPSIS-Ckmeans", fontsize = 20, fontweight =
'bold')
```

```
plt.xlabel("Número de clusters", fontsize = 14, fontweight = 'bold')
plt.ylabel("WCSS (inertia)", fontsize = 14, fontweight = 'bold')
plt.plot(range(1, 1+len(l)), l, c = 'black', lw = 2)
plt.scatter(range(1, 1+len(l)), l, c = 'black', lw = 2)
plt.xticks(range(1,1+len(l)))
plt.grid()
plt.show()

# Dividindo os clusters
c1 = df[df['Cluster'] == 0]
c2 = df[df['Cluster'] == 1]
c3 = df[df['Cluster'] == 2]

# Estatística descritivas dos clusters
print('Estatísticas cluster 1 para TOPSIS-Ckmeans:')
print(c2.describe())
print('_____
_____')
print('Estatísticas cluster 2 para TOPSIS-Ckmeans:')
print(c3.describe())
print('_____
_____')
print('Estatísticas cluster 0 para TOPSIS-Ckmeans:')
print(c1.describe())

# Gerando os Boxplots
sns.set(style="whitegrid")

# Gráfico 1: Recency por Cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Cluster', y='Recency', palette='Set3')
plt.title('Recência - TOPSIS-Ckmeans')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por Cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Cluster', y='Frequency', palette='Set2')
plt.title('Frequência - TOPSIS-Ckmeans')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por Cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Cluster', y='Monetary', palette='Set1')
plt.title('Valor Monetário - TOPSIS-Ckmeans')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()
```

```
# Mostrar os gráficos
plt.tight_layout()
plt.show()
```

APÊNDICE C – CÓDIGO EM PYTHON DA MATRIZ DE COMPARAÇÃO

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

df1 = pd.read_excel('clusters_kmeans_confusionmatrix.xlsx')
print(df1)
df2 = pd.read_excel('ckmeans_clusters_confusionmatrix.xlsx')
print(df2)

# Estatísticas descritivas dos clusters do K-Means
print('Estatísticas cluster 0 para kmeans:')
print(df1[df1['Cluster']==0].describe())
print("_____________________________________________________")
print('Estatísticas cluster 1 para kmeans:')
print(df1[df1['Cluster']==1].describe())
print("_____________________________________________________")
print('Estatísticas cluster 2 para kmeans:')
print(df1[df1['Cluster']==2].describe())

# Estatísticas descritivas dos clusters do TOPSIS-Ckmeans
print('Estatísticas cluster 0 para topsiscluster:')
print(df2[df2['Cluster']==0].describe())
print("_____________________________________________________")
print('Estatísticas cluster 1 para topsiscluster:')
print(df2[df2['Cluster']==1].describe())
print("_____________________________________________________")
print('Estatísticas cluster 2 para topsiscluster:')
print(df2[df2['Cluster']==2].describe())

df1[["Frequency", "Cluster"]]

# Boxplot para  visualizar os dados do KMeans
sns.set(style="whitegrid")

# Gráfico 1: Recency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1, x='Cluster', y='Recency', palette='Set3')
plt.title('Boxplot de Recency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1, x='Cluster', y='Frequency', palette='Set2')
plt.title('Boxplot de Frequency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
```

```
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1, x='Cluster', y='Monetary', palette='Set1')
plt.title('Boxplot de Monetary por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()

# Boxplot para os dados do TopsisCluster
sns.set(style="whitegrid")

# Gráfico 1: Recency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df2, x='Cluster', y='Recency', palette='Set3')
plt.title('Boxplot de Recency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df2, x='Cluster', y='Frequency', palette='Set2')
plt.title('Boxplot de Frequency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df2, x='Cluster', y='Monetary', palette='Set1')
plt.title('Boxplot de Monetary por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()

# Alterando o dataframe df1 para ter a mesma ordem dos bloxplot do TOPSIS-
Ckmeans
mapeamento_cluster = {2: 0, 1: 2, 0: 1}

# Criando um novo DataFrame com os clusters substituídos
df1_novo = df1.copy()
df1_novo['Cluster'] = df1_novo['Cluster'].replace(mapeamento_cluster)

# Exibir os primeiros registros do novo DataFrame
print(df1_novo.head())

# Boxplot para os dados df1_novo
```

```
sns.set(style="whitegrid")

# Gráfico 1: Recency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1_novo, x='Cluster', y='Recency', palette='Set3')
plt.title('Boxplot de Recency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.tight_layout()
plt.show()

# Gráfico 2: Frequency por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1_novo, x='Cluster', y='Frequency', palette='Set2')
plt.title('Boxplot de Frequency por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary por cluster
plt.figure(figsize=(8, 5))
sns.boxplot(data=df1_novo, x='Cluster', y='Monetary', palette='Set1')
plt.title('Boxplot de Monetary por Cluster')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.tight_layout()
plt.show()

# Suponha que você tenha dois DataFrames:
# rfm1 e rfm2, com as colunas: Recency, Frequency, Monetary, cluster

# Passo 1: Adicionar coluna identificadora
df1_novo['Fonte'] = 'Kmeans'
df2['Fonte'] = 'Topsis-Ckmeans'

# Passo 2: Concatenar os dois DataFrames
rfm_comparado = pd.concat([df1_novo, df2], ignore_index=True)

# Passo 3: Criar os 3 gráficos separados por métrica

# Estilo dos gráficos
sns.set(style="whitegrid")

# Gráfico 1: Recency
plt.figure(figsize=(10, 5))
sns.boxplot(data=rfm_comparado, x='Cluster', y='Recency', hue='Fonte',
palette='pastel')
plt.title('Recência por cluster')
plt.xlabel('Cluster')
plt.ylabel('Recency')
plt.legend(title='Fonte')
plt.tight_layout()
plt.show()
```

```
# Gráfico 2: Frequency
plt.figure(figsize=(10, 5))
sns.boxplot(data=rfm_comparado, x='Cluster', y='Frequency', hue='Fonte',
palette='Set2')
plt.title('Frequência por cluster')
plt.xlabel('Cluster')
plt.ylabel('Frequency')
plt.legend(title='Fonte')
plt.tight_layout()
plt.show()

# Gráfico 3: Monetary
plt.figure(figsize=(10, 5))
sns.boxplot(data=rfm_comparado, x='Cluster', y='Monetary', hue='Fonte',
palette='Set3')
plt.title('Valor Monetário por cluster')
plt.xlabel('Cluster')
plt.ylabel('Monetary')
plt.legend(title='Fonte')
plt.tight_layout()
plt.show()


from sklearn.metrics import confusion_matrix

# Gerando a matriz de comparação com base nos dois métodos
# Extraindo os clusters
y_true = df1_novo['Cluster']
y_pred = df2['Cluster']

# Gerar matriz de comparação
matriz_confusao = confusion_matrix(y_true, y_pred)

# Transformar em DataFrame para visualização mais clara
matriz_df = pd.DataFrame(
    matriz_confusao,
    index=[f'K_means_{i}' for i in sorted(y_true.unique())],
    columns=[f'Topsis_Ckmeans_{i}' for i in sorted(y_pred.unique())]
)

# Plotar a matriz
plt.figure(figsize=(8, 6))
sns.heatmap(matriz_df, annot=True, fmt='d', cmap='Blues')
plt.title('Matriz de Comparação')
plt.xlabel('Cluster Topsis_Ckmeans')
plt.ylabel('Cluster K_means')
plt.tight_layout()
plt.show()
```