LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

APOIO À TOMADA DE DECISÃO EM UMA REDE VAREJISTA: UMA ABORDAGEM HÍBRIDA ENTRE MACHINE LEARNING NÃO SUPERVISIONADO E APOIO MULTICRITÉRIO À DECISÃO

Adilson Vilarinho Terra[1,a], Claudio de Souza Rocha Junior[2,a], Carlos Francisco Simões Gomes[3,a], Marcos dos Santos[4,b]
[a]Universidade Federal Fluminense (UFF), R. Passo da Pátria, 156 - São Domingos, Niterói – RJ
[b]Instituto Militar de Engenharia (IME), Praça Gen. Tibúrcio, 80 - Urca, Rio de Janeiro – RJ
[1]adilsonvilarinho@id.uff.br, [2]cs_junior@id.uff.br, [3]cfsg1@bol.com.br,
[4]marcosdossantos@ime.eb.br

RESUMO

Este artigo apresenta uma modelagem híbrida para gerenciamento de pontos de venda de uma rede varejista, integrando técnicas de *machine learning* não supervisionadas com o Apoio Multicritério à Decisão (AMD). Integrou-se o método não hierárquico de clusterização *K-means* e o método AMD AHP-Gaussiano. Aplicou-se a modelagem proposta em uma rede varejista do segmento farmacêutico com mais de 70 pontos de venda na cidade do Rio de Janeiro. Academicamente, o trabalho contribui para a literatura em modelagem híbrida, enquanto no âmbito prático, oferece uma ferramenta estratégica para redes varejistas, melhorando sua eficiência operacional e competitividade de mercado. A modelagem proposta combina o poder do *machine learning*, especialmente em identificar semelhanças comportamentais entre observações e variáveis de um banco de dados, com a eficácia do AMD em avaliar e priorizá-los. Esta abordagem não apenas facilita o processo decisório, mas também contribui para o desenvolvimento de estratégias de marketing mais direcionadas e eficazes.

PALAVRAS CHAVE. Apoio Multicritério à Decisão, Machine Learning, Varejo.

Tópicos: Apoio Multicritério à Decisão e Machine Learning.

ABSTRACT

This paper presents a hybrid modeling approach for managing retail outlets in a retail network, integrating unsupervised machine learning techniques with Multi-Criteria Decision Support (MCDM). The non-hierarchical K-means clustering method and the MCDM AHP-Gaussian method were integrated. The proposed modeling was applied to a pharmaceutical retail network with over 70 outlets in Rio de Janeiro city. Academically, this work contributes to the literature on hybrid modeling, while in a practical context, it offers a strategic tool for retail networks, enhancing their operational efficiency and market competitiveness. The proposed model combines the power of machine learning, particularly in identifying behavioral similarities among observations and variables in a database, with the effectiveness of MCDM in evaluating and prioritizing them. This approach not only facilitates the decision-making process but also contributes to the development of more targeted and effective marketing strategies.

KEYWORDS. Multi Criteria Decision Making, Machine Learning, Retail.

Paper topics: Multi Criteria Decision Making and Machine Learning.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

1. Introdução

O setor varejista, caracterizado por sua complexidade operacional e grande volume de dados, enfrenta desafios significativos na tomada de decisões estratégicas. A natureza dinâmica e competitiva do mercado varejista impõe uma pressão constante para aprimorar a eficiência operacional e otimizar as estratégias de gestão de pontos de venda. A gestão eficaz desses estabelecimentos requer uma abordagem analítica robusta para lidar com o grande volume de dados disponíveis. Nesse contexto, a utilização de técnicas de análise de dados, como Apoio Multicritério à Decisão (AMD) (MAÊDA *et al*., 2021), e a análise de cluster (CHENG; CHEN, 2009) emergem como ferramentas essenciais para auxiliar na avaliação criteriosa de múltiplas variáveis envolvidas na tomada de decisões.

O AMD é uma abordagem metodológica que se destaca pela capacidade de lidar com a complexidade inerente a decisões que envolvem múltiplos critérios e variáveis (SANTOS; COSTA; GOMES, 2021). O AMD consiste em um vasto conjunto de métodos matemáticos projetados para incorporar ativamente diversos critérios na avaliação de múltiplas alternativas em problemas de decisão. A abordagem multicritério facilita um processo de decisão mais consciente (BELTON; STEWART, 2002). Além disso, os métodos multicritério são fundamentais para tornar o processo decisório mais sistemático, lógico e eficaz (HATAMI-MARBINI; TAVANA, 2011). Assim, o AMD oferece uma estrutura sólida para avaliar opções complexas e diversas, permitindo que os tomadores de decisão compreendam melhor as implicações e os *trade-offs* de suas escolhas em diferentes cenários e contextos.

A clusterização, ou análise de cluster, é uma técnica não supervisionada de *machine learning* (técnica exploratória) que busca identificar padrões e estruturas intrínsecas em grandes conjuntos de dados (ONER; OZTAYSI, 2018). O objetivo da clusterização é agrupar as observações de um banco de dados em grupos homogêneos internamente e heterogêneos entre si. Ao agrupar observações similares, a análise de *cluster* facilita a compreensão de estruturas subjacentes e a segmentação de dados complexos (WANG *et al*., 2009). A análise de cluster é uma ferramenta fundamental na análise exploratória de dados, revelando grupos latentes que podem dar apoio na tomada de decisões estratégicas (GÜÇDEMIR; SELIM, 2015). No contexto varejista, a análise de cluster pode ser vital na identificação de perfis de consumidores, comportamentos de compra e características específicas de pontos de venda, fornecendo insights para a gestão eficiente desses estabelecimentos. As técnicas não supervisionadas de aprendizagem de máquina são recomendadas quando se deseja analisar e extrair padrões de conjuntos de dados severamente volumosos (KARMITSA; BAGIROV; TAHERI, 2018).

A combinação das técnicas de clusterização e do AMD apresenta-se como uma abordagem híbrida factível para enfrentar os desafios decisórios no varejo. Nesse trabalho, utilizou-se uma abordagem de modelagem híbrida, mesclando o método não supervisionado de machine learning intitulado K-means e o método multicritério AHP-Gaussiano. O principal objetivo com a clusterização é agrupar pontos de venda em grupos internamente homogêneos e heterogêneos entre si, utilizando o poder analítico do machine learning para descobrir padrões e tendências ocultas nos dados dos PDVs. O AHP-Gaussiano auxiliará na avaliação e priorização desses PDVs com base em múltiplos critérios e com base na estratégia a ser utilizada pela rede varejista. Este modelo híbrido visa oferecer uma ferramenta para varejistas no gerenciamento de seus pontos de venda.

2. Fundamentação Teórica

No processo de tomada de decisão, é comum enfrentar a necessidade de escolher entre diversas alternativas, as quais devem ser viáveis para atingir o objetivo proposto. Essas alternativas são então avaliadas e comparadas com base em critérios específicos, influenciados por diferentes atributos (EKER, 2022). Nesse contexto, os métodos de AMD desempenham um papel fundamental, pois incorporam julgamentos de valor, além de considerar aspectos técnicos, para avaliar alternativas e solucionar problemas práticos. Dessa forma, os métodos de AMD são caracterizados por uma abordagem altamente multidisciplinar (ALAMOODI *et al*., 2023).

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

2.1 Método AHP-Gaussiano

O método AHP-Gaussiano consiste em uma derivação do tradicional AHP (SANTOS; COSTA; GOMES, 2021). Segundo os autores, ele se baseia em uma análise de sensibilidade utilizando o fator Gaussiano, eliminando a necessidade de avaliação por pares ou uso da escala de Saaty para definir os pesos de cada critério. Por meio desse método, os pesos podem ser determinados a partir dos dados quantitativos das alternativas em cada critério da própria matriz de decisão.

O primeiro passo do método AHP-Gaussiano consiste na elaboração da matriz de decisão. Se houver um conjunto de *n* alternativas e *m* critérios, a matriz de decisão D pode ser representada pela equação 1.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nm} \end{bmatrix} \quad (1)$$

Após a criação da matriz de decisão, é imprescindível proceder com a normalização. Essa etapa tem como objetivo equalizar as diferentes escalas dos critérios, viabilizando comparações equitativas entre eles, conforme demonstrado pela equação 2.

$$d_{ij}^{norm} = \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}} \quad (2)$$

Agora, com base na matriz de decisão normalizada, conforme definida pela equação 2, procede-se ao cálculo da média μ_j para cada critério *j*, cuja expressão é apresentada na equação 3.

$$u_j = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^{norm} \quad (3)$$

O desvio padrão de cada critério é calculado a partir dos valores normalizados das alternativas. Dentro do contexto do método AHP-Gaussiano, o desvio padrão é empregado no cálculo do fator Gaussiano, o qual é essencial para determinar o peso de cada critério. A equação 4 apresenta a representação do desvio padrão para o critério *j*.

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (d_{ij}^{norm} - u_j)^2} \quad (4)$$

Para cada critério, é calculado o fator Gaussiano, que é determinado pela divisão do desvio padrão pela média dos valores normalizados correspondentes a esse critério. A equação 5 apresenta a representação do fator Gaussiano para o critério *j*.

$$G_j = \frac{\sigma_j}{\mu_j} \quad (5)$$

Em seguida, no procedimento do método, o fator Gaussiano para cada critério *j* é normalizado dividindo-o pela soma dos fatores de todos os critérios, como exemplificado pela equação 6.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

$$w_j = \frac{G_j}{\sum_{k=1}^{m} G_k} \tag{6}$$

Nesta fase do método, é realizado o processo de ponderação da matriz de decisão com os pesos normalizados obtidos na etapa anterior, como demonstrado pela equação 7.

$$D_{ij}^{pond} = d_{ij}^{norm} \times w_j \tag{7}$$

A fase final do método AHP-Gaussiano envolve a obtenção do ranking das alternativas. Para gerar o ranking, é necessário calcular a pontuação global para cada alternativa, conforme representado pela equação 8, onde a pontuação global R_i é obtida pela soma das pontuações ponderadas em todos os critérios m para cada alternativa *i*.

$$R_i = \sum_{j=1}^{m} D_{ij}^{pond} \tag{8}$$

Depois de calcular a pontuação global para cada alternativa, procede-se à classificação das alternativas com base nessas pontuações. A alternativa que obtém a pontuação global mais alta é considerada a mais preferível ou adequada, seguida pela alternativa com a segunda maior pontuação, e assim por diante. Como resultado, obtém-se um ranking ou lista ordenada de todas as alternativas, organizadas da mais preferível para a menos preferível, com base na análise multicritério realizada.

2.1 Método K-means

Sua abordagem é fundamentada na minimização da soma dos quadrados das distâncias entre os pontos de dados e os centroides dos clusters. Em primeiro lugar, o algoritmo seleciona os centroides iniciais e, em seguida, atribui cada ponto de dados ao cluster cujo centroide está mais próximo (O'NEILL *et al*., 2019). Posteriormente, os centroides são recalculados com base na média dos pontos pertencentes a cada cluster. Esse processo de atribuição e atualização dos centroides é repetido iterativamente até que ocorra a convergência, isto é, quando não houver mais alterações nos clusters ou quando o critério de parada pré-definido for alcançado. O método K-means é amplamente utilizado em uma variedade de aplicações, como segmentação de mercado, análise de imagens, processamento de sinais e reconhecimento de padrões (KOIZUMI; WIDDERSHEIM, 2019).

Para utilizar o K-means, é necessário realizar a padronização das variáveis antes de iniciar a análise, garantindo que todas estejam na mesma escala. Uma técnica comumente utilizada para padronização é o Z-Score. Como é possível observar na equação 9, o Z-Score é calculado utilizando média aritmética e o desvio padrão de cada variável.

$$ZX_{ji} = \frac{X_{ji} - \bar{X}_j}{S_j} \tag{9}$$

Após a realização da padronização pelo Z-Score, inicia-se o esquema de agloração. O K-means possui um esquema de aglomeração não-hierárquico, ou seja, a quantidade K de clusters é escolhida a priori, sendo usada de base para a identificação dos centros de aglomeração, de modo que as observações são arbitrariamente alocadas aos K clusters para o cálculo dos centroides iniciais. Nas etapas seguintes, as observações vão sendo comparadas pela proximidade aos centroides dos outros clusters. Se houver realocação a outro cluster por estar mais próximo, os centroides devem ser recalculados (em ambos os clusters). Por isso, trata-se de um processo iterativo.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

Existem diversos métodos distintos para a escolha do número de clusters a priori (HANCER; KARABOGA, 2017). Neste trabalho, utilizou-se o Método de Elbow para a escolha do número de clusters a priori. O Método de Elbow é um dos métodos clássicos para a escolha do número de clusters a priori para o conjunto de dados analisado. Sua ideia central é especificar K=2 como o número de clusters inicial e, em seguida, aumentar K em etapas de 1 até o máximo especificado para o número potencial de clusters adequados estimados e, finalmente, distinguir o número potencial de clusters K correspondente. O número adequado de clusters K é obtido pelo fato de que antes de alcançar K, o custo diminui rapidamente para o chamado valor máximo de custo e, após exceder K, continua a aumentar com o valor máximo de custo quase inalterado, com um ponto de "cotovelo" explícito.

O procedimento K-means encerra-se quando não for possível realocar qualquer observação por estar mais próxima do centroide de outro cluster: indica que a soma dos quadrados de cada ponto até o centro do cluster alocado foi minimizada. A soma total dos quadrados dentro dos clusters é representada pela equação 10.

$$SS = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{10}$$

Após a conclusão do método K-means, é essencial examinar se os valores de uma variável quantitativa específica no conjunto de dados diferem entre os grupos definidos, isto é, se a variabilidade entre os clusters é significativamente maior do que a variabilidade dentro de cada cluster. Para isso, o teste F da análise de variância de um fator (conhecido como *one-way analysis of variance ou one-way* ANOVA) pode ser utilizado. Sua hipótese nula e alternativa são formuladas da seguinte maneira:

a) H_0: a média da variável em estudo é a mesma em todos os grupos formados.
b) H_1: a média da variável em estudo difere em pelo menos um dos grupos em comparação com os outros.

Portanto, um único teste F é aplicado para cada variável para avaliar se há pelo menos uma diferença entre todas as possíveis comparações. Uma vantagem significativa desse teste é que não são necessários ajustes em relação a discrepâncias nas dimensões dos grupos para realizar diversas comparações. A rejeição da hipótese nula a um nível de significância específico não indica qual grupo é estatisticamente diferente dos demais em relação à variável em estudo, porém indica que aquela variável foi estatisticamente relevante para a formação de pelo menos um cluster. A estatística F correspondente a esse teste é expressa pela equação 11.

$$F = \frac{\frac{\sum_{k=1}^{k} N_k \,.\, (\bar{X}_k - \bar{X})^2}{K - 1}}{\frac{\sum_{ki} (X_{ki} - \bar{X}_k)^2}{n - K}} \tag{11}$$

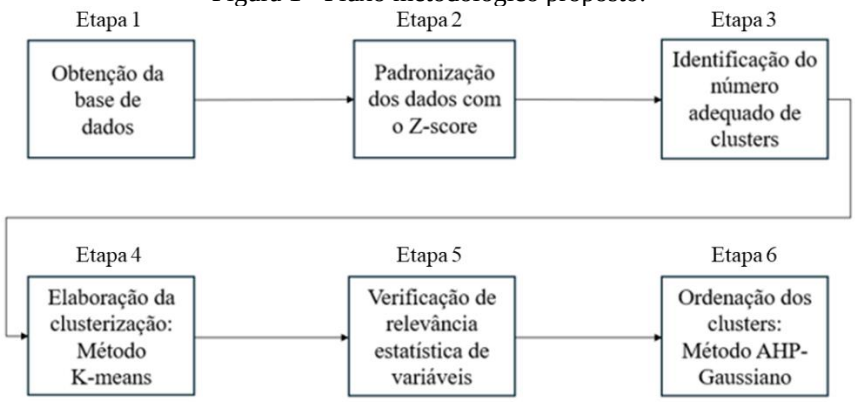
Neste contexto, N representa o número de observações no *k-ésimo* cluster, enquanto $\overline{X}_k$ representa a média da variável *X* nesse mesmo cluster. $\overline{X}$ é a média geral da variável *X*, e X_{ki} indica o valor assumido pela variável *X* para uma determinada observação *i* no *k-ésimo* cluster. Além disso, *K* representa o número de clusters a serem comparados, e *n* indica o tamanho da amostra. De maneira resumida, a estatística F é a divisão da variabilidade entre clusters pela variabilidade dentro dos clusters. Por meio da estatística F, é possível identificar as variáveis cujas médias diferem significativamente entre os grupos, destacando aquelas que mais contribuem para a formação de pelo menos um dos *K* clusters. Além disso, é possível identificar as variáveis que não influenciam significativamente na formação dos clusters sugeridos, com base em um determinado nível de significância.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

3. Metodologia

Este estudo pode ser caracterizado como uma abordagem mista qualitativa-quantitativa, que combina elementos de estudo de caso com modelagem matemática (CRESWELL; CRESWELL, 2017; WILL M. BERTRAND; FRANSOO, 2002). O foco da pesquisa está no desenvolvimento de uma abordagem híbrida entre os métodos AHP-Gaussiano e K-means. Os detalhes específicos do estudo de caso realizado em uma rede varejista de farmácias são abordados no estudo de caso. O fluxo metodológico que guia a abordagem híbrida proposta, está representado na figura 1.

Figura 1 - Fluxo metodológico proposto.



4. Estudo de Caso

Aplicou-se o método AHP-Gaussiano e o K-means de maneira híbrida para agrupar 72 farmácias em clusters com comportamentos semelhantes e depois ordená-los de acordo com a estratégia de marketing da rede varejista, visando apoiar o processo de tomada de decisão da rede no que tange a operação de múltiplos PDVs com características distintas entre si.

A Tabela 1 sintetiza os critérios, ou seja, as variáveis do banco de dados da rede varejista, disponíveis para analisar as 72 farmácias (alternativas) deste estudo.

Tabela 1 – Definição das variáveis (critérios) disponíveis no banco de dados para o estudo.

Critério	Definição
Calçada	Número diário de pessoas que passam em frente à cada uma das 72 farmácias da rede.
Interior	Número diário de pessoas que circulam pelo interior de cada uma das 72 farmácias da rede.
Caixa	Número diário de pessoas que vão ao caixa ou *check-out* de cada uma das 72 farmácias da rede.
Adesão	Taxa diária de adesão de cada uma das 72 farmácias da rede. É a divisão da variável "Interior" pela variável "Calçada". Representa a taxa de pessoas que entra na farmácia em relação às que passam na calçada.
Transações	Número diário de transações realizado no caixa de cada uma das 72 farmácias analisadas.
Faturamento	Faturamento diário em reais de cada uma das 72 farmácias analisadas.
SKU	*Stock Keeping Unit* (SKU) representa quantos tipos de produtos diferentes cada uma das farmácias possui em seu mix.
Atendimento	Variável métrica que varia de zero a cem para cada farmácia. Consiste em uma avaliação ou nota dada pelos clientes para o atendimento de cada uma das 72 farmácias.
Sortimento	Variável métrica que varia de zero a cem para cada farmácia. Consiste em uma avaliação ou nota dada pelos clientes para o sortimento de cada uma das 72 farmácias.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

Organização	Variável métrica que varia de zero a cem para cada farmácia. Consiste em uma avaliação ou nota dada pelos clientes para a organização de cada uma das 72 farmácias.

Com a definição dos critérios na tabela 1, foi obtida a base de dados disponível para o estudo de caso. Ao todo, foram analisadas 72 farmácias da cidade do Rio de Janeiro, localizadas em bairros distintos, avaliadas à luz dos dez critérios estabelecidos na tabela 1. A tabela 2 apresenta as primeiras dez linhas da base de dados disponível, contendo o valor dos critérios para cada uma das farmácias analisadas. Essa base de dados será submetida ao método K-means nas etapas subsequentes, visando agrupar as mais de 70 farmácias em clusters antes da análise multicritério com o método AHP-Gaussiano.

Tabela 2 – Dez primeiras linhas do banco de dados utilizado no estudo.

Loja	Calcada	Interior	Caixa	Adesão	Transações	Faturamento	SKU	Atendimento	Sortimento	Organização
1	17080	2603	431	0,1524	370	21405	7037	17	8	2
2	10838	1621	369	0,1495	391	17635	7220	22	10	7
3	13649	2609	299	0,1911	362	22766	8223	4	17	10
4	15563	1469	427	0,0943	305	16694	8461	13	10	20
5	12193	2352	313	0,1928	394	25109	7434	6	23	12
6	11437	1879	315	0,1642	322	18277	7602	13	4	6
7	8653	2793	405	0,3227	275	18545	6824	4	19	21
8	10414	2497	432	0,2397	303	16552	7608	15	2	5
9	18144	1974	316	0,1087	272	23341	8456	9	6	24
10	5314	1982	367	0,3729	326	25961	7031	22	6	2

Na análise dos dados coletados para o presente estudo, identificou-se que cada variável da base de dados está expressa em uma unidade de medida distinta, por isso, as variáveis precisam ser padronizadas. Essa diversidade de escalas pode distorcer os resultados das análises estatísticas, levando a interpretações equivocadas sobre as relações entre as variáveis. Para garantir a comparabilidade dos dados, padronizou-se com o Z-Score (equação 9). As dez primeiras linhas da base de dados padronizada pelo Z-Score estão representadas na figura 2.

Figura 2 – Base de dados padronizada pelo *Z-Score*.

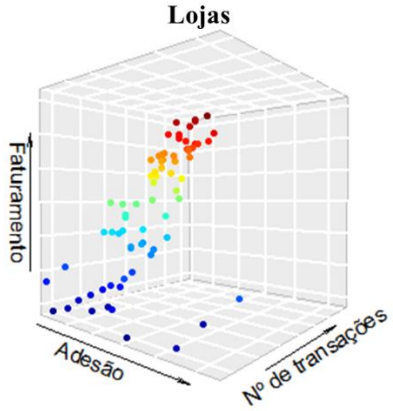
	Calcada	Interior	Caixa	Adesão	Transações	Faturamento	SKU	Atendimento	Sortimento	Organização
1	-0.9463107351	-0.95274815	-0.99886349	0.039170943	-1.08643405	-1.28049394	-1.29862527	-1.10008710	-1.28682181	-1.428888700
2	-1.3398423719	-1.47415460	-1.36682244	-0.018201330	-0.95316386	-1.56678704	-1.25596969	-0.93637013	-1.21633785	-1.263593092
3	-1.1626207393	-0.94956237	-1.78225997	0.823582304	-1.13720364	-1.17713986	-1.02217985	-1.52575123	-0.96964398	-1.164415727
4	-1.0419511538	-1.55486109	-1.02260277	-1.135144979	-1.49893699	-1.63824640	-0.96670429	-1.23106068	-1.21633785	-0.833824511
5	-1.2544153665	-1.08602006	-1.69917246	0.858968296	-0.93412527	-0.99921288	-1.20608831	-1.46026444	-0.75819208	-1.098297484
6	-1.3020779614	-1.33716594	-1.68730282	0.279882058	-1.39105161	-1.51803368	-1.16692909	-1.23106068	-1.42778974	-1.296652214
7	-1.4775973586	-0.85186503	-1.15316685	3.488189519	-1.68932297	-1.49768181	-1.34827356	-1.52575123	-0.89916001	-0.800765390
8	-1.3665737744	-1.00903032	-0.99292867	1.807892425	-1.51162939	-1.64902985	-1.16553055	-1.16557389	-1.49827371	-1.329711335
9	-0.8792300460	-1.28672438	-1.68136800	-0.843524079	-1.70836157	-1.13347447	-0.96786975	-1.36203426	-1.35730578	-0.701588025
10	-1.6881071528	-1.28247667	-1.37869208	4.504382274	-1.36566681	-0.93451216	-1.30002381	-0.93637013	-1.35730578	-1.428888700

O próximo passo, antes de aplicar o K-means, é identificar o número adequado de clusters. Nesse estudo, utilizou-se o Método de *Elbow* para a definição do número de clusters. De maneira preliminar, gerou-se uma visualização em três dimensões para a os dados padronizados, representando variáveis chave como faturamento, adesão e número de transações nas farmácias. O intuito é verificar se é possível identificar visualmente um padrão que oriente a definição do número de clusters, antes do Método de *Elbow*. Com isso, observou-se padrões de agrupamento naturais, indicativos de comportamentos de mercado semelhantes entre as farmácias. Essa visualização também facilita a detecção de *outliers*, os quais podem representar casos atípicos de

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

comportamento de uma farmácia. Portanto, gerou-se um gráfico em três dimensões com as variáveis “Faturamento”, “Adesão” e “Transações”, representado na figura 3.

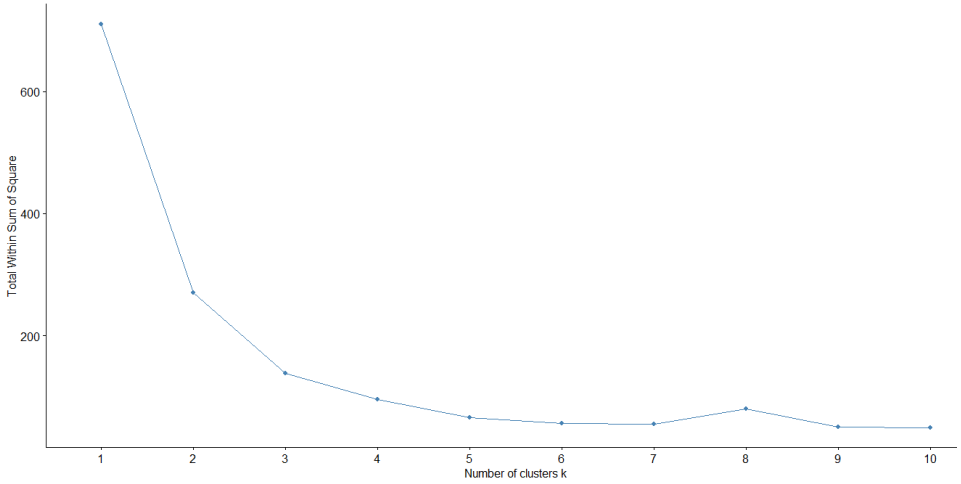
Figura 3 – Visualização das lojas em três dimensões.



A partir dessa visualização em três dimensões, pode-se identificar alguns padrões de agrupamento entre as farmácias. De maneira inicial, observa-se a formação de aproximadamente seis clusters e também algumas farmácias com comportamentos mais distintos das demais.

O Método de *Elbow* envolve a plotagem dos valores de uma métrica de distorção ou inércia, que representa a soma das distâncias ao quadrado dos pontos de dados até o centro de seu respectivo cluster, em função do número de clusters. À medida que o número de clusters aumenta, a inércia total tende a diminuir, pois os pontos estarão, em média, mais próximos de seus centros de cluster. O número adequado de clusters é identificado no lugar onde a taxa de decréscimo muda abruptamente, indicando que adicionar mais clusters além desse ponto não traz benefícios significativos em termos de redução da inércia. Na figura 4 está representado o gráfico gerado a partir do método de *Elbow* para definir o número adequado de clusters para a análise dos dados das farmácias.

Figura 4 – Método de *Elbow* para identificação do número de *clusters*.



A análise do gráfico do método de *Elbow* guia a determinação do número adequado de clusters no presente estudo. A partir da análise da curva, nota-se que o declínio acentuado da inércia total se estabiliza após o ponto correspondente a quatro clusters, indicando que o aumento do número de clusters além desse ponto não resulta em uma diminuição substancial da inércia. Este fenômeno é identificado pelo "cotovelo" no gráfico, que ocorre no eixo x onde K=4. A presença deste "cotovelo" sugere que a clusterização das farmácias em quatro grupos distintos é suficiente para capturar a estrutura essencial dos dados, oferecendo um equilíbrio entre a precisão da clusterização e a simplicidade do modelo.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

Portanto, selecionar quatro clusters permite uma categorização adequada das farmácias, promovendo um agrupamento que reflete diferenças fundamentais nos dados disponíveis na base de dados sem subdividir excessivamente as farmácias em grupos menores que poderiam levar a uma sobreposição de características e a um aumento desnecessário da complexidade do modelo. Esta decisão metodológica para a definição do número de clusters embasa-se na procura por uma solução que não apenas destila a variabilidade dos dados, mas que também é prática e interpretável para fins analíticos e estratégicos subsequentes.

Com a definição do número adequado de clusters estabelecida em quatro, conforme sugerido pela análise do método de *Elbow*, o próximo passo metodológico no estudo envolve a implementação do algoritmo K-means para efetuar a clusterização das 72 farmácias. O K-means é um método de clusterização particional que tem por objetivo dividir o conjunto de dados em K grupos distintos, minimizando a variância intra-cluster e maximizando a variância inter-cluster. O K-means irá agrupar as farmácias de modo que as que pertencem a um mesmo cluster terão características homogêneas, enquanto farmácias em clusters diferentes exibirão atributos distintos em termos das variáveis analisadas no banco de dados.

Utilizando o K-means, conforme descrito na equação 10, procedeu-se à segmentação das farmácias nos quatro clusters distintos com base nas variáveis disponíveis na base de dados e descritas na tabela 1. A formação dos clusters seguiu uma distribuição que alinha as farmácias em conjuntos homogêneos: o primeiro cluster abrange as lojas de 1 a 18, o segundo engloba as lojas 19 a 36, o terceiro compreende as lojas 37 a 53, e o quarto e último cluster inclui as lojas 54 a 72.

A tabela 3 representa a média de cada uma das variáveis da tabela 1 separadas por cada um dos quatro clusters. A tabela 3 que será a tabela, ou matriz de decisão, que será submetida à avaliação multicritério pelo método AHP-Gaussiano.

Tabela 3 – Média das variáveis dentro de cada *cluster*.

Cluster K	Média Calcada	Média Interior	Média Caixa	Média Adesão	Média Transações	Média Faturamento	Média SKU	Média Atendimento	Média Sortimento	Média Organização
1	25959,22	3583,05	545,00	0,142	483,11	32262,33	10658,44	37,61	38,55	40,00
2	37623,59	5154,82	665,58	0,138	604,52	45311,88	14594,29	63,00	36,88	36,23
3	11591,22	1985,61	372,27	0,191	331,27	21042,72	7052,83	11,05	13,16	11,38
4	52366,53	6775,94	806,52	0,130	738,42	53969,89	17941,78	89,26	86,68	90,26

Com isso, após a clusterização das farmácias, a nova matriz de decisão passa a ter 4 alternativas e não mais 72, como no início da análise. Cada um dos quatro clusters formados é composto com farmácias de perfis similares em cada um dos atributos (critérios) analisados. Analisando a tabela 3, observa-se que o cluster 4, por exemplo, é composto pelas farmácias que obtiveram as maiores notas médias de Atendimento (89,26), Sortimento (86,68) e Organização (90,26). O cluster 4 também corresponde ao grupo de farmácias com maior média de Faturamento (R$53.969,89) e número médio de Transações (738,42). Entretanto, apesar do maior desempenho nos critérios citados anteriormente, o cluster 4 teve e menor taxa média de Adesão (13,0%), o que indica que a média de pessoa que passa na sua Calçada é alta (52.366,53), entretanto, uma taxa pequena de fato vai ao Interior da loja (6.77,94).

Seguindo para a etapa 5, que é a verificação da relevância estatística das variáveis, calculou-se a estatística F, conforme descrito na equação 11, para cada uma das dez variáveis desse estudo de caso. A estatística F consiste na divisão da variabilidade entre os clusters pela variabilidade dentro dos clusters. O objetivo de uma análise de cluster é fazer com que a variabilidade entre clusters seja maior do que a variabilidade dentro do cluster, dessa forma, agrupando observações homogêneas internamente e heterogêneas entre si. Portanto, quanto maior o valor da estatística F de cada variável, mas ela contribui para a formação de pelo menos um dos clusters.

Uma vez calculadas as estatísticas F de cada uma das dez variáveis, é possível analisar a rejeição ou não da hipótese nula dos testes F da análise de variância de um fator (conhecido como

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

one-way analysis of variance ou *one-way* ANOVA) para cada uma das variáveis. As hipóteses nula e alternativa utilizadas são:

a) H_0: a média da variável em estudo é a mesma em todos os grupos formados.

b) H_1: a média da variável em estudo difere em pelo menos um dos grupos em comparação com os outros.

Para o teste, foram utilizados três graus de liberdade para a variabilidade entre os grupos (K - 1 = 2) e sessenta e oito graus de liberdade para a variabilidade dentro dos grupos (n - K = 2), onde K é o número de clusters formados (4) e n é o número de farmácias totais (72). Para aceitar ou rejeitar a hipótese nula, deve-se verificar se o nível de significância da estatística F calculada é menor que 0,05 (5%). Se o nível de significância da estatística F calculada for menor que 0,05, rejeita-se a hipótese nula H_0 e aceitasse a hipótese alternativa H_1, concluindo que existe pelo menos uma diferença entre os grupos para a variável em análise. A Tabela 4 apresenta, os resultados da análise de variância de um fator (ANOVA), com as variabilidades de cada variável, as estatísticas F e os respectivos níveis de significância.

Tabela 4 – Análise de variância de um fator (ANOVA).

	Variabilidade entre *clusters*	Variabilidade dentro dos *clusters*	Estatística F calculada	Nível de significância de F
Calcada	21,957	0,075	291,1	2×10^{-16}
Interior	21,979	0,074	295,2	2×10^{-16}
Caixa	21,971	0,075	293,8	2×10^{-16}
Adesão	5,635	0,796	7,083	$3{,}3 \times 10^{-4}$
Transações	22,30	0,06	370	2×10^{-16}
Faturamento	22,141	0,067	328,9	2×10^{-16}
SKU	22,30	0,06	370,7	2×10^{-16}
Atendimento	22,229	0,063	350,5	2×10^{-16}
Sortimento	21,985	0,074	296,4	2×10^{-16}
Organização	22,227	0,064	350	2×10^{-16}

A tabela 4 apresenta reduzidos níveis de significância para o teste F. Valores reduzidos para o nível de significância de F em cada variável, mostram que que a probabilidade (p-valor) de obter um resultado tão extremo quanto o observado no teste, assumindo que a hipótese nula seja verdadeira, é menor do que o nível de significância de F da tabela 4. Isso é uma evidência forte contra a hipótese nula, sugerindo que o resultado do teste é estatisticamente significativo e as médias dos clusters analisados são estatisticamente significativas. Um baixo nível de significância indica que há uma diferença muito significativa entre as médias dos clusters que estão sendo comparados, e a hipótese nula H_0 de que todas as médias de grupo são iguais pode ser rejeitada com alta confiança, para todas as variáveis.

Portanto, nesse caso, conclui-se que para essa variável em questão, existe pelo menos um cluster que apresenta média estatisticamente diferente dos demais. A partir da tabela 10 é possível observar que todas as estatísticas F calculadas para cada uma das variáveis, é superior a 2,74 (F crítico). Por isso, rejeita-se a hipótese nula para todas as variáveis, e conclui-se que todas foram estatisticamente relevantes para a formação de pelo menos um dos quatro clusters.

A sexta e última etapa da metodologia proposta, é a ordenação dos clusters com o método multicritério AHP-Gaussiano. Uma vez que todos as variáveis (critérios) foram estatisticamente relevantes para a formação de pelo menos um dos clusters, decide-se não remover nenhuma variável da análise multicritério. Com isso, a tabela 3, com 4 alternativas e 10 critérios, será a matriz de decisão submetida ao método AHP-Gaussiano. Utilizou-se a plataforma computacional do método AHP-Gaussiano, disponível em https://decision-making.shinyapps.io/gaussian_ahp/ para a obtenção do ranking das alternativas, conforme a tabela 5. Para a obtenção do ranking, considerou-se todos os critérios como sendo monotônicos de lucro, ou seja, buscou-se maximizar o resultado dos dez critérios analisados.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

Tabela 5 – Ordenação das alternativas com o método AHP-Gaussiano.

Alternativa	Pontuação Global (AHP-Gaussiano)	Ranking
Cluster 4	0,4149854	1º
Cluster 2	0,2656235	2º
Cluster 1	0,2127837	3º
Cluster 3	0,1066074	4º

A tabela 5 apresenta a ordenação dos clusters resultante da aplicação do método AHP-Gaussiano. Os resultados indicam uma hierarquia na pontuação global das alternativas, com o cluster 4 posicionando-se na primeira posição com uma pontuação superior de 0,4149854. Este valor denota uma preferência distinta em relação aos outros clusters, conforme avaliado pelos critérios estabelecidos neste estudo. Em sequência, o cluster 2 segue com uma pontuação de 0,2656235, colocando-o em segundo lugar no ranking e manifestando uma performance menor que o cluster 4, mas ainda assim superior aos demais. O cluster 1 e o cluster 3, com pontuações globais de 0,2127837 e 0,1066074, respectivamente, são classificados em terceiro e quarto lugares.

A obtenção do ranking da tabela 5 resulta de um processo de maximização de todos os critérios estabelecidos, refletindo uma estratégia de priorização que visa a otimização das variáveis desse estudo, consideradas benéficas à performance das farmácias em questão. Destaca-se que tal ordenação é influenciada pela orientação adotada nos critérios de decisão. Caso o enfoque fosse alternado para a minimização de um ou mais critérios, por exemplo, reduzir o Faturamento ou o número de Transações, a hierarquia representada poderia sofrer alterações. A flexibilidade na manipulação desses critérios permite adaptar a análise em consonância com objetivos estratégicos e de marketing distintos da rede varejista farmacêutica em questão. Assim, cada cluster, representando um agrupamento diverso de farmácias, poderia ser reavaliado e reclassificado com base em tais objetivos variáveis, demonstrando a natureza dinâmica da modelagem de decisão e sua capacidade de alinhamento com metas estratégicas alternativas.

5. Considerações Finais

A partir do presente trabalho, conclui-se que a combinação de técnicas de *machine learning* não supervisionado com métodos de Apoio Multicritério à Decisão oferece uma abordagem factível para a otimização da gestão de PDV's em redes varejistas. Esta modelagem híbrida não apenas facilita a identificação de padrões significativos nos dados, mas também permite uma avaliação criteriosa e a priorização eficaz de estratégias de mercado. Destaca-se a aplicabilidade do modelo no contexto de uma rede de farmácias, onde foi possível agrupar as lojas em clusters homogêneos que refletem comportamentos e características específicas, orientando assim as decisões estratégicas da rede.

Além disso, os resultados obtidos reforçam a relevância da modelagem híbrida para fornecer insights operacionais e estratégicos que são cruciais para o sucesso no ambiente competitivo do varejo. As implicações práticas deste estudo são vastas, permitindo que gestores não só compreendam melhor o desempenho de seus pontos de venda, mas também implementem intervenções mais direcionadas. Isso demonstra que a integração de métodos quantitativos na análise de grandes volumes de dados pode transformar significativamente a eficiência operacional e a competitividade no mercado varejista, abrindo novas perspectivas para futuras pesquisas e aplicações práticas no setor.

Referências

MAÊDA, S. M. DO N. *et al*. Multi-criteria analysis applied to aircraft selection by Brazilian Navy. Production, v. 31, 2021.

CHENG, C.-H.; CHEN, Y.-S. Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications, v. 36, n. 3, p. 4176–4184, abr. 2009.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br

SANTOS, M. DOS; COSTA, I. P. DE A.; GOMES, C. F. S. MULTICRITERIA DECISION-MAKING IN THE SELECTION OF WARSHIPS: A NEW APPROACH TO THE AHP METHOD. International Journal of the Analytic Hierarchy Process, v. 13, n. 1, 19 maio 2021.

BELTON, V.; STEWART, T. J. Multiple Criteria Decision Analysis. Boston, MA: Springer US, 2002.

HATAMI-MARBINI, A.; TAVANA, M. An extension of the Electre I method for group decision-making under a fuzzy environment. Omega, v. 39, n. 4, p. 373–386, ago. 2011.

ONER, S. C.; OZTAYSI, B. An interval type 2 hesitant fuzzy MCDM approach and a fuzzy c means clustering for retailer clustering. Soft Computing, v. 22, n. 15, p. 4971–4987, 17 ago. 2018.

WANG, G. *et al*. Data Mining for Customer Segmentation in Personal Financial Market. In: [s.l: s.n.]. p. 614–621.

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, 13 jul. 2015.

KARMITSA, N.; BAGIROV, A. M.; TAHERI, S. Clustering in large data sets with the limited memory bundle method. Pattern Recognition, v. 83, p. 245–259, nov. 2018.

EKER, M. Determining the Appropriate Tool for In-Stand Debarking with Analytical Hierarchy Process. European Journal of Forest Engineering, v. 8, n. 2, p. 85–92, 30 dez. 2022.

ALAMOODI, A. H. *et al*. Systematic review of MCDM approach applied to the medical case studies of COVID-19: trends, bibliographic analysis, challenges, motivations, recommendations, and future directions. Complex & Intelligent Systems, v. 9, n. 4, p. 4705–4731, 3 ago. 2023.

O'NEILL, S. *et al*. Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline. Health Informatics Journal, v. 25, n. 4, p. 1722–1738, 17 dez. 2019.

KOIZUMI, M.; WIDDERSHEIM, M. M. Specialties and strategies in academic libraries: a cluster analysis approach. Library Management, v. 40, n. 1/2, p. 45–58, 14 jan. 2019.

HANCER, E.; KARABOGA, D. A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. Swarm and Evolutionary Computation, v. 32, p. 49–67, fev. 2017.

CRESWELL, J. W.; CRESWELL, J. D. Research design: Qualitative, quantitative, and mixed methods approaches. [s.l.] [s.l.] Sage publications, 2017.

WILL M. BERTRAND, J.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. International Journal of Operations & Production Management, v. 22, n. 2, p. 241–264, fev. 2002.

DOI: 10.59254/sbpo-2024-193920
https://proceedings.science/p/193920?lang=pt-br