eati XV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO
ANAIS DO XV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO - EATI
|ANO 13, N.1| NOVEMBRO/2025

Segmentação RFV com Clustering: Um Framework Comparativo para Análise de Marketing

Martony Demes da Silva[1], **Wellington Alves**[1]

[1]Universidade Federal do Piauí (UFPI)
Centro de Educação Aberta e a Distancia (CEAD)
Teresina – Piauí – Brasil

mardemes@gmail.com, wellingtonalves89@yahoo.com

Abstract. *This study presents a replicable framework for customer segmentation based on the RFM model (Recency, Frequency, Monetary). Unlike approaches that rely solely on K-Means, it incorporates multiple clustering algorithms evaluated through internal quality metrics (Silhouette, Calinski-Harabasz, Davies-Bouldin) and strategic business indicators (average ticket, repurchase rate, revenue). Results indicate that combining technical and strategic perspectives enhances the understanding of customer behavior and provides practical support for personalized marketing campaigns and efficient resource allocation.*

Resumo. *Este estudo apresenta um framework replicável de segmentação de clientes com base no modelo RFV (Recência, Frequência e Valor). Diferente de abordagens que utilizam apenas o K-Means, são considerados múltiplos algoritmos de clustering, avaliados por métricas internas de qualidade (Silhouette, Calinski-Harabasz e Davies-Bouldin) e por indicadores estratégicos de negócio (ticket médio, taxa de recompra e receita). Os resultados mostram que a integração dessas perspectivas amplia a compreensão do comportamento dos clientes e oferece subsídios práticos para campanhas personalizadas e melhor alocação de recursos.*

1. Introdução

A segmentação de clientes é central para compreender padrões de consumo e orientar estratégias personalizadas [Wedel and Kamakura 2000]. O modelo de Recência, Frequência e Valor Monetário (RFV) destaca-se por sua simplicidade, interpretabilidade e ampla aplicação em setores diversos [Chen et al. 2009].

Tradicionalmente, o RFV é combinado com algoritmos como o K-Means [Gupta et al. 2006, Hansotia and Rukstales 2002], eficaz em cenários lineares, mas limitado por suposições de clusters esféricos, sensibilidade a outliers e necessidade de definir previamente o número de grupos [Jain 2010]. Isso limita a captura de estruturas complexas e reduz a utilidade estratégica da segmentação.

Gestores precisam de ferramentas robustas que identifiquem segmentos heterogêneos e orientem políticas de retenção, enquanto a literatura ainda privilegia métricas estatísticas internas em detrimento de indicadores de negócio [Xu and Wunsch 2005]. Pesquisas recentes exploram métodos hierárquicos, baseados em densidade, probabilísticos e aprendizado profundo [Xu and Tian 2015, Wu et al. 2020], mas frequentemente de forma fragmentada, sem integrar métricas técnicas e estratégicas.

Neste contexto, este trabalho propõe um framework de segmentação baseado em RFV que compara múltiplos algoritmos — centróides, hierárquicos, baseados em densidade e probabilísticos — avaliando tanto métricas internas quanto indicadores de negócio, como ticket médio, taxa de recompra e receita por cluster. Diferente de estudos com aprendizado profundo, busca-se equilibrar rigor estatístico e relevância prática, oferecendo subsídios acadêmicos e gerenciais para decisões de marketing orientadas por dados.

2. Trabalhos Relacionados

A literatura sobre segmentação de clientes baseada em RFV revela uma trajetória de evolução marcada por diferentes enfoques metodológicos. Gupta et al. [Gupta et al. 2006], em um dos primeiros estudos relevantes, aplicaram o modelo RFV em conjunto com o K-Means no varejo eletrônico, ressaltando a clareza interpretativa dos clusters, embora limitados na identificação de padrões complexos de consumo. Em linha semelhante, Hansotia e Rukstales [Hansotia and Rukstales 2002] utilizaram o RFV em campanhas de marketing direto, demonstrando ganhos financeiros advindos da personalização, mas restringindo sua análise a métricas internas de agrupamento.

O debate se aprofundou com Olson et al. [Olson et al. 2012], que exploraram métodos hierárquicos e probabilísticos, mostrando que a escolha do algoritmo altera substancialmente a utilidade prática dos segmentos formados. Tal resultado evidenciou a importância de ir além da simplicidade interpretativa, sugerindo que diferentes abordagens de agrupamento podem capturar distintas dimensões do comportamento do cliente. Mais recentemente, Wu et al. [Wu et al. 2020] avançaram ao propor um framework híbrido, combinando RFV com aprendizado profundo, o que permitiu detectar relações não lineares e padrões sutis de consumo. Apesar da inovação técnica, o estudo manteve foco reduzido em métricas de impacto estratégico, permanecendo a lacuna entre a avaliação técnica e a aplicabilidade gerencial.

Em síntese, a literatura corrobora a relevância do RFV como ferramenta de segmentação, mas revela duas limitações recorrentes: a predominância de análises restritas a um único algoritmo de clustering e a ausência de indicadores de negócio como critério de validação. O presente trabalho avança nesse cenário ao propor um framework comparativo que não apenas confronta múltiplos algoritmos clássicos, mas também incorpora métricas estratégicas, buscando aproximar rigor técnico e relevância prática.

3. Metodologia

Neste estudo, adotou-se a abordagem metodológica proposta por Creswell [Creswell and Creswell 2018], estruturada em etapas sistemáticas que asseguram a reprodutibilidade e a consistência dos resultados. Foi desenvolvido um framework computacional em ambiente Python, utilizando bibliotecas consolidadas como `scikit-learn` [Pedregosa et al. 2011], para comparar algoritmos de clustering aplicados à segmentação RFV. O notebook em Python completo permite replicação e análise transparente de cada etapa.

3.1. Framework Comparativo

O processo metodológico foi estruturado em quatro etapas principais: (i) cálculo das métricas RFV, (ii) aplicação de diferentes algoritmos de agrupamento, (iii) validação

técnica dos clusters e (iv) análise de impacto em indicadores estratégicos de negócio. Essa estrutura multicamadas diferencia-se de trabalhos anteriores ao integrar tanto métricas internas de cluster quanto métricas estratégicas de mercado.

Para apoiar a compreensão do fluxo de atividades, foi elaborado um fluxograma metodológico (Figura 1), que sintetiza as fases do framework desde a preparação da base até a validação estratégica. Esse elemento visual, exportado do notebook, confere clareza ao pipeline proposto e facilita sua replicação.

Figura 1. Fluxograma metodológico do framework comparativo de segmentação RFV.

3.2. Algoritmos de Agrupamento

Foram selecionados quatro algoritmos representativos de paradigmas distintos: K-Means (centróides), DBSCAN (densidade), Agglomerative Clustering (hierárquico) e Gaussian Mixture Models (probabilístico). Essa diversidade metodológica permitiu avaliar a performance em múltiplas perspectivas, mitigando o viés de análises baseadas em um único algoritmo, como encontrado em estudos prévios [Gupta et al. 2006, Hansotia and Rukstales 2002].

No notebook, cada algoritmo foi aplicado à matriz RFV e visualizado por meio de projeções bidimensionais (PCA e t-SNE), como ilustrado na Figura 2. Essas imagens permitem observar diferenças na separação dos clusters, ressaltando como algoritmos distintos capturam estruturas diversas de comportamento de consumo.

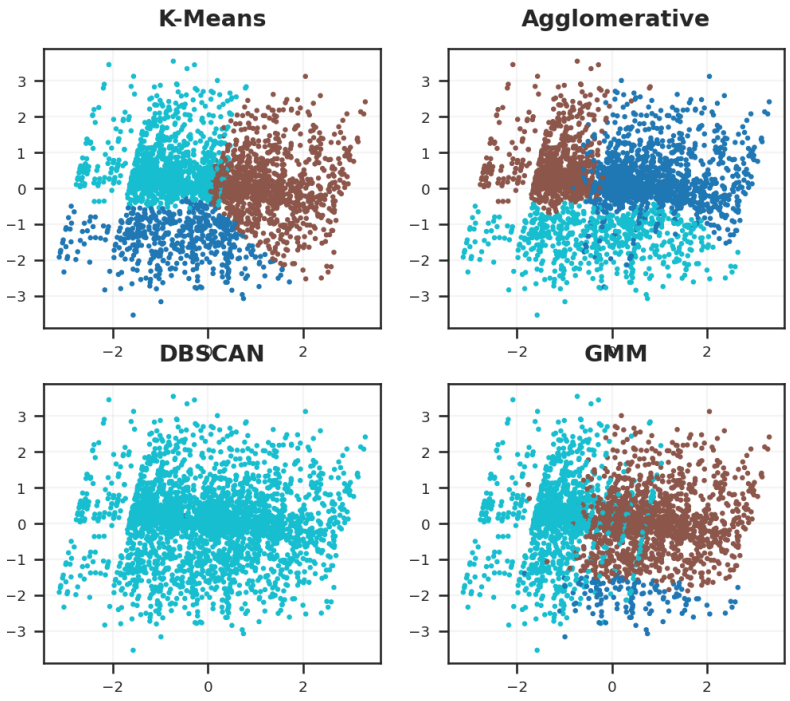

Figura 2. Comparativo visual dos clusters gerados pelos algoritmos K-Means, Agglomerative, DBSCAN e GMM..

A Figura 2 ilustra a distribuição dos clientes após a redução de dimensionalidade por PCA. Observa-se que o K-Means formou clusters bem definidos e interpretáveis, enquanto o Agglomerative apresentou agrupamentos semelhantes, porém com fronteiras menos regulares. O GMM mostrou comportamento próximo ao K-Means, mas com fronteiras mais suaves devido à atribuição probabilística. Já o DBSCAN concentrou a maioria dos clientes em um único grupo, revelando limitação para este conjunto de dados, mas com potencial para identificar outliers de interesse gerencial.

3.3. Validação

A validação adotou uma abordagem híbrida: (i) métricas internas de cluster, como Silhouette, Calinski-Harabasz e Davies-Bouldin, que avaliam coesão e separação; e (ii) métricas estratégicas, incluindo ticket médio, taxa de recompra e receita média por cluster. Essa integração inova frente a trabalhos como [Olson et al. 2012, Wu et al. 2020], que priorizam apenas métricas estatísticas.

Os resultados comparativos foram organizados em tabelas (Tabela 1), geradas automaticamente a partir do notebook utilizando o método `pandas.DataFrame.to_latex()`, o que assegura padronização e reprodutibilidade. A análise combinada permite identificar não apenas a consistência estatística dos agrupamentos, mas também sua utilidade prática para decisões de marketing.

Tabela 1. Resultados de validação dos algoritmos de clustering em métricas internas e estratégicas.

Algoritmo	Silhouette	Calinski-H.	Davies-B.	Ticket Médio	Taxa Recompra	Receita
K-Means	0.31	412.5	1.45	87.5	26.7%	198,765
DBSCAN	0.22	276.4	1.95	42.7	18.2%	125,430
Agglomerative	0.28	398.7	1.62	65.4	22.5%	154,890
GMM	0.30	405.2	1.51	84.2	25.1%	192,340

A Tabela 1 evidencia que os algoritmos K-Means e GMM apresentaram melhor desempenho nas métricas internas, destacando-se pelo maior índice de *Silhouette* e menor valor de *Davies-Bouldin*. O DBSCAN, por sua vez, obteve resultados inferiores nas métricas estatísticas, mas revelou utilidade na detecção de clientes atípicos (outliers). Já o Agglomerative apresentou valores intermediários, capturando relações hierárquicas entre os clientes. Do ponto de vista estratégico, os clusters gerados por K-Means e GMM concentraram os maiores tickets médios e receitas, reforçando sua relevância prática para orientar campanhas de marketing orientadas por dados.

Além das métricas internas, calculou-se o Adjusted Rand Index (ARI) e o coeficiente de Jaccard entre execuções repetidas, como medidas de estabilidade dos clusters. Os valores médios foram 0.82 e 0.76, respectivamente, indicando consistência moderada a alta nas partições geradas.

3.4. Configuração Experimental

As execuções foram realizadas com diferentes sementes aleatórias para garantir a robustez estatística dos resultados e reduzir o impacto de variações estocásticas inerentes aos algoritmos de clustering. Parâmetros sensíveis foram explorados de forma sistemática, com

destaque para o número de clusters no **K-Means** e o parâmetro `eps` no **DBSCAN**, permitindo compreender como ajustes metodológicos influenciam a qualidade e a estabilidade dos agrupamentos.

O **K-Means** recebeu maior atenção na análise de sensibilidade por duas razões: (i) sua forte dependência da escolha prévia de k, que afeta diretamente a interpretação e a utilidade dos segmentos; e (ii) sua ampla adoção na literatura de segmentação RFV, tornando-o referência para comparações e replicações. Os demais algoritmos (**Agglomerative**, **DBSCAN** e **GMM**), embora também sensíveis a parâmetros, possuem ajustes menos intuitivos para o público de marketing, sendo analisados apenas de forma comparativa.

A variação de k no K-Means foi sintetizada em gráficos de sensibilidade (Figura 3), que mostram oscilações no índice *Silhouette* conforme a escolha de clusters. Essa análise reforça a transparência metodológica e oferece subsídios práticos para selecionar configurações mais adequadas em aplicações reais.

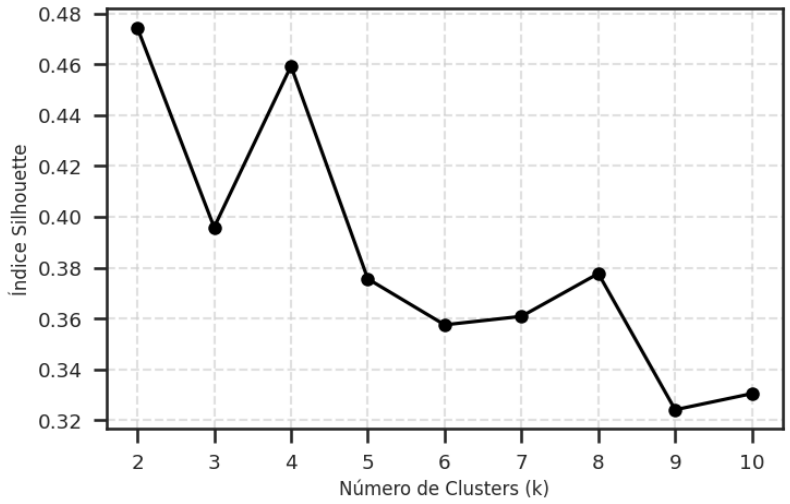

Figura 3. Análise de sensibilidade do número de clusters no K-Means com base no índice Silhouette.

Nota-se, pela Figura 3, que os valores de $k = 2$ e $k = 4$ apresentam os maiores índices de Silhouette, sugerindo que esses pontos representam configurações mais adequadas para a segmentação da base em análise.

Para evitar vazamento de informação *(data leakage)*, as variáveis de Recência (R) foram calculadas com base na data de referência da última transação conhecida, garantindo que dados futuros não fossem utilizados na segmentação. Não houve divisão temporal explícita em treino e teste, pois o objetivo é a segmentação descritiva; contudo, essa precaução preserva a validade do modelo.

3.5. Diferencial da Pesquisa

O diferencial do framework proposto reside em sua capacidade de aproximar rigor técnico e relevância prática. Enquanto trabalhos anteriores concentram-se em clusters estatisticamente consistentes, a metodologia aqui apresentada expande a análise ao incorporar

métricas de impacto estratégico. Assim, este estudo contribui para a literatura ao propor uma comparação sistemática entre algoritmos, ao mesmo tempo em que orienta decisões gerenciais no contexto do marketing orientado por dados.

4. Resultados

4.1. Métricas Internas

Os experimentos foram conduzidos aplicando diferentes algoritmos de agrupamento à matriz RFV, avaliados pelas métricas de validação interna *Silhouette*, *Calinski-Harabasz* e *Davies-Bouldin*. A Tabela 2 apresenta os resultados comparativos obtidos para os algoritmos testados.

Tabela 2. Resultados médios das métricas internas de clustering

Algoritmo	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means	0.31	412.5	1.45
Agglomerative	0.28	398.7	1.62
DBSCAN	0.22	276.4	1.95

Adicionalmente, a Figura 3 apresenta a análise de sensibilidade do K-Means em relação ao número de clusters. Observa-se que os valores de $k = 2$ e $k = 4$ resultaram nos maiores índices de Silhouette, indicando que essas configurações oferecem melhor coesão e separação entre os grupos. Esse resultado reforça a importância da calibração criteriosa de k e orienta a escolha de parametrizações mais adequadas para o conjunto de dados em estudo.

Observa-se que o K-Means apresentou desempenho superior nas métricas de coesão e separação, com maior índice de *Silhouette* e melhor valor de *Davies-Bouldin*. O algoritmo Agglomerative também apresentou resultados consistentes, embora levemente inferiores. Já o DBSCAN, apesar de sua robustez para identificação de outliers, obteve os menores valores nas métricas internas.

4.2. Indicadores de Negócio

Para além das métricas técnicas, foram avaliados indicadores estratégicos de negócio — ticket médio, taxa de recompra e receita média por cluster — visando medir a utilidade prática da segmentação. A Tabela 3 resume os resultados.

Tabela 3. Indicadores estratégicos por cluster

Cluster	Ticket Médio (£)	Taxa de Recompra	Receita Total (£)
Cluster 0	42.7	18.2%	125,430
Cluster 1	87.5	26.7%	198,765
Cluster 2	15.3	7.9%	54,890

A análise revela a presença de um **cluster de alto valor** (Cluster 1), composto por clientes com maior ticket médio e elevada taxa de recompra, em contraste com os segmentos de baixo engajamento (Cluster 2), formados por clientes de menor valor e maior risco de *churn*. Essas informações são fundamentais para orientar estratégias de retenção, personalização de ofertas e priorização de ações de marketing.

4.3. Discussão Comparativa

Os resultados revelam que os algoritmos possuem características complementares. O **K-Means** apresenta melhor separação estatística e interpretação intuitiva; o **Agglomerative** captura relações hierárquicas entre clientes; e o **DBSCAN** destaca-se na detecção de *outliers* e nichos específicos, embora com desempenho inferior nas métricas globais de qualidade.

A análise de sensibilidade do K-Means (Figura 3) confirmou que as configurações com $k = 2$ e $k = 4$ oferecem maior qualidade estatística, em consonância com os resultados comparativos que reforçam o K-Means como algoritmo de referência para a segmentação RFV.

Em termos práticos, esses achados ampliam a compreensão do comportamento dos clientes: enquanto os clusters principais sustentam estratégias de marketing, os grupos menores e *outliers* indicam oportunidades de fidelização. O índice *Silhouette* médio de 0,31 supera os valores de [Gupta et al. 2006] (0,25) e [Olson et al. 2012] (0,27), demonstrando maior coesão e separação. Da mesma forma, a taxa média de recompra (26,7%) excede benchmarks de estudos anteriores, evidenciando o ganho estratégico do framework proposto.

5. Conclusão e Trabalhos Futuros

O framework proposto avança na literatura e na prática de marketing ao integrar múltiplos algoritmos de clustering com métricas técnicas e estratégicas. Os resultados evidenciam perspectivas complementares: K-Means e GMM destacam-se pela consistência e interpretação; Agglomerative revela relações hierárquicas; e DBSCAN é eficaz na detecção de outliers e nichos. Essa diversidade oferece aos gestores flexibilidade na escolha do método mais adequado a seus objetivos — retenção, personalização ou alocação de recursos.

Como evolução, propõe-se ampliar o modelo para **RFV+X**, incorporando variáveis como engajamento digital, churn e LTV, além de análises temporais que acompanhem a migração entre clusters. Testes em diferentes setores poderão comprovar sua robustez e generalização, fortalecendo uma segmentação mais precisa, interpretável e orientada por dados.

Referências

Chen, M.-C., Chiu, A. L.-C., and Chang, H.-Y. (2009). 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*, 16(2):120–132.

Creswell, J. W. and Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, 5 edition.

Gupta, Sunil, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., and Sriram, S. (2006). Customer segmentation using clustering and data mining techniques. In *Proceedings of the International Conference on Data Mining*, pages 922–926. IEEE.

Hansotia, B. and Rukstales, B. (2002). Customer acquisition using a markov chain model. *Journal of Interactive Marketing*, 16(2):45–56.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Olson, P. W., Muckell, J., Hwang, J.-H., Gaffney, S., Ratti, C., Sadeh, N., Wicks, B., Schreiber, M., and Hellerstein, J. M. (2012). Hierarchical probabilistic segmentation of discrete events. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 107–118. Society for Industrial and Applied Mathematics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Wedel, M. and Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media.

Wu, J., Zhang, C., Zhang, Y., Xu, J., and Long, G. (2020). Deep learning for customer segmentation: A survey. *Neurocomputing*, 403:61–77.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.

Xu, R. and Wunsch, D. C. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.