Catalogação na publicação
Seção de Catalogação e Classificação

J95s Junior, Geraldo Figueiredo de Santana.
Segmentação de clientes e otimização de estratégias: abordagens avançadas com análise RFM e algoritmos de clusterização / Geraldo Figueiredo de Santana Junior. - João Pessoa, 2024.
15 f. : il.

Orientação: Yuri de Almeida Malheiros Barbosa.
TCC (Graduação) - UFPB/Informática.

1. Clusterização. 2. RFM. 3. Segmentação. I. Barbosa, Yuri de Almeida Malheiros. II. Título.

UFPB/CI CDU 004.421

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

Segmentação de Clientes e Otimização de Estratégias: Abordagens Avançadas com Análise RFM e Algoritmos de Clusterização

Geraldo Figueiredo de Santana Junior[1], **Yuri de Almeida Malheiros Barbosa**[1]

[1]Centro de Informática
Universidade Federal da Paraíba (UFPB) - João Pessoal, PB - Brasil

geraldo.figueiredosj@gmail.com

Abstract. *This article discusses customer segmentation and strategy optimization through RFM analysis and clustering algorithms. RFM analysis (Recency, Frequency, and Monetary Value) identifies high-value customers, helping prioritize marketing efforts for retention and loyalty. The study's motivation is the need for companies to better understand their customers, improve their marketing strategies, and boost business growth. Combining RFM analysis with data mining techniques provides valuable insights for customer relationship management and business operations optimization. The study applies RFM analysis using the k-Means method to evaluate its effectiveness, enabling the creation of distinct customer allocation profiles, providing important information for investment management and optimization, demonstrating to be an effective approach for customer segmentation in financial contexts and for strategic decision-making in the market.*

Resumo. *Este artigo discute a segmentação de clientes e a otimização de estratégias por meio da análise RFM e algoritmos de clusterização. A análise RFM (Recência, Frequência e Valor Monetário) identifica clientes de alto valor, ajudando a priorizar esforços de marketing para retenção e fidelização. A motivação do estudo é a necessidade das empresas de entender melhor seus clientes, melhorar suas estratégias de marketing e impulsionar o crescimento dos negócios. A combinação da análise RFM com técnicas de mineração de dados oferece insights valiosos para a gestão de relacionamento com o cliente e otimização das operações comerciais. O estudo aplica a análise RFM usando o método k-Means para avaliar sua eficácia, permitindo a criação de perfis distintos de alocação de clientes fornecendo informações importantes para a gestão e otimização de investimentos, demonstrando ser uma abordagem eficaz para segmentação de clientes em contextos financeiros e para a tomada de decisões estratégicas no mercado.*

1. Introdução

A segmentação de clientes é um aspecto fundamental das estratégias de marketing modernas, permitindo que as empresas adaptem suas abordagens a diferentes grupos de clientes com base em seus comportamentos e características. Um método amplamente utilizado para a segmentação de clientes é a análise RFM (*Recency, Frequency and Monetary Value*), focando em Recência, Frequência e Valor Monetário. Esta abordagem tem se

mostrado eficaz em diversas indústrias, incluindo varejo, *e-commerce* e serviços financeiros, destacando sua versatilidade e aplicabilidade em diferentes setores Ernawati e Sarno, 2021.

A análise RFM é uma técnica de segmentação de clientes que se concentra em três dimensões principais: Recência (R), Frequência (F) e Valor Monetário (M) Blattberg et al., 2008. Recência refere-se ao tempo desde a última transação de um cliente, onde clientes mais recentes são considerados mais propensos a responder a campanhas de marketing. Frequência mede o número de transações em um determinado período, identificando clientes que compram com maior regularidade. Valor Monetário avalia o total gasto pelos clientes, permitindo a identificação daqueles que geram mais receita. Por exemplo, em um contexto de mercado financeiro, um cliente que comprou ações há uma semana (alta Recência), que realiza transações mensalmente (alta Frequência) e que investe uma quantia significativa em cada compra (alto Valor Monetário) seria considerado de alto valor para a corretora de valores. A análise RFM facilita a segmentação dos clientes, permitindo que as empresas direcionem suas estratégias de marketing e recursos de maneira mais eficiente, aumentando a retenção e a fidelidade dos clientes e pode ser utilizada em diversos setores, como varejo, telecomunicações e saúde. Além disso, ao integrar a análise RFM com técnicas de mineração de dados, como algoritmos de clusterização, é possível obter *insights* mais profundos e acionáveis sobre os padrões de comportamento dos clientes.

A utilização da análise RFM na segmentação de clientes oferece diversas vantagens e motivos para as empresas adotarem essa abordagem. Primeiramente, a análise RFM permite uma segmentação mais precisa dos clientes, com base em seu comportamento de compra, o que possibilita a personalização de estratégias de marketing e a oferta de produtos e serviços mais adequados a cada segmento Aggelis e Christodoulakis, 2005. Além disso, a identificação de clientes de alto valor, com base nos critérios RFM, permite que as empresas priorizem esforços de marketing e foquem em reter esses clientes lucrativos, contribuindo para o aumento da fidelidade do cliente e a maximização do retorno sobre o investimento Christy et al., 2021. A análise RFM também auxilia na previsão de padrões de comportamento de compra e na tomada de decisões estratégicas, proporcionando insights valiosos para a gestão de relacionamento com o cliente e a otimização das operações comerciais Cheng e Chen, 2009.

Estudos recentes Aggelis e Christodoulakis, 2005; Derya, 2011; Gustriansyah et al., 2020 têm investigado métodos para otimizar a análise RFM, como o uso de algoritmos de clusterização como o *k-Means* para aprimorar a segmentação de produtos e a gestão de estoque Christy et al., 2021. O principal objetivo desta pesquisa é utilizar a análise RFM em conjunto com o método *k-Means* e avaliar sua eficácia utilizando uma fonte de dados real. Ao incorporar a análise RFM em técnicas de mineração de dados, as empresas podem obter *insights* mais profundos sobre o comportamento do cliente, aprimorar estratégias de marketing personalizadas e melhorar a retenção de clientes Christy et al., 2021.

A integração da análise RFM com técnicas de mineração de dados oferece uma ferramenta poderosa para as empresas aprimorarem a segmentação de clientes, prever padrões de compra e melhorar a satisfação e fidelidade do cliente Chang et al., 2011. A pesquisa em andamento e os avanços na análise RFM destacam sua importância nas

práticas de marketing modernas e seu potencial para impulsionar o crescimento e o sucesso das empresas.

Na literatura enfatiza-se a importância da análise de lucratividade do cliente por meio da pontuação RFM, especialmente em setores como *e-banking*, onde a retenção de clientes e a geração de receita são críticas Aggelis e Christodoulakis, 2005. Ao identificar clientes de alto valor com base nos critérios RFM, as empresas podem priorizar seus esforços de marketing e adaptar seus serviços para atender às necessidades específicas desses segmentos valiosos.

Após a identificação dos clientes de alto valor por meio da análise RFM, as empresas podem adotar estratégias adicionais para maximizar o retorno sobre esses investimentos Chang et al., 2011. Uma abordagem comum é a aplicação de técnicas de otimização de carteira, inspiradas nos princípios estabelecidos por Markowitz, 1952. Essas técnicas visam a alocação eficiente de ativos dentro de cada segmento identificado pela análise RFM. Ao utilizar o otimizador de média-variância, por exemplo, as empresas podem determinar a combinação ideal de investimentos para cada carteira, equilibrando os riscos e os retornos esperados. Essa abordagem permite que as empresas otimizem não apenas a lucratividade de cada cliente, mas também a gestão global de seus portfólios, garantindo uma alocação de recursos eficaz e alinhada aos objetivos estratégicos da organização.

Utilizamos o otimizador de média-variância estabelecido por Markowitz, 1952, que consiste em: se $\mathbf{w}$ é o vetor de pesos dos ativos com retornos esperados $\boldsymbol{\mu}$, então o retorno do portfólio é igual à soma ponderada dos retornos de cada ativo, ou seja, $\mathbf{w}^T\boldsymbol{\mu}$. A volatilidade do portfólio, por sua vez, é determinada pela matriz de covariância $\boldsymbol{\Sigma}$ e é expressa por $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$. Esse processo de otimização de portfólio é formulado como um problema de otimização convexa, podendo ser resolvido utilizando técnicas de programação quadrática.

Ao estabelecermos um retorno alvo como $\boldsymbol{\mu}^*$, o problema de otimização do portfólio é então formulado, buscando minimizar a volatilidade do portfólio sujeito a essa restrição de retorno alvo. Variando o retorno alvo, obtemos diferentes conjuntos de pesos para os ativos, resultando em diferentes carteiras. O conjunto de todas essas carteiras ótimas é conhecido como a fronteira eficiente (*Efficient Frontier*). Essa abordagem nos permite avaliar e comparar objetivamente as diferentes estratégias de investimento, contribuindo para uma gestão mais eficaz dos portfólios financeiros.

A equação da média-variância para a otimização do portfólio é dada pela seguinte expressão:

$$\min_{\mathbf{w}} \quad \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$$

sujeita às restrições:

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\mu} &\geq \mu^* \\ \mathbf{w}^T\mathbf{1} &= 1 \\ w_i &\geq 0 \quad \forall i \end{aligned}$$

Onde:

- $\mathbf{w}$ é o vetor de pesos dos ativos no portfólio,
- $\mathbf{\Sigma}$ é a matriz de covariância dos retornos dos ativos,
- $\boldsymbol{\mu}$ é o vetor de retornos esperados dos ativos,
- μ^* é o retorno alvo do portfólio,
- $\mathbf{1}$ é um vetor de uns,
- w_i é o peso do ativo i.

Essa formulação busca minimizar a volatilidade do portfólio, representada pela expressão $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$, sujeita a uma restrição de retorno alvo e às restrições de que a soma dos pesos é igual a 1 e que os pesos de cada ativo são não negativos.

Os principais indicadores obtidos por meio da saída do otimizador de média-variância incluem o retorno anual, volatilidade anual e o índice de Sharpe. O retorno anual representa o percentual de crescimento médio do investimento ao longo de um ano. Já a volatilidade anual indica a medida da flutuação dos retornos de um investimento ao longo do tempo, proporcionando insights sobre o risco associado ao portfólio. O índice de Sharpe, por sua vez, é uma métrica de desempenho ajustada ao risco, calculada pela divisão do retorno excedente pelo desvio padrão dos retornos, fornecendo uma avaliação da eficiência do portfólio em relação ao risco assumido Sharpe, 1998.

2. Trabalhos Relacionados

A segmentação de clientes com base no modelo RFM tem sido objeto de estudo em diversas áreas, incluindo varejo, e-commerce e serviços financeiros. Chang et al., 2011 destacam a relevância da análise RFM na compreensão do comportamento do cliente e na personalização das estratégias de marketing, abrangendo diferentes setores comerciais.

Por outro lado, estudos como Gustriansyah et al., 2020 e Christy et al., 2021 exploram métodos de clusterização, como o *K-Means* e o Fuzzy *C-Means*, para aprimorar a segmentação de clientes e melhorar os processos de gestão de estoque. Enquanto o primeiro enfatiza a integração da análise RFM com técnicas de mineração de dados, o segundo destaca a importância da retenção de clientes e compara diferentes algoritmos de clusterização.

Ernawati e Sarno, 2021 e Cheng e Chen, 2009 examinam a aplicação da análise RFM em diferentes contextos, como unidades de *e-banking* e segmentação de clientes, respectivamente. Ambos enfatizam a importância da identificação de clientes de alto valor e o uso de modelos RFM modificados para personalizar as estratégias de marketing.

Além disso, estudos como Markowitz, 1952 discutem a construção de carteiras de investimento eficientes, destacando a diversificação e otimização do portfólio como elementos-chave. Embora este estudo se concentre em investimentos financeiros, suas conclusões sobre a maximização do retorno esperado têm implicações relevantes na gestão de clientes e estratégias de marketing.

Por fim, Aggelis e Christodoulakis, 2005 explora a segmentação de clientes com base na análise RFM e sua extensão para outros algoritmos de clusterização, como *K-Means* e Fuzzy *C-Means*. O estudo ressalta a importância da personalização das estratégias de marketing, com base no comportamento de compra dos clientes e na análise do desempenho dos clientes em cada segmento.

Na Tabela 1 é possível visualizar a sumarização dos trabalhos relacionados ao proposto aqui e visualizar as principais diferenças em relação aos métodos de análise utilizados, aos ojetivos da pesqisa e os resultados e conclusões obtidos pelos autores.

Artigo	Métodos de Análise	Objetivos da Pesquisa	Resultados e Conclusões
Ernawati e Sarno, 2021	RFM, *K-Means*, LEM2	Demonstração do procedimento com base de dados da C-company	Superioridade e robustez do método proposto
Gustriansyah et al., 2020	RFM, *K-Means*	Otimização da clusterização em análise RFM	Identificação de grupos de clientes com base em seus comportamentos de compra
Chang et al., 2011	Análise Estatística	Investigação de padrões de comportamento do consumidor	Identificação de tendências significativas
Christy et al., 2021	Análise de RFM	Classificação de clientes com base em RFM	Geração de rankings de clientes
Markowitz, 1952	Teoria de Portfólio	Desenvolvimento de modelo de otimização de portfólio	Introdução do conceito de diversificação
Cheng e Chen, 2009	Análise de Dados	Estudo sobre eficiência operacional	Recomendações para melhoria de processos
Derya, 2011	*K-Means*	Segmentação de clientes	Identificação de perfis de clientes distintos
Ernawati e Sarno, 2021	RFM, textitK-Means	Clusterização de clientes com base em RFM	Geração de regras de classificação eficazes

Tabela 1. Tabela comparativa dos trabalhos relacionados

Este estudo se insere no contexto da análise de dados e segmentação de clientes, seguindo a abordagem de combinação de RFM e *K-Means*. Ao comparar com pesquisas anteriores, como os trabalhos de Gustriansyah et al., 2020 e Ernawati e Sarno, 2021, que demonstram a clusterização de clientes e a robustez do método proposto, e o estudo de Aggelis e Christodoulakis, 2005, que foca na otimização da clusterização em análise RFM com o algoritmo *K-Means*, observamos uma variedade de métodos e objetivos. Enquanto estes estudos exploram diferentes aspectos da análise de dados e segmentação de clientes, o presente trabalho se destaca por sua aplicação prática na clusterização da base de clientes de uma empresa parceira, calculando a performance da carteira de cada cliente e gerando carteiras de referência para cada grupo, com base na proporção de cada produto. Essa abordagem contribui para a personalização das estratégias de marketing e vendas, auxiliando na tomada de decisões estratégicas e no aprimoramento das práticas comerciais da empresa.

3. Metodologia

Esta seção detalha o processo seguido, desde a obtenção e preparação dos dados, até a avaliação dos portfólios financeiros para cada grupo identificado. Inicialmente, descrevemos como os dados foram adquiridos e preparados, destacando a colaboração com uma empresa parceira na coleta e armazenamento dos dados. Em seguida, abordamos as etapas de limpeza de dados, onde identificamos *outliers* e assimetrias e aplicamos técnicas de remoção desses dados e normalização para garantir a qualidade dos dados utilizados na análise.

Posteriormente, exploramos a clusterização dos dados usando o algoritmo *K-Means*, com destaque para a utilização da aceleração por GPU para tornar o processo

mais eficiente. A seleção do número ideal de grupos é baseada no método do cotovelo, que avalia a inércia para diferentes números de grupos. Os resultados obtidos são então visualizados por meio do algoritmo t-SNE (*t-distributed Stochastic Neighbor Embedding*), proporcionando uma interpretação mais acessível dos grupos identificados.

Por fim, apresentamos a avaliação dos portfólios financeiros para cada grupo, utilizando o otimizador de média-variância da biblioteca PyPortfolioOpt. Discutimos os indicadores financeiros utilizados na avaliação das carteiras destacando sua relevância na seleção das melhores carteiras para cada segmento de clientes.

3.1. Obtenção e Preparação dos Dados

A base de dados utilizada neste estudo foi cedida por uma corretora de investimentos renomada no mercado financeiro, com uma significativa presença na bolsa de valores americana. Com mais de 800 mil clientes registrados e uma custódia de ativos que ultrapassa os 3 bilhões de dólares, a empresa é reconhecida por sua posição de destaque e influência no setor. Os dados fornecidos, que abrangem informações detalhadas sobre transações financeiras e posições de custódia dos clientes, bem como dados de mercado relacionados aos preços de ativos negociados, foram anonimizados para proteger a identidade dos clientes, conforme acordos estabelecidos pela Lei Geral de Proteção de Dados (LGPD) e outras regulamentações de privacidade.

Os dados fornecidos pela empresa incluem uma variedade de informações sobre os clientes e dados históricos do mercado. No que diz respeito às informações dos clientes, a base de dados contém detalhes como o valor total investido na carteira de cada cliente, bem como o valor investido por produto de investimento específico. Cada produto de investimento é identificado por meio de um código único, que pode ser o *CUSIP* (*Committee on Uniform Securities Identification Procedures*) ou o símbolo do ativo, dependendo do tipo de ativo. O CUSIP é um identificador alfanumérico de nove dígitos atribuído a cada título negociado publicamente nos Estados Unidos e no Canadá, enquanto o *símbolo* é o símbolo do *ticker* utilizado para identificar ações em bolsas de valores. Além disso, a base de dados registra todas as transações realizadas pelos clientes, incluindo a identificação do produto, a data da transação e o valor da aplicação. Em relação aos dados históricos do mercado, são fornecidos identificadores únicos para cada produto, juntamente com o preço de fechamento do dia correspondente. Esses dados detalhados proporcionam uma visão abrangente das atividades de investimento dos clientes e das tendências de mercado ao longo do tempo, essenciais para a análise e modelagem realizadas neste estudo.

Para adquirir os dados necessários, utilizamos rotinas de extração, definidas por um time específico da empresa parceira. Essas rotinas, compostas por *scripts* e *pipelines* de dados, são executadas em intervalos regulares, garantindo a obtenção de informações atualizadas e precisas das diversas fontes de dados disponíveis.

A variedade de fontes de dados inclui bases de dados MySQL, PostgreSQL e Datastore. Através dessas rotinas de extração, conseguimos recuperar os dados de forma sistemática e eficiente, mantendo a consistência e a integridade das informações armazenadas.

Além disso, os dados históricos dos produtos oferecidos pela empresa parceira foram obtidos internamente, diretamente de sua plataforma. Essa abordagem assegurou a

confiabilidade e a precisão dos dados, uma vez que não dependemos de terceiros para sua recuperação.

Uma vez extraídos, os dados foram armazenados no BigQuery. Essa plataforma centralizada oferece recursos escaláveis e avançados para armazenamento e análise de dados, garantindo que possamos manter um conjunto completo e atualizado de dados para análise.

3.2. Criação do Dataframe RFM

Após a obtenção dos dados, o próximo passo foi criar o dataframe RFM com base em três métricas principais: Recência, Frequência e Valor Monetário. A coluna de Recência foi construída calculando a quantidade de dias desde a última compra realizada por cada cliente, o que permite entender a proximidade temporal das interações. Em seguida, a coluna de Frequência foi determinada pelo número total de transações realizadas por cada cliente ao longo do período analisado, refletindo seu engajamento e frequência de interação. Finalmente, a coluna de Valor Monetário foi formada pelo montante total investido ou gasto pelos clientes em suas transações ou investimentos, proporcionando uma medida quantitativa do seu envolvimento financeiro.

3.3. Limpeza de Dados

Após a coleta dos dados, uma análise visual foi realizada, como mostrado na Figura 1, revelando assimetrias no *dataframe* com os dados. Para lidar com essas questões, foi adotada a técnica de transformação logarítmica em todo o conjunto de dados. Essa técnica consiste na aplicação do logaritmo natural a cada valor do conjunto de dados. O logaritmo natural, uma função matemática que mapeia números positivos para números reais, possui a capacidade de comprimir valores grandes e expandir valores pequenos. Essa característica auxilia na redução da assimetria nos dados, tornando a distribuição mais simétrica e facilitando a interpretação e análise dos dados, especialmente quando estes apresentam distribuição assimétrica ou uma ampla faixa de valores Gustriansyah et al., 2020.

Nas Figuras 1 e 3, podemos observar uma assimetria, onde a maioria dos dados está concentrada à esquerda, resultando em uma cauda longa na distribuição. Esse tipo de distribuição assimétrica pode prejudicar a interpretação dos resultados e comprometer a eficácia de modelos estatísticos, especialmente aqueles que assumem uma distribuição normal dos dados.

Após a aplicação da transformação logarítmica, é evidente uma melhoria na distribuição dos dados nas três colunas da base de dados, como ilustrado nas Figuras 2 e 4.

3.4. Método de Agrupamento

O agrupamento é uma etapa essencial na análise de dados, permitindo a identificação de grupos similares dentro de um conjunto de dados. Utilizamos o algoritmo *K-Means*, uma técnica amplamente empregada em análises de dados, para realizar a clusterização dos dados normalizados do *dataframe* RFM.

Utilizamos a aceleração por GPU para aplicar o algoritmo *K-Means*, o que nos permitiu realizar a clusterização de forma significativamente mais rápida e eficiente. Esta

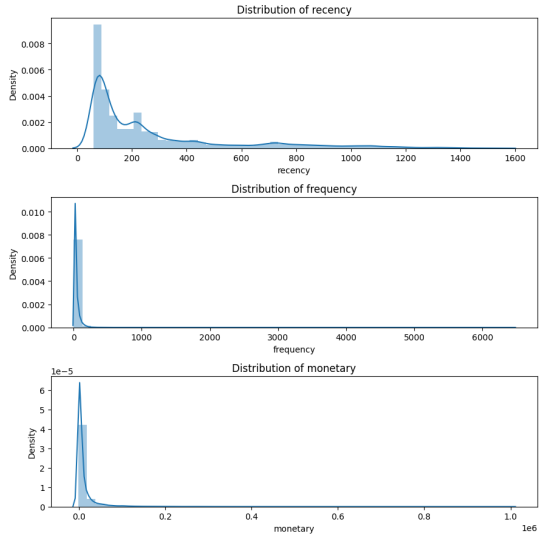

Figura 1. Base de dados antes de normalização

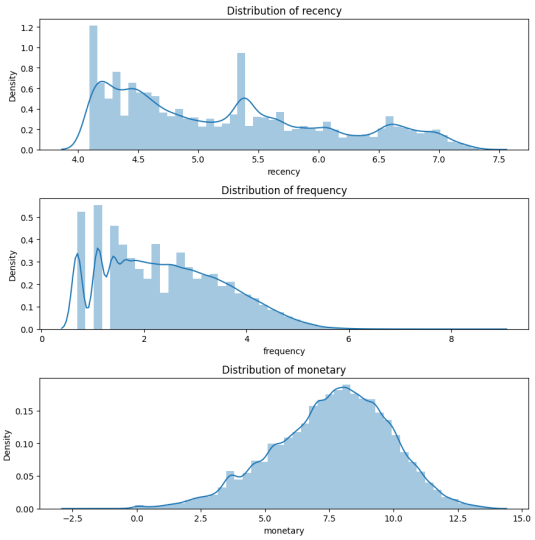

Figura 2. Base de dados normalizada

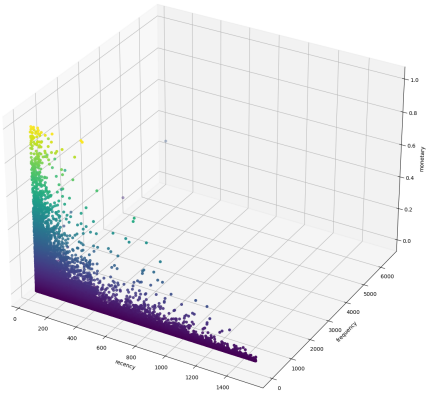

Figura 3. *Dataframe* RFM antes de normalização

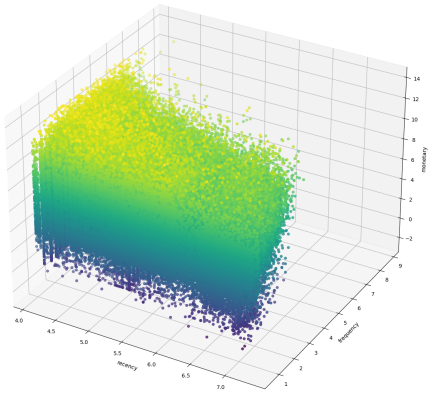

Figura 4. *Dataframe* RFM normalizado

técnica aproveita o poder de processamento paralelo das unidades de processamento gráfico (GPU), permitindo que múltiplos cálculos sejam executados simultaneamente. Isso resulta em tempos de execução mais curtos e maior capacidade de lidar com conjuntos de dados grandes.

Com o *dataframe* normalizado e *outliers* removidos, aplicamos o algoritmo *K-Means* utilizando a biblioteca *Scikit-learn* (*sklearn*). Ao definir diferentes números de grupos, podemos explorar diferentes estruturas nos dados e identificar agrupamentos que melhor representam as características dos dados normalizados e livres de *outliers*.

Para determinar o número ideal de *clusters* em uma análise de clusterização, é essencial empregar uma técnica que avalie os resultados para diferentes valores de *K*. Uma abordagem comum é o método do cotovelo (*elbow method*), amplamente utilizado em análises de agrupamento (*clustering*) para identificar o número ótimo de *clusters* em um conjunto de dados. Este método implica analisar graficamente o número de *clusters*

em relação a uma métrica de avaliação, como a Soma dos Quadrados dos Erros (SSE) ou distância euclidiana, e identificar o ponto de inflexão na curva que se assemelha a um cotovelo Syakur et al., 2018.

No escopo deste estudo, a implementação do algoritmo *K-Means* foi conduzida utilizando a biblioteca *Scikit-learn*(*sklearn.cluster*). Com o intuito de explorar diversas configurações de agrupamento, foram aplicadas iterações do algoritmo *K-Means* com um conjunto variável de *clusters*, variando de 1 a 10. A fim de avaliar a eficácia dos agrupamentos, utilizou-se a distorção como métrica de avaliação, calculando a soma das distâncias euclidianas de cada ponto em relação ao centróide do cluster correspondente. Para este cálculo, foi empregada a função *cdist* da biblioteca *Scipy* (*scipy.spatial.distance*).

Observamos que um número entre dois e quatro grupos seria o ideal para o nosso conjunto de dados, conforme ilustrado na Figura 5.

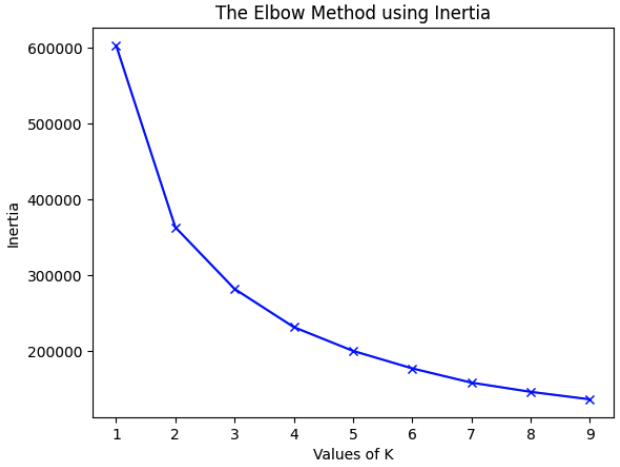

Figura 5. Avaliação de clusters utilizando o método do cotovelo

Esses resultados nos fornecem *insights* valiosos para a definição do número ótimo de grupos a serem utilizados na segmentação de clientes RFM, permitindo uma análise mais precisa e uma melhor compreensão dos padrões de comportamento dos clientes.

Para aprimorar a compreensão dos resultados obtidos com a clusterização, é essencial contar com ferramentas que proporcionem uma visualização clara e intuitiva dos grupos identificados. Nesse contexto, exploramos uma técnica conhecida como t-SNE, que se destaca por sua habilidade em converter conjuntos de dados de alta dimensionalidade em uma matriz de similaridades e posteriormente visualizar esses dados de maneira altamente eficiente Van der Maaten e Hinton, 2008. Ao contrário de outras abordagens, que frequentemente enfrentam desafios na captura simultânea da estrutura local e global dos dados, o t-SNE demonstra uma notável capacidade em preservar a estrutura local com precisão, como a identificação de *clusters* em diversas escalas Van der Maaten e Hinton, 2008.

Optamos por utilizar a implementação do t-SNE disponível na biblioteca *cuML*, uma biblioteca de aprendizado de máquina acelerada por *GPU* para Python, projetada para funcionar eficientemente em hardware NVIDIA CUDA. Ao utilizar o algoritmo

t-SNE da cuML, pudemos utilizar a capacidade de processamento paralelo das GPUs, resultando em tempos de execução significativamente reduzidos em comparação com implementações tradicionais em CPU. O único parâmetro configurado no t-SNE foi *n_components* que corresponde a dimensionalidade da saída do algoritmo, definimos como 2 para visualizar na Figura 6.

3.5. Avaliação do Portfólio

Após a segmentação da base de clientes em grupos, avançamos para a etapa de avaliação dos portfólios financeiros. Nessa fase, nosso objetivo foi identificar e selecionar a melhor carteira para cada grupo resultante da segmentação. Para realizar essa avaliação, utilizamos a biblioteca *PyPortfolioOpt*, uma ferramenta em Python voltada para análise e otimização de portfólios com base em algoritmos avançados e dados históricos dos ativos.

4. Resultados e Discussões

Nesta seção, são apresentados os resultados obtidos através do método de agrupamento aplicado aos dados RFM. A visualização dos agrupamentos formados foi realizada utilizando a técnica t-SNE (t-Distributed Stochastic Neighbor Embedding), que proporciona uma representação visual dos dados em um espaço bidimensional Van der Maaten e Hinton, 2008.

4.1. Resultados do Método de Agrupamento

A visualização dos agrupamentos formados foi realizada com a técnica t-SNE e pode ser observada na Figura 6, que são utilizadas para entender como interpretar as nuances e padrões que emergem da distribuição dos pontos.

Cada ponto no gráfico representa uma observação individual dos dados do *dataframe* RFM, correspondendo a um cliente específico. A disposição dos pontos no espaço bidimensional é determinada pelo algoritmo t-SNE, que busca preservar as relações de proximidade entre os pontos originais de alta dimensão em um espaço de menor dimensão. Assim, pontos que estão próximos no gráfico t-SNE são aqueles que compartilham características semelhantes nos dados originais.

Ao interpretar essa visualização, é crucial considerar a separabilidade dos grupos. Em um cenário ideal, os pontos pertencentes a diferentes grupos devem estar bem separados, indicando que os grupos são distintos e facilmente distinguíveis. Além disso, a estrutura global do gráfico também é relevante. Observar se existem regiões densas de pontos ou padrões de dispersão pode fornecer *insights* sobre a distribuição dos dados e a presença de agrupamentos naturais. Por exemplo, grupos de pontos agrupados em uma área específica podem indicar a presença de subgrupos ou segmentos distintos de clientes.

Ao compararmos as duas visualizações disponíveis, é evidente que a presença de mais de 3 grupos resulta em regiões menos distintas, o que sugere uma maior incerteza na categorização dos clientes. Ao examinar a Figura 6, parece que a opção de 3 grupos é a mais apropriada. No entanto, ao analisarmos as médias de cada atributo para diferentes números de agrupamentos (Tabelas 2 e 3), observamos que a divisão em 4 grupos define perfis mais diferenciados com base na média de cada coluna. Na Seção 4, aprofundaremos a análise dos resultados obtidos e discutiremos suas implicações mais detalhadamente.

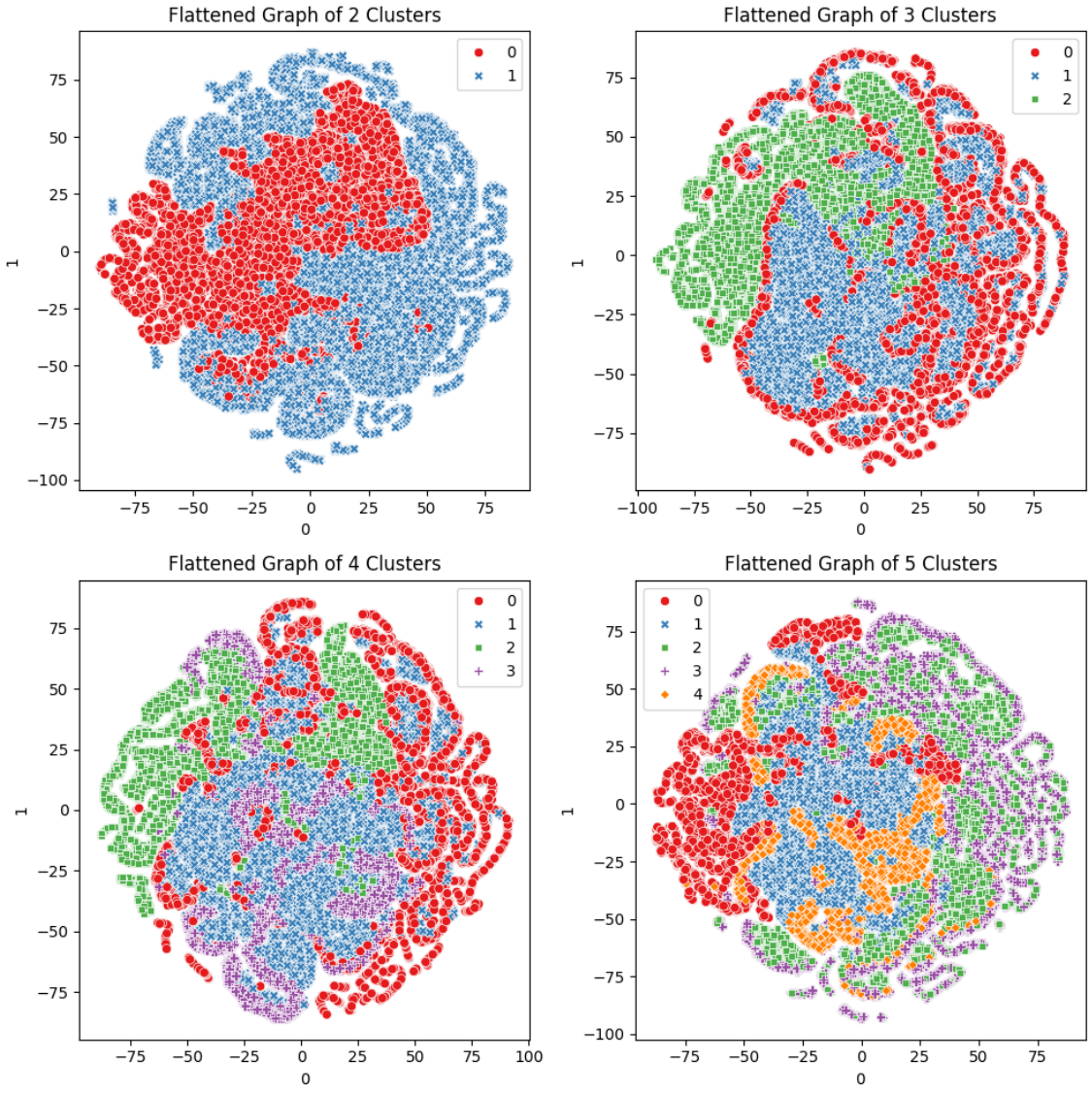

Figura 6. Visualização bidimensional dos resultados da clusterização

4.2. Avaliação dos *clusters*

Após determinarmos o número ideal de agrupamentos, é crucial avaliar os portfólios de cada ponto em cada agrupamento. Para isso, aplicamos o otimizador de média-variância a cada portfólio, juntamente com os dados históricos dos ativos. Como parâmetro de entrada para o otimizador, devemos definir os pesos de cada ativo na carteira, que representam a porcentagem investida da carteira em cada um dos ativos.

Através do *PyPortfolioOpt*, foi viabilizada a análise de cada portfólio dos agrupamentos de dados definidos, permitindo a avaliação de sua performance por meio de indicadores extraídos do otimizador de média-variância. Esses indicadores oferecem uma forma objetiva de visualizar a performance da carteira, permitindo uma análise fundamentada com base em dados históricos do portfólio selecionado.

A clusterização baseada em RFM resulta na análise dos *clusters*, essencial para delinear o perfil médio de cada cliente em cada agrupamento Gustriansyah et al., 2020. A

Cluster	Recência	Frequência	Valor Monetário
0	31,55	45,55	33.915,08
1	367,29	13,95	7.765,48
2	239,82	5,75	215,75

Tabela 2. Médias de cada *Cluster* para cada atributo RFM com *K*=3

Cluster	Recência	Frequência	Valor Monetário
0	53,59	52,43	49.435,29
1	406,79	13,15	5.874,79
2	287,14	4,76	157,05
3	23,25	20,76	2.320,71

Tabela 3. Médias de cada *Cluster* para cada atributo RFM com *K*=4

Tabela 3 fornece *insights* valiosos sobre a posição de cada cliente em relação aos atributos de recência, frequência e valor monetário.

Os clientes do *Cluster* 0 são os mais atrativos segundo a metodologia RFM. Este grupo apresenta uma movimentação financeira substancial na plataforma, uma frequência de compra elevada e uma recência relativamente menor, indicando um alto nível de engajamento recente. Esse perfil sugere que os clientes deste *cluster* estão altamente envolvidos com a plataforma e representam uma oportunidade significativa para estratégias de fidelização e aumento do valor do cliente.

É recomendável realizar uma análise detalhada dos perfis dos clientes dentro deste cluster 0. Identificar padrões de comportamento e preferências específicas pode ajudar a adaptar as ofertas e experiências da plataforma para melhor atender às necessidades desses clientes. Além disso, estratégias direcionadas para retenção e incentivo à repetição de compra podem ser altamente eficazes para maximizar o valor desses clientes.

Por outro lado, os clientes do *Cluster* 2 representam um grupo menos atraente em termos de valor do cliente. Com uma movimentação financeira modesta, baixa frequência de compra e uma recência mais longa, esses clientes podem estar menos engajados com a plataforma e podem exigir estratégias de reengajamento para aumentar seu valor ao longo do tempo.

O *Cluster* 3, por sua vez, apresenta um perfil intermediário, com uma atividade financeira média, frequência de compra e recência relativamente baixas em comparação com o *Cluster* 0, mas superiores ao *Cluster* 1. Estratégias focadas em incentivar a frequência de transações e aumentar o engajamento podem ser eficazes para atrair e reter os clientes deste grupo, maximizando assim seu potencial de valor para a plataforma.

Por fim, o *Cluster* 1 representa um grupo com características distintas, com uma recência muito alta, frequência de compra baixa e uma movimentação financeira significativa. Esses clientes podem representar oportunidades específicas de vendas adicionais ou *cross-selling*, uma vez que estão ativos na plataforma e demonstram um forte potencial de valor. Estratégias personalizadas para maximizar o valor desses clientes podem incluir ofertas exclusivas, programas de fidelidade ou recomendações de produtos relevantes.

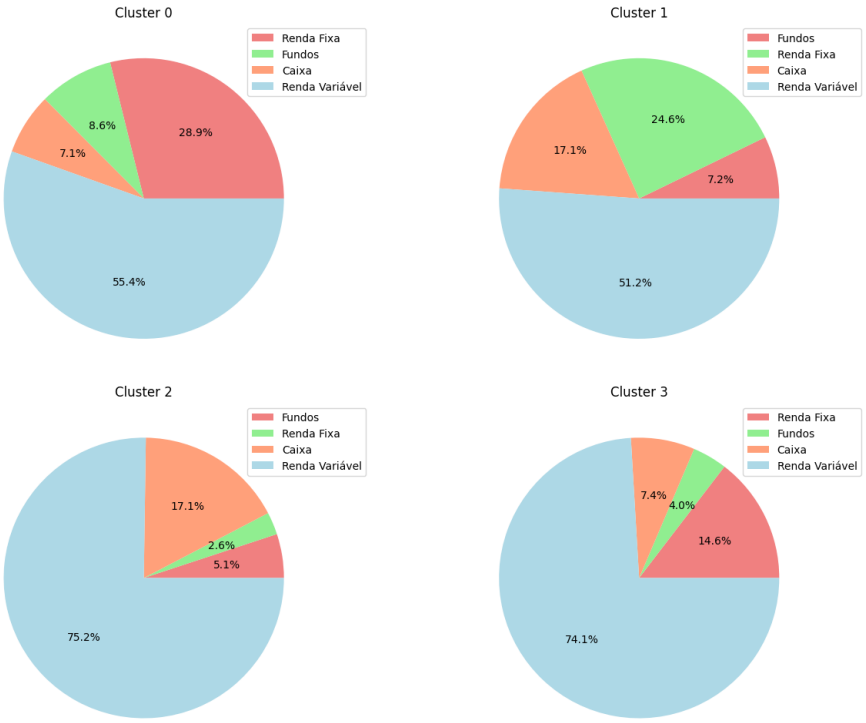

Figura 7. Alocação por produto de investimentos em cada *cluster*

4.3. Alocação de Investimentos por Classe de Ativos

Após a análise do comportamento dos clientes em cada *cluster*, torna-se fundamental examinar a distribuição por produto de investimento para compreender as semelhanças e diferenças entre os grupos. Na Figura 7, apresentamos uma visão detalhada da alocação percentual de cada produto de investimento oferecido pela corretora em cada *cluster*.

Buscar uma diversificação de portfólio é uma das estratégias confiáveis para mitigar o que for possível de risco, quando falamos de investimento financeiro Markowitz, 1952. Ao analisarmos visualmente a Figura 7, podemos observar que os *Clusters* 0 e 1 conseguem alocar uma parte significativa de seus investimentos em Renda Variável (ações e ETFs - Fundos de Índice), ao mesmo tempo em que mantêm números relevantes em Renda Fixa. Isso cria uma sensação de diversificação mais clara em comparação aos outros grupos.

4.4. Performance dos portfólios

A Tabela 4 mostra os percentis do retorno anual das carteiras de investimento para cada cluster resultante da clusterização baseada em RFM. Para cada cluster, os percentis representam os valores abaixo dos quais 10%, 25%, 50%, 75% e 90% dos retornos anuais das carteiras estão situados. Além disso, são fornecidas as médias simples e ponderadas dos retornos anuais das carteiras em cada cluster. A média ponderada considera o tamanho do investimento em cada carteira, oferecendo uma medida mais precisa do desempenho médio ponderado.

Ao examinar a Tabela 4, é evidente que o *Cluster* 3 se destaca com uma diferença entre os percentis de 10 e 90 em comparação com os outros *clusters*, sugerindo um de-

Cluster	10º	25º	50º	75º	90º	Média	Média Ponderada
0	12,63%	-3,58%	0,84%	5,25%	12,62%	1,64%	1,16%
1	14,04%	-11,35%	-0,65%	5,72%	14,04%	-1,72%	0,44%
2	12,29%	-4,59%	1,14%	5,53%	12,29%	1,14%	1,57%
3	21,09%	-5,93%	1,44%	7,47%	21,09%	2,76%	3,04%

Tabela 4. Percentis de Retorno Anual de Carteiras de Investimento

sempenho maior nessas faixas específicas. No entanto, essa discrepância não se estende uniformemente para os demais percentis. Uma análise mais aprofundada das médias simples e ponderadas revela uma tendência interessante: em três dos quatro *Clusters* (1, 2 e 3), a média ponderada supera a média simples, resultando em um aumento do retorno anual médio. Por outro lado, no Agrupamento 0, observa-se uma diminuição no valor do retorno anual. Essa variação nos resultados pode ser explicada pela distribuição do tamanho do investimento em cada carteira dentro de cada *cluster*. Conforme observado na Tabela 3, o *Cluster* 3 se destaca por ter a média de valor investido mais elevada em comparação com os outros *clusters*.

5. Conclusão

A presente pesquisa representa uma aplicação de clusterização baseada em RFM em um contexto real de dados, especificamente relacionados à portfólios do mercado financeiro. Empregamos o método de clusterização *K-Means* para agrupar os clientes com base no comportamento descrito pelo RFM, o que permitiu a análise detalhada da saúde financeira dos portfólios agrupados. Como resultado, pudemos construir perfis de alocação distintos para cada agrupamento, fornecendo *insights* valiosos para a gestão e otimização dos investimentos. Este estudo demonstra a eficácia e a relevância da abordagem de clusterização RFM para entender e segmentar a base de clientes em contextos financeiros, oferecendo um instrumento robusto para a tomada de decisões estratégicas no mercado.

Apesar dos resultados significativos obtidos neste estudo, é crucial reconhecer algumas limitações que podem impactar a interpretação e a generalização dos achados. Primeiramente, a sensibilidade a quantidade de *clusters* foi examinada apenas por uma única métrica. Além disso, devido às divergências entre o número de contas ativas e o número total de contas criadas, não foi possível utilizar 100% da base de dados disponível. Adicionalmente, a análise foi conduzida de forma estática, considerando apenas o estado das carteiras em um determinado dia, sem levar em conta sua evolução patrimonial ao longo do tempo. Destaca-se, portanto, a importância de investigações futuras que possam abordar essas questões de forma mais abrangente.

Para futuras pesquisas, o objetivo é aprimorar a qualidade e confiabilidade dos agrupamentos, além de gerar carteiras de referência para cada grupo. Para alcançar esse objetivo, é essencial: (i) incorporar técnicas de avaliação da clusterização, (ii) incorporar diferentes análises da quantidade de clusters e (iii) aplicar a técnica da fronteira eficiente em todo o agrupamento, definindo assim uma carteira de referência para cada segmento de clientes.

Referências

Aggelis, V., & Christodoulakis, D. (2005). Customer clustering using rfm analysis. *Proceedings of the 9th WSEAS International Conference on Computers*, 2.

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). RFM Analysis. Em *Database Marketing: Analyzing and Managing Customers* (pp. 323–337). Springer New York. https://doi.org/10.1007/978-0-387-72579-6_12

Chang, C. L., Lin, C. J., & Hsu, C. W. (2011). Clustering optimization in RFM analysis. *Expert Systems with Applications*, *38*(9), 11349–11356.

Cheng, C. H., & Chen, L. H. (2009). Customer clustering using RFM analysis. *Expert Systems with Applications*, *36*(3), 5946–5953.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1251–1257.

Derya, B. (2011). Incorporating RFM analysis into data mining techniques for market intelligence. *Expert Systems with Applications*, *38*(12), 14908–14914.

Ernawati, L., & Sarno, R. (2021). Knowledge-Oriented Applications in Data Mining: A Case Study of RFM Analysis in Direct Marketing. *Journal of Physics: Conference Series*, *1869*(1), 012085.

Gustriansyah, R., Suhandi, N., & Antony, F. (2020). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, *7*(1), 77–91.

Sharpe, W. F. (1998). The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, *3*, 169–185.

Syakur, M., Khotimah, B. K., Rochman, E., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: materials science and engineering*, *336*, 012017.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).