UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA - CCET

Departamento de Computação

Trabalho de Conclusão de Curso - TCC

Felipe Lopes Duarte

Comparação entre Técnicas de Redução de Dimensionalidade em Séries Temporais: Um Foco na Setorização de Ativos em Índices no Mercado Financeiro

São Carlos - SP

2025

Felipe Lopes Duarte

Comparação entre Técnicas de Redução de Dimensionalidade em Séries Temporais: Um Foco na Setorização de Ativos em Índices no Mercado Financeiro

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Federal de São Carlos, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientação Prof. Dr. Alexandre Levada

São Carlos - SP

2025

Dedico este trabalho a Fernando e Susana Duarte, cujo amor e apoio incondicional tornaram possível cada passo desta jornada. Sou eternamente grato por acreditarem em mim e transformarem meu sonho em realidade. Nós conseguimos.

Agradecimentos

Primeiramente, quero expressar minha eterna gratidão ao meu pai, Fernando, por ter me proporcionado uma educação de excelência, abrindo as portas para que eu pudesse construir as bases sólidas da minha vida adulta em uma das melhores universidades do país. O apoio emocional e financeiro que recebi durante esses cinco anos de graduação foram indispensáveis para que eu pudesse chegar até aqui. Te amo, pai! Agradeço profundamente a minha mãe, Susana, minha maior inspiração pessoal, por ter me ensinado o verdadeiro significado do amor e do cuidado. Seus ensinamentos moldaram a pessoa que sou hoje e foram cruciais para o meu desenvolvimento como homem. Nunca encontrarei palavras capazes de expressar toda a minha gratidão. Obrigado por tudo, mãe!

A minha companheira de vida, Williane, deixo um agradecimento especial pelas inúmeras ligações, viagens e visitas a São Carlos. Você sempre esteve ao meu lado, me incentivando e me mostrando o que é amar de verdade. Seu apoio foi fundamental para que eu pudesse superar os inúmeros desafios e continuasse seguindo em frente. Obrigado por viver essa fase comigo. O nosso futuro juntos será incrível.

Aos irmãos que a vida universitária me deu – Brainer, Gabriel, Murilo, Pedro, Romeu, Thiago e Vitor – minha gratidão eterna. Vocês se tornaram parte da minha família e foram companheiros indispensáveis ao longo dessa jornada. As risadas, conversas e os momentos que compartilhamos ficarão para sempre na minha memória. Sentirei muita falta da convivência com vocês. Aos professores que contribuíram para a minha formação, deixo meu reconhecimento e agradecimento. Em especial, agradeço ao professor Dr. Alexandre Levada por sua paciência, atenção e didática, que foram essenciais para o desenvolvimento deste projeto. Seu constante acompanhamento e orientação precisa foram fundamentais para alcançarmos as metas estabelecidas e o sucesso deste trabalho.

Por fim, quero agradecer a todas as pessoas que, mesmo não citadas nominalmente, tiveram um impacto na minha vida ao longo dessa trajetória. Cada palavra de incentivo, gesto de apoio ou simples presença fizeram a diferença e me ajudaram a chegar até aqui. Sou grato a todos que, de alguma forma, contribuíram para a realização deste sonho.

"Crê em ti mesmo, age e verá os resultados.
Quando te esforças, a vida também se esforça para te ajudar."
(Chico Xavier)

Resumo

Este estudo analisou a aplicação do método *Uniform Manifold Approximation and Projection* (UMAP) em comparação com o *Principal Component Analysis* (PCA) e *t-distributed Stochastic Neighbor Embedding* (t-SNE) na setorização de ativos em índices financeiros, com foco em séries temporais. O problema central aborda como a escolha entre métodos lineares, representados pelo PCA, e não lineares, como t-SNE e UMAP, impacta a preservação de informações cruciais para a setorização eficaz de ativos em índices como o Ibovespa e o S&P 500. A pesquisa preenche uma lacuna na literatura ao explorar a especificidade dessa escolha, destacando sua importância técnica e estratégica ao balancear eficiência e eficácia. Para conduzir os experimentos, foram analisados diferentes horizontes temporais, aplicando técnicas de redução de dimensionalidade para transformar séries temporais financeiras e, posteriormente, agrupando os ativos resultantes com algoritmos de clusterização como K-Means, HDBSCAN e GMM. As métricas utilizadas incluíram o Índice de Silhueta, que avalia a consistência dos *clusters*, e o Índice de Calinski-Harabasz, que mede a separação entre eles. Os resultados ressaltam a Redução de Dimensionalidade como uma ferramenta essencial para superar os desafios impostos pela Maldição da Dimensionalidade em análises financeiras. Métodos como o UMAP demonstraram ser especialmente eficazes ao revelar padrões estruturais em dados complexos e multidimensionais, superando limitações de técnicas lineares. Este trabalho não apenas reforça a relevância dessas técnicas em aplicações práticas, mas também oferece uma sólida base para futuras pesquisas e para o desenvolvimento de soluções voltadas a análise de dados financeiros de alta dimensionalidade.

Palavras-chave: UMAP, PCA, t-SNE, Redução de Dimensionalidade, Séries Temporais, Setorização de Ativos, Clustering, Mercado de Capitais, Algoritmos Não Lineares, Índices Financeiros.

Abstract

This study analyzed the application of the Uniform Manifold Approximation and Projection (UMAP) method in comparison to Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) in the clustering of assets in stock indexes, focusing on time series. The central problem addresses how the choice between linear methods, represented by PCA, and non-linear methods, such as t-SNE and UMAP, impacts the preservation of crucial information for the effective clustering of assets in indexes like Ibovespa and S&P 500. The research fills a gap in the literature by exploring the specificity of this choice, highlighting its technical and strategic importance in balancing efficiency and effectiveness. To conduct the experiments, different time horizons were analyzed by applying dimensionality reduction techniques to transform financial time series and subsequently clustering the resulting assets using algorithms such as K-Means, HDBSCAN, and GMM. The metrics used included the Silhouette Index, which evaluates the consistency of the *clusters*, and the Calinski-Harabasz Index, which measures their separation. The results emphasize Dimensionality Reduction as an essential tool to overcome the challenges posed by the Curse of Dimensionality in financial analyses. Methods such as UMAP proved particularly effective in revealing structural patterns in complex and multidimensional data, overcoming the limitations of linear techniques. This work not only reinforces the relevance of these techniques in practical applications but also provides a solid basis for future research and the development of solutions focused on analyzing high-dimensional financial data.

Keywords: UMAP, PCA, t-SNE, Dimensionality Reduction, Time Series, Asset Clustering, Market Segmentation, Capital Markets, Non-Linear Algorithms, Financial Ratios.

Lista de Ilustrações

Sumário

1 Introdução

A constante evolução do mercado financeiro ao longo das últimas décadas tornou mais complexas e diversificadas as operações e a variedade de ativos disponíveis. Nesse ambiente multifacetado, investidores enfrentam o desafio de identificar as aplicações que melhor se adequam às suas estratégias de investimento. Essa ampla oferta de opções exige métodos cada vez mais robustos para identificar oportunidades de valorização e otimizar a tomada de decisão.

Com isso, tanto investidores individuais quanto gestores de carteiras, buscam métodos quantitativos para aprimorar suas escolhas. A análise dos dados de mercado, frequentemente marcados por ruídos e alta dimensionalidade, apresentam inúmeros desafios. A complexidade dos conjuntos de dados financeiros, que incluem múltiplas variáveis inter-relacionadas, torna difícil a classificação e interpretação precisa.

A segmentação de ativos, fundamental para a construção de carteiras diversificadas, é um dos pilares da Teoria Moderna do Portfólio, proposta por Harry Markowitz na década de 50. Essa teoria inovadora introduziu conceitos como a diversificação eficiente e a relação entre risco e retorno, afirmando que a alocação estratégica de ativos permite maximizar o retorno esperado para um determinado nível de risco ou, alternativamente, minimizar o risco para um retorno específico. Nesse contexto, a escolha criteriosa de ativos e a diversificação são essenciais para proteger e construir um patrimônio.

As técnicas de redução de dimensionalidade, como PCA, t-SNE e UMAP, emergem como ferramentas sofisticadas que possibilitam a extração e a visualização de informações relevantes em dados de alta dimensionalidade. Ao agrupar ativos com características semelhantes, essas técnicas facilitam a representação de dados complexos em um espaço de menor dimensão, preservando suas principais propriedades e tornando mais claras as correlações entre variáveis.

A aplicação dessas técnicas permite uma análise aprofundada das correlações entre ativos, contribuindo para uma alocação mais eficiente em carteiras e uma gestão mais estratégica. Em um cenário financeiro dinâmico, a redução de dimensionalidade se torna um aliado indispensável para pessoas físicas e gestores, ajudando-os a avaliar a qualidade e o desempenho de ativos, em alinhamento com suas estratégias de diversificação e controle de risco.

1.1 Organização do Trabalho

Este trabalho está estruturado em seis capítulos. O Capítulo 1 apresenta a proposta de pesquisa, contextualizando o tema. O Capítulo 2 define os objetivos a serem explorados ao longo do estudo. O Capítulo 3 é dedicado a fundamentação teórica, abordando conceitos de aprendizado de máquina, agrupamento e redução de dimensionalidade, além de fornecer justificativas para a escolha dos algoritmos utilizados e das métricas adotadas para avaliar os resultados. O Capítulo 4 descreve a metodologia empregada na execução dos experimentos, detalhando os procedimentos necessários para alcançar os objetivos propostos. O Capítulo 5 concentra-se na análise dos resultados obtidos. Por fim, o Capítulo 6 apresenta as conclusões do trabalho, sintetizando os principais achados e contribuições.

2 Objetivos

2.1 Objetivos Gerais

Ao concluir este estudo, o principal objetivo é estabelecer uma condição verificável por meio de métricas específicas do aprendizado de máquina e da redução de dimensionalidade na análise dos *clusters* gerados e dos seus parâmetros intrínsecos, evidenciando a eficácia da escolha entre métodos lineares (PCA) e não lineares (t-SNE e UMAP). Esta análise tem como objetivo aplicar, comparar e avaliar os dados de setorização no contexto de diferentes mercados de capitais: o brasileiro (Ibovespa - IBOV) e o norte-americano (S&P 500).

2.2 Objetivos Específicos

Os objetivos específicos podem ser descritos como:

- Comparar a eficiência dos métodos de redução de dimensionalidade (PCA, t-SNE e UMAP) na identificação e segmentação de *clusters* de ativos financeiros em séries temporais, avaliando os resultados por meio de métricas do aprendizado de máquina.
- Examinar como a aplicação de técnicas de redução de dimensionalidade podem melhorar a visualização e interpretação das relações entre ativos financeiros, facilitando a identificação de padrões relevantes e de oportunidades estratégicas em mercados distintos.

3 Fundamentação Teórica

Os conceitos teóricos e operacionais que embasam este estudo serão detalhados neste capítulo.

3.1 Aprendizado de Máquina

O Aprendizado de Máquina (AM), um campo essencial da Inteligência Artificial, integra conceitos de estatística, matemática e ciência da computação para criar modelos que solucionem problemas complexos a partir de dados. Segundo Faceli e colaboradores (FACELI et al., 2021), o AM pode ser descrito como "a indução de uma função ou hipótese capaz de resolver um problema com base em dados que representam suas instâncias". De forma complementar, Russell e Norvig (RUSSELL; NORVIG, 2016) definem o AM como a habilidade de analisar dados, construir um modelo baseado neles e utilizá-lo como hipótese explicativa ou ferramenta para resolução de problemas. Tanto uma perspectiva quanto a outra enfatizam a importância do AM como um método eficaz para lidar com grandes conjuntos de dados, extrair *insights* valiosos e oferecer soluções inovadoras em diversas áreas do saber.

Dentro do AM, destacam-se o aprendizado supervisionado, que utiliza dados rotulados para treinar modelos e realizar previsões em novos dados, e o aprendizado não supervisionado, que explora padrões ocultos nos dados sem a necessidade de rótulos. Hardt e Recht (HARDT; RECHT, 2021) enfatizam que o AM é essencial para generalizar regras e padrões a partir de exemplos específicos, mostrando sua adaptabilidade e eficácia em cenários diversos. Combinando flexibilidade e escalabilidade, o AM tem se mostrado indispensável para resolver desafios relacionados a complexidade e alta dimensionalidade dos dados.

3.2 Aprendizado não Supervisionado

O aprendizado não supervisionado explora estruturas subjacentes em conjuntos de dados sem rótulos explícitos. Em contextos de agrupamento, os algoritmos identificam padrões e organizam os dados em grupos com características semelhantes, mesmo sem uma categorização definida previamente (ALPAYDIN, 2020). Essa abordagem é valiosa em cenários onde a rotulagem manual não é prática ou possível, permitindo *insights* sobre as relações ocultas entre os elementos analisados.

Neste trabalho, o aprendizado não supervisionado será o foco principal, com ênfase

em técnicas modernas de agrupamento e métodos para detecção de comunidades. Essas estratégias serão detalhadas nos próximos capítulos, evidenciando sua relevância na análise e organização de dados complexos.

3.3 Algoritmos de Agrupamento

Os algoritmos de clusterização são ferramentas do aprendizado não supervisionado que agrupam dados em conjuntos conhecidos como *clusters*. O objetivo é criar grupos em que os itens compartilhem alta similaridade dentro do mesmo *cluster*, enquanto apresentam diferenças significativas em relação aos itens de *clusters* distintos (JAMES, 2013). Essa abordagem é amplamente utilizada para revelar padrões ocultos e organizar informações complexas sem a necessidade de rótulos prévios.

Esses algoritmos podem ser classificados em dois grandes tipos: particionais e hierárquicos. Algoritmos particionais segmentam os dados em um número específico de *clusters*, geralmente pré-determinado, sem considerar relações hierárquicas entre os grupos. Por outro lado, algoritmos hierárquicos organizam os dados em uma estrutura de dendrograma, permitindo uma análise detalhada das relações entre os pontos e formando subgrupos em diferentes níveis (LU, 2010). A seguir, serão explorados os métodos usados por cada tipo, destacando suas vantagens e limitações.

3.3.1 K-Means

O K-Means é um dos algoritmos de agrupamento mais conhecidos e amplamente aplicados, devido a sua eficiência computacional. O processo básico envolve a atribuição iterativa de pontos de dados a *clusters* com base na menor distância entre esses pontos e os centróides dos *clusters*. Após a atribuição, os centróides são recalculados para refletir a média dos pontos atribuídos a cada grupo, e esse ciclo continua até que os centróides não sofram mudanças significativas. Essa abordagem é especialmente eficaz em situações onde os *clusters* possuem formatos esféricos e tamanhos semelhantes, tornando-o uma escolha popular para muitos problemas práticos (LU, 2010).

Apesar de sua eficácia em cenários específicos, o K-Means apresenta limitações importantes. Primeiramente, o número de *clusters* (k) deve ser definido previamente pelo usuário, o que nem sempre é trivial quando o conhecimento sobre os dados é limitado. Além disso, o algoritmo é sensível a *outliers*, que podem distorcer significativamente os resultados, e a escolha inicial dos centróides, pode levar a soluções subótimas. Outro ponto crítico é a necessidade de normalização das variáveis em dados com escalas diferentes, pois o K-Means depende da distância euclidiana, que é sensível a magnitudes discrepantes. Essas características podem comprometer seu desempenho em dados com alta dimensionalidade ou estruturas complexas (AGGARWAL, 2018).

Os passos para a execução do K-Means incluem:

1. **Definir o número de *clusters*:** O usuário especifica o valor de k, representando o número de grupos esperados.
2. **Inicializar os centróides:** Selecionam-se k pontos aleatórios como centróides iniciais.
3. **Atribuir pontos aos *clusters*:** Cada ponto é atribuído ao *cluster* cujo centróide está mais próximo, medido pela distância euclidiana.
4. **Atualizar os centróides:** Os centróides são recalculados como a média dos pontos atribuídos a cada *cluster*.
5. **Iterar até a convergência:** Repete-se o processo de atribuição e atualização até que os centróides não mudem ou atinjam um critério de parada.

3.3.2 HDBSCAN

O HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo avançado de agrupamento baseado em densidade, projetado para lidar com dados que apresentam densidades variáveis e *clusters* de tamanhos ou formas irregulares. Diferentemente de métodos como K-Means, o HDBSCAN não requer a definição prévia do número de *clusters* e é capaz de tratar pontos fora dos grupos como ruído, tornando-o mais robusto em cenários com *outliers* e estruturas complexas (MCINNES et al., 2017). Além disso, ele constrói uma hierarquia de *clusters*, permitindo a análise de relações mais detalhadas entre os dados.

A principal força do HDBSCAN está em sua flexibilidade para lidar com diferentes densidades nos dados, mas o algoritmo também apresenta desafios. Ele exige o ajuste criterioso de parâmetros, como o tamanho mínimo do *cluster*, que determina a granularidade dos agrupamentos, e pode ser mais intensivo computacionalmente, especialmente em grandes conjuntos de dados ou após a aplicação de técnicas de redução de dimensionalidade. Apesar disso, sua capacidade de identificar *clusters* bem definidos em dados não lineares o torna uma escolha valiosa para análises complexas (CAMPELLO; MOULAVI; SANDER, 2013).

Os passos básicos para a aplicação do HDBSCAN incluem:

1. **Construir o grafo de vizinhança:** Conectar os pontos dos dados com base na densidade local.
2. **Transformar as distâncias:** Ajustar as distâncias entre pontos para refletir suas densidades locais.

3. **Criar a hierarquia de *clusters*:** Organizar os dados em diferentes níveis de granularidade, representados como uma árvore hierárquica.

4. **Condensar a hierarquia:** Identificar os *clusters* mais estáveis e significativos, eliminando ruídos e agrupamentos instáveis.

5. **Classificar os pontos:** Atribuir os pontos aos *clusters* ou classificá-los como ruído, com base na densidade local e na hierarquia.

3.3.3 GMM

O GMM (*Gaussian Mixture Model*) é um algoritmo de agrupamento probabilístico que modela os dados como uma combinação de múltiplas distribuições normais (gaussianas). Ele é útil em cenários onde os *clusters* possuem fronteiras difusas, já que, ao contrário do K-Means, o GMM atribui uma probabilidade de pertencimento a cada cluster em vez de uma classificação determinística. Essa abordagem probabilística torna o GMM altamente flexível, permitindo identificar agrupamentos mais complexos em dados com variações sutis (BISHOP, 2006).

Sua flexibilidade aumenta o risco de *overfitting* - quando o algoritmo se ajusta muito de perto aos seus dados de treinamento -, especialmente em conjuntos de dados com ruídos ou com características que não seguem distribuições gaussianas. Além disso, a exigência computacional do GMM é maior do que a de métodos como o K-Means, devido a necessidade de estimar os parâmetros das distribuições gaussianas para cada *cluster*. Outro aspecto crítico é a inicialização, que pode influenciar fortemente os resultados, já que o algoritmo pode convergir para soluções locais subótimas (MURPHY, 2012).

Os passos para a aplicação do GMM incluem:

1. **Definir o número de *clusters*:** Especificar o número de distribuições gaussianas que serão ajustadas aos dados.

2. **Inicializar os parâmetros:** Determinar os valores iniciais para as médias, variâncias e pesos das gaussianas.

3. **Calcular as probabilidades de pertencimento:** Para cada ponto, calcular a probabilidade de pertencer a cada *cluster* com base nos parâmetros das gaussianas.

4. **Reestimar os parâmetros:** Atualizar as médias, variâncias e pesos das gaussianas com base nas probabilidades calculadas.

5. **Iterar até a convergência:** Repetir os passos de cálculo e atualização até que as mudanças nos parâmetros sejam insignificantes.

O GMM é eficaz em análises exploratórias de dados onde os *clusters* possuem características complexas ou não seguem formas regulares.

3.4 Medidas de Avaliação

Avaliar métodos de aprendizado não supervisionado não é uma tarefa simples, uma vez que não existem classes predefinidas ou rótulos para comparar os resultados. Essa ausência de um padrão de referência requer o uso de critérios que analisem características intrínsecas dos agrupamentos ou realizem comparações entre diferentes configurações de algoritmos. Os critérios para validação de agrupamentos são classificados em três categorias: (i) internos, que avaliam a qualidade do agrupamento com base apenas nas propriedades dos dados; (ii) externos, que verificam a aderência dos agrupamentos a uma estrutura conhecida previamente; e (iii) relativos, que comparam múltiplos agrupamentos gerados para determinar o mais adequado a um contexto específico (JAIN; DUBES, 1988). Neste trabalho, a análise foi feita baseada em dois índices: o índice Silhouette, que mede o quanto os pontos estão bem agrupados em relação a proximidade com outros *clusters*, e o índice Calinski–Harabasz, que avalia a relação entre a compactação interna dos *clusters* e a separação entre eles.

3.4.1 Coeficiente de Silhouette

O coeficiente de Silhouette avalia a qualidade dos *clusters* formados, medindo a coesão dentro dos *clusters* e a separação entre eles. Para cada ponto, calcula-se a proximidade média com os pontos do mesmo *cluster* ($a(i)$, coesão) e a proximidade média com os pontos do *cluster* mais próximo ($b(i)$, separação). A pontuação varia de -1 a 1, onde:

- Valores próximos a 1 indicam que o ponto está bem atribuído ao seu *cluster*;
- Valores próximos a 0 indicam sobreposição de *clusters*;
- Valores negativos sugerem que o ponto foi atribuído de maneira incorreta a um *cluster*.

A fórmula utilizada para calcular o coeficiente de Silhouette é:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.1}$$

Onde:

- $a(i)$: É a distância média do ponto i para todos os outros pontos dentro do mesmo *cluster*. Representa a coesão do *cluster* ao qual o ponto pertence.

- $b(i)$: É a distância média do ponto i para todos os pontos do *cluster* mais próximo, ou seja, aquele que não contém i, mas que está mais próximo em termos de distância. Representa a separação entre os *clusters*.

Em geral, o coeficiente de Silhouette é amplamente utilizado para avaliar a eficácia dos agrupamentos devido a sua simplicidade interpretativa e ampla aplicabilidade.

3.4.2 Índice de Calinski–Harabasz

O índice Calinski–Harabasz avalia a qualidade dos *clusters* com base na relação entre a dispersão interna dos *clusters* e a separação entre eles. Ele é calculado como a razão entre a soma da variância entre os *clusters* e a soma da variância dentro dos *clusters*, ajustada pelo número de *clusters* e de amostras. Valores mais altos indicam *clusters* mais compactos e bem separados, sugerindo uma melhor qualidade no agrupamento. Essa métrica é amplamente utilizada para comparar diferentes configurações de agrupamento e determinar a configuração mais adequada para os dados analisados.

Sendo:

- k: Número de *clusters*;
- n_q: Número de pontos no *cluster* q;
- c_q: Centro do *clusters* q;
- n_E: Número total de pontos (*data points*);
- c_E: Centro de todos os pontos.

Between-cluster Dispersion, B:

$$B = \sum_{q \in k} n_q (c_q - c_E)(c_q - c_E)^T \tag{3.2}$$

Within-cluster Dispersion, W:

$$W = \sum_{q \in k} \sum_{x \in \text{cluster} q} (x - c_q)(x - c_q)^T \tag{3.3}$$

Calinski-Harabasz Score, S:

$$S = \frac{B}{W} \times \frac{n_E - k}{k - 1} \tag{3.4}$$

Onde:

- B: Dispersão entre *clusters*, medindo a distância entre os centroides dos *clusters* e o centro global c_E, ponderada pelo número de pontos em cada *cluster* n_q;
- W: Dispersão dentro dos *clusters*, medindo a soma das distâncias entre os pontos x e o centro c_q do *cluster* correspondente;
- S: O valor do índice de Calinski-Harabasz, utilizado para avaliar a qualidade do agrupamento.

O índice de Calinski–Harabasz é particularmente útil em análises comparativas de diferentes configurações de agrupamento, permitindo selecionar a configuração que apresenta a melhor relação entre compactação dos *clusters* e separação entre eles.

3.5 Maldição da Dimensionalidade

A maldição da dimensionalidade refere-se a um fenômeno descrito por Richard E. Bellman, no qual o aumento das dimensões do espaço de análise torna os dados mais dispersos e difíceis de interpretar. À medida que a dimensionalidade aumenta, o volume do espaço cresce exponencialmente, tornando os dados mais esparsos e complexos, assim dificultando a identificação de padrões e correlações relevantes. Em tais casos, a quantidade de dados necessária para uma análise precisa aumenta exponencialmente, tornando o processamento computacionalmente inviável para muitas aplicações (AGHABOZORGI; Seyed Shirkhorshidi; Ying Wah, 2015). Métodos de redução de dimensionalidade ajudam a mitigar os efeitos da maldição da dimensionalidade ao identificar e descartar variáveis redundantes ou irrelevantes, otimizando a análise dos dados.

3.6 Redução de Dimensionalidade

A redução de dimensionalidade (RD) é uma técnica amplamente utilizada para transformar conjuntos de dados de alta dimensionalidade em representações mais compactas, preservando a maior quantidade possível de informações relevantes. Segundo Ghojogh et al. (2023), essas técnicas são fundamentais para lidar com a complexidade dos dados modernos, permitindo que padrões estruturais sejam identificados em um espaço reduzido, facilitando análises, visualizações e aplicações em aprendizado de máquina (GHOJOGH et al., 2023).

Essa transformação busca uma equivalência entre os conjuntos de dados em diferentes dimensões, ainda que uma relação exata de igualdade seja impossível devido a perda de detalhes ao reduzir o número de variáveis. Assim, métodos de RD tornam os dados mais acessíveis para tarefas como agrupamento e classificação (LANDALUCE-CALVO; MODROÑO-HERRÁN, 2020).

A redução de dimensionalidade pode ser aplicada de duas formas principais:

- **Seleção de características:** Consiste em selecionar apenas as variáveis mais relevantes do conjunto de dados original, descartando aquelas que são menos significativas.
- **Transformação dimensional:** Busca reduzir a redundância nos dados de entrada ao criar um conjunto menor de novas variáveis. Essas novas variáveis são combinações das originais, preservando essencialmente as mesmas informações contidas no conjunto inicial.

Segundo Wang e colaboradores (WANG et al., 2021), técnicas de redução de dimensionalidade como t-SNE e UMAP enfrentam um desafio fundamental: o equilíbrio entre a preservação das estruturas locais e globais dos dados. A escolha dos componentes a serem preservados e a formulação das funções de perda são aspectos críticos para o sucesso de algoritmos que buscam representar adequadamente os dados em espaços de menor dimensionalidade.

3.7 Algoritmos de Redução de Dimensionalidade

Esta seção visa detalhar o funcionamento dos algoritmos de redução de dimensionalidade utilizados neste trabalho.

3.7.1 PCA (*Principal Component Analysis*)

O PCA é uma técnica linear fundamental para redução de dimensionalidade, amplamente utilizada para conjuntos de dados de alta dimensionalidade com relações lineares entre variáveis. O principal objetivo do PCA é transformar os dados originais em um novo sistema de coordenadas onde a maior variação dos dados é capturada nos primeiros eixos, denominados componentes principais. Esses componentes são ordenados pela quantidade de variabilidade que retêm, permitindo representar os dados em um espaço reduzido com o mínimo de perda de informação.

Processo do PCA:

1. **Padronização dos Dados:** Para que variáveis com diferentes escalas não dominem a análise, o PCA requer a padronização dos dados. Cada variável é transformada para ter média zero e variância unitária, eliminando o efeito de magnitude.
2. **Cálculo da Matriz de Covariância:** A matriz de covariância mede como as variáveis estão relacionadas entre si. Ela mostra quais variáveis tendem a variar juntas e quais são independentes. Essa matriz é fundamental para encontrar as direções de maior variabilidade nos dados.

3. **Autovalores e Autovetores:** A análise da matriz de covariância, por meio do cálculo de autovalores e autovetores, permite decompor a variabilidade dos dados em componentes principais. Cada autovetor, associado a um autovalor, define um eixo ao longo do qual a variância é máxima. Dessa forma, os autovetores ordenados por seus autovalores (do maior para o menor) representam as direções de maior interesse para a análise, capturando a maior parte da variabilidade original dos dados.

4. **Seleção de Componentes Principais:** Com os autovalores ordenados em ordem, selecionam-se os autovetores correspondentes aos maiores autovalores para compor uma matriz de transformação. O número de componentes a ser mantido geralmente é escolhido com base na proporção de variabilidade acumulada desejada.

5. **Transformação dos Dados:** A matriz é utilizada para projetar os dados originais no espaço dos componentes principais, reduzindo a dimensionalidade ao mesmo tempo em que se preserva a variabilidade máxima possível. Essa projeção facilita a identificação de *clusters* e padrões em contextos financeiros.

De maneira resumida, o PCA é particularmente eficaz para detectar variações lineares nos dados e possui baixo custo computacional, o que o torna ideal para dados com estruturas lineares ou aproximadamente lineares.

3.7.2 t-SNE (*t-distributed Stochastic Neighbor Embedding*)

O t-SNE é uma técnica não linear projetada para preservar as proximidades locais entre pontos em espaços de alta e baixa dimensionalidade. Esse método é eficaz na visualização de agrupamentos em dados complexos, uma vez que prioriza a fidelidade das relações locais, o que o diferencia do PCA. Segundo Maaten e Hinton (MAATEN; HINTON, 2008), o t-SNE é capaz de criar mapas bidimensionais ou tridimensionais que revelam a estrutura dos dados em diferentes escalas, sendo especialmente útil para dados de alta dimensionalidade que residem em múltiplas variedades relacionadas de baixa dimensionalidade. O t-SNE é amplamente aplicado em conjuntos de dados onde a estrutura local é mais significativa do que as distâncias absolutas.

Processo do t-SNE:

1. **Construção da Distribuição de Similaridade em Alta Dimensão:** O t-SNE inicia medindo a proximidade entre os dados em seu estado original, de alta dimensão. Essa medida de proximidade é representada por uma probabilidade, que indica o quão semelhantes dois pontos são. A escolha da distribuição Gaussiana e o ajuste da perplexidade permitem controlar o nível de detalhe dessa medida.

2. **Mapeamento para o Espaço de Baixa Dimensão:** Em seguida, o algoritmo tenta mapear esses dados para um espaço de menor dimensão, geralmente bidimensional ou tridimensional, para facilitar a visualização. A distribuição t-Student é usada para modelar a relação entre os pontos nesse novo espaço, evitando aglomerações indesejadas.

3. **Minimização da Função de Perda:** O objetivo final é encontrar uma representação no espaço de baixa dimensão que preserve ao máximo a estrutura dos dados originais. Para isso, o t-SNE minimiza uma função de perda que compara as probabilidades de similaridade nos dois espaços. Essa minimização é feita através de um algoritmo de otimização, como o gradiente descendente.

Parâmetros Importantes:

- **Perplexidade:** Controla o equilíbrio entre a densidade local e a preservação global de proximidades. A escolha depende do número de pontos no conjunto de dados.
- **Número de Iterações e Taxa de Aprendizado:** Estes parâmetros influenciam a convergência e a qualidade do resultado final.

O t-SNE é extremamente útil para capturar estruturas locais em dados de alta dimensionalidade, porém é computacionalmente caro e não preserva bem as relações globais, o que limita seu uso para grandes conjuntos de dados.

3.7.3 UMAP (*Uniform Manifold Approximation and Projection*)

O UMAP é uma técnica não linear que busca preservar tanto as relações locais quanto as globais dos dados. Desenvolvido com base em topologia e teoria de grafos, o UMAP é ideal para conjuntos de dados de alta dimensionalidade e grandes volumes, sendo mais eficiente e escalável em comparação ao t-SNE. Segundo McInnes e colaboradores (MCINNES; HEALY; MELVILLE, 2020), o UMAP é baseado em geometria Riemanniana e topologia algébrica, resultando em um algoritmo prático e escalável, competitivo com o t-SNE em qualidade de visualização, mas com desempenho superior em termos de tempo de execução. Além disso, como destacado por Ghojogh e colaboradores (GHOJOGH et al., 2021), o UMAP é amplamente reconhecido como uma das técnicas de estado da arte para redução de dimensionalidade, combinando teoria algébrica e topologia para criar representações robustas e flexíveis dos dados. Ele ainda se diferencia por suportar variantes, como o DensMAP para preservação de densidade e o Parametric UMAP, que integra aprendizado profundo.

Processo do UMAP:

1. **Construção de uma Representação Fuzzy:** O UMAP começa criando um grafo de vizinhança aproximada no espaço de alta dimensionalidade. Cada ponto é conectado aos seus vizinhos mais próximos, formando uma estrutura fuzzy de similaridade. A densidade local é ajustada com base em uma distribuição de vizinhança que suaviza as conexões, garantindo uma transição fluida entre regiões.

2. **Mapeamento para o Espaço de Baixa Dimensão:** A estrutura fuzzy é projetada em um espaço de baixa dimensão. O UMAP ajusta as distâncias entre pontos para preservar tanto as estruturas locais quanto as globais da estrutura original. A função de perda é baseada na entropia cruzada, que mede a discrepância entre as distribuições fuzzy no espaço original e no espaço reduzido.

3. **Otimização por Gradiente Descendente:** Para otimizar a correspondência entre as representações no espaço de alta e baixa dimensionalidade, o UMAP utiliza gradiente descendente para minimizar a função de perda e ajustar as posições dos pontos.

Parâmetros Principais:

- **Número de Vizinhos:** Define o tamanho do contexto local a ser preservado, o que impacta a preservação de estruturas locais e globais.
- **Mínimo de Distância:** Controla a densidade do layout no espaço de baixa dimensionalidade, permitindo que pontos próximos fiquem mais agrupados ou dispersos.

O UMAP oferece uma flexibilidade superior, capturando relações não lineares detalhadas e preservando tanto a estrutura local quanto a global dos dados.

3.8 Índice Financeiros

O Ibovespa e o S&P 500 são índices financeiros que desempenham papéis centrais em seus respectivos mercados, oferecendo um panorama abrangente do comportamento das ações mais relevantes. O Ibovespa, o índice de referência do mercado brasileiro, foi criado em 1968 e é amplamente reconhecido como um termômetro da economia nacional. Ele é composto por uma cesta de ações das empresas mais líquidas listadas - que possuem o maior volume financeiro de negociação - cobrindo diversos setores como bancos, energia, infraestrutura, consumo e *commodities*. Esse índice não apenas mede o desempenho agregado dessas empresas, mas também reflete a interação dinâmica entre fatores econômicos internos, como políticas fiscais e monetárias, e eventos globais.

A metodologia de composição do Ibovespa é baseada em critérios de liquidez, ajustados quadrimestralmente, assegurando sua capacidade de captar mudanças estruturais

no mercado brasileiro. Como índice de retorno total, incorpora não apenas a valorização dos ativos, mas também os dividendos pagos, permitindo uma visão integrada do retorno proporcionado aos investidores. Além disso, o Ibovespa também reflete o caráter emergente do mercado brasileiro, caracterizado por alta volatilidade, exposição a choques externos e interdependência com variáveis macroeconômicas.

Por outro lado, o S&P 500, criado em 1957, é amplamente considerado um dos mais importantes índices globais devido a sua representatividade e robustez. Ele reflete o desempenho das 500 maiores empresas de capital aberto dos Estados Unidos, abrangendo aproximadamente 80% do valor total do mercado acionário americano. Sua metodologia de ponderação por capitalização de mercado ajustada por *free float* - porcentagem das ações de uma empresa que está disponível para negociação livremente no mercado - garante que o índice capture o impacto real das empresas no mercado, enquanto sua diversificação setorial – incluindo tecnologia, saúde, energia, e serviços financeiros – permite que ele funcione como um termômetro da economia americana e, muitas vezes, global.

As empresas que compõem o índice têm operações e receitas altamente internacionalizadas, tornando-o sensível a mudanças em cadeias de suprimentos, inovações tecnológicas e políticas globais. Além disso, como os mercados norte-americanos frequentemente lideram tendências globais, o S&P 500 é frequentemente usado como referência para estratégias de investimento e estudos.

A escolha do Ibovespa e do S&P 500 como objetos de estudo é justificada pela complementaridade entre seus contextos econômicos. O Ibovespa, representativo de um mercado emergente com alta volatilidade e suscetibilidade a eventos locais, contrasta com o S&P 500, que reflete um mercado mais maduro e eficiente. Essa dualidade permite avaliar a capacidade das técnicas de redução de dimensionalidade em lidar com diferentes desafios, como a presença de padrões não lineares, sazonalidades e tendências globais que desafiam os modelos tradicionais de análise.

Além disso, a interseção entre o aprendizado de máquina e a análise de diferentes mercados de capitais oferece uma oportunidade única para explorar novas abordagens na análise de dados financeiros. Técnicas como PCA, t-SNE e UMAP, ao reduzir a dimensionalidade dos dados, facilitam a visualização de padrões complexos e a identificação de relações não lineares entre os ativos. Os insights obtidos a partir dessa análise podem ser utilizados para aprimorar a tomada de decisão de investidores e gestores de portfólio, permitindo a construção de estratégias de investimento mais eficazes e a gestão de riscos mais precisa.

4 Metodologia

Este capítulo tem como objetivo apresentar a metodologia empregada na pesquisa, detalhando os dados utilizados, os algoritmos aplicados e os procedimentos experimentais.

4.1 Explicação do Escopo de Trabalho

Este trabalho investiga como diferentes técnicas de redução de dimensionalidade influenciam na setorização de ativos financeiros em séries temporais, aplicadas a índices como o Ibovespa e o S&P 500.

Ao aplicar essas técnicas, os resultados foram utilizados como base para algoritmos de clusterização, que agrupam os ativos de forma significativa. Esses agrupamentos, avaliados pelas métricas citadas anteriormente, possibilitam a identificação de padrões ocultos e facilitam a segmentação de ativos nos índices Ibovespa e S&P 500. Assim, este trabalho não apenas avalia a eficácia das técnicas de redução de dimensionalidade, mas também evidencia suas contribuições para a análise de séries temporais financeiras e o suporte a tomada de decisões no mercado financeiro.

4.2 Extração do Conjunto de Dados

A extração dos dados históricos foi realizada utilizando a *Application Programming Interface* (API) do Yahoo Finance. Essa API é amplamente utilizada para coletar informações financeiras detalhadas, oferecendo uma interface prática para análise de ativos em diferentes mercados. A escolha dessa ferramenta se dá pela sua acessibilidade, robustez e capacidade de integrar-se facilmente a fluxos de trabalho baseados em Python.

Por meio dessa API é possível definir o período histórico de análise por meio da seleção de datas de início e término. Essa decisão é sensível, pois o intervalo escolhido impacta diretamente os dados retornados, influenciando a análise de clusterização. Por exemplo, mudanças significativas no mercado, como crises financeiras ou períodos de alta volatilidade, podem afetar os padrões identificados nos dados. Além disso, um intervalo inadequado pode introduzir ruídos ou perder tendências importantes, afetando a qualidade das conclusões obtidas.

Para evitar inconsistências causadas por ativos que não estavam listados no período selecionado, a data inicial mais comum entre todos os *tickers* - código de um ativo na bolsa de valores - foi utilizada como critério de filtragem. Assim, apenas os ativos com dados completos a partir dessa data foram considerados, garantindo a uniformidade dos

dados analisados. Essa etapa é fundamental para evitar que ativos recém-listados ou com histórico incompleto introduzam discrepâncias nos *clusters* formados, preservando a integridade da análise.

4.2.1 Atributos do Conjunto de Dados

Após a extração dos dados históricos, um *Dataframe* foi estruturado de forma a facilitar a análise e o pré-processamento dos dados financeiros.

O índice do *Dataframe* é um *MultiIndex*, isto é, ele possui dois níveis de hierarquia para as colunas. No primeiro nível, estão as métricas financeiras, como *Open, High, Low, Close, Adjusted Close* e Volume, e, no segundo nível, os *tickers* correspondentes de cada ativo. Essa estrutura é especialmente útil para organizar grandes volumes de dados, permitindo que diferentes dimensões sejam acessadas e manipuladas com facilidade.

Atributos do *Dataframe*:

- **Data**: Representa as datas de negociação, incluindo apenas dias úteis no intervalo especificado.
- **Open**: Preço de abertura das ações em cada dia de negociação.
- **High**: O preço mais alto registrado durante o dia.
- **Low**: O preço mais baixo registrado durante o dia.
- **Close**: O preço de fechamento das ações no final do dia.
- **Adjusted Close**: Preço de fechamento ajustado, corrigido para eventos corporativos, como (i) *splits* - processo no qual as empresas dividem suas ações para aumentar o número de ativos em circulação - ou (ii) dividendos - parcela do lucro líquido que uma empresa de capital aberto ou fechado distribui para seus acionistas.
- **Volume**: Número de ações negociadas durante o dia, fornecendo uma medida da liquidez e atividade do mercado.

4.3 Pré-Processamento

No pré-processamento, valores faltantes foram preenchidos por meio de interpolação linear, utilizando a média dos valores adjacentes na mesma coluna, preservando a continuidade dos dados financeiros. Linhas duplicadas também foram removidas para evitar redundâncias e assegurar a integridade do conjunto de dados. Além disso, os valores extraídos foram convertidos para *arrays* NumPy, permitindo operações matemáticas e análises subsequentes de forma mais eficiente e otimizada, adequando os dados para as etapas seguintes do estudo.

4.4 Análise do Atributo Destaque

Para a redução dimensional, os log-retornos foram selecionados como meta-atributo devido à sua eficácia na normalização dos dados. Isso ocorre porque eles são menos influenciados pelas diferentes magnitudes de preço dos ativos, proporcionando uma representação mais uniforme. O log-retorno é calculado conforme fórmula abaixo, onde P_0 é o preço inicial e P_1 é o preço final do ativo.:

$$r = \ln\left(\frac{P_1}{P_0}\right) \tag{4.1}$$

O log-retorno captura a variação proporcional entre os preços de forma que distorce menos a análise em comparação aos retornos simples. Em vez de calcular a variação percentual simples, que é dada por $\frac{P_1 - P_0}{P_0}$, o log-retorno utiliza o logaritmo natural da razão entre os preços, o que possui algumas vantagens adicionais.

Os log-retornos são preferidos em análises financeiras por várias razões. Primeiramente, eles são aditivos ao longo do tempo. Isso significa que os retornos em períodos sucessivos podem ser somados diretamente, o que facilita o cálculo do retorno total em um intervalo de tempo maior. Essa propriedade é especialmente útil quando se analisa o desempenho de um ativo ao longo de vários períodos, pois elimina a necessidade de multiplicar os retornos individuais. Para períodos $T_1, T_2, \ldots, T_n$, o retorno total é simplesmente a soma dos log-retornos de cada período:

$$r_{\text{total}} = r_1 + r_2 + \cdots + r_n \tag{4.2}$$

Além disso, os log-retornos possuem a vantagem de aproximar o comportamento dos preços de ativos de uma forma mais contínua, já que as variações percentuais podem ser pequenas, mas ainda assim influenciam diretamente o valor do log-retorno. Isso é particularmente útil em mercados financeiros, onde os preços podem variar amplamente e o comportamento de ativos com grandes variações pode ser mais bem representado dessa maneira.

Neste contexto, o uso do *Adjusted Close* foi preferido ao *Close*, pois o primeiro ajusta os preços históricos para eventos corporativos, como pagamento de dividendos e *splits* de ações. Esse ajuste é fundamental, pois o *Close* pode refletir quedas ou variações bruscas no preço que não correspondem a mudanças reais no valor do ativo, mas sim a ajustes mecânicos.

Por exemplo, quando uma empresa paga dividendos, o preço de fechamento (*Close*) geralmente sofre uma queda equivalente ao valor do dividendo, o que pode ser interpretado erroneamente como uma perda no valor do ativo. Por outro lado, o *Adjusted Close* corrige os preços anteriores para refletir esse pagamento de forma precisa, apresentando uma

medida mais fiel do retorno real do investidor. Da mesma forma, o *Adjusted Close* também ajusta os preços durante *splits* de ações, evitando distorções na análise.

Esse ajuste no preço garante que o cálculo dos log-retornos seja mais confiável, uma vez que ele elimina distorções e possibilita que os retornos reflitam com precisão o comportamento real dos ativos ao longo do tempo. Como os log-retornos são baseados em proporções, eles capturam a volatilidade e o comportamento relativo dos ativos, revelando padrões importantes, como *clusters* de ativos com comportamentos semelhantes. Essa característica é relevante para o agrupamento e análise setorial, já que, ao representar a variação diária de cada ativo, a redução de dimensionalidade facilita a visualização de relações estruturais entre ativos, identificando grupos com características similares, como setores com perfis de risco semelhantes.

4.5 Estruturação das Janelas de Análise

Os experimentos foram estruturados com base em diferentes janelas temporais, buscando capturar tanto as oscilações de curto prazo quanto os padrões de longo prazo que refletem ciclos completos de mercado. Janelas mais curtas são particularmente úteis para identificar volatilidades imediatas e eventos de impacto súbito, como mudanças abruptas na política monetária ou flutuações no preço de *commodities*. Já os períodos mais longos permitem observar tendências estruturais e a resiliência de setores durante crises econômicas e recuperações. Essa abordagem é fundamental para compreender as nuances do comportamento de ativos em índices como o Ibovespa e o S&P 500, onde a dinâmica de mercado e a natureza das empresas listadas variam significativamente. A escolha das diferentes janelas impacta diretamente a setorização final, pois define a granularidade das informações analisadas, influenciando a formação dos *clusters* e, consequentemente, a interpretação das relações entre ativos ao longo do tempo.

Para janelas curtas, foram consideradas análises trimestrais e anuais, que capturam volatilidades de curto prazo e movimentos sazonais. Já as janelas longas englobam períodos de 10 anos, possibilitando a identificação de ciclos econômicos completos e tendências estruturais. Essa estrutura foi aplicada separadamente para os ambos os índices e a base de dados utilizada reflete informações extraídas até 9 de dezembro de 2024.

4.6 Algoritmos e Bibliotecas Utilizadas

Com base nas informações obtidas, os métodos de redução de dimensionalidade PCA, t-SNE e UMAP, serão aplicados, seguidos da utilização de três diferentes técnicas de agrupamento para avaliar a formação de *clusters*: K-Means, HDBSCAN e GMM. Cada técnica possui características distintas, com limitações e qualidades que impactam

diretamente nos resultados do agrupamento, avaliados por meio das métricas de Índice de Silhueta e Índice de Calinski-Harabasz.

As implementações utilizam bibliotecas de referência para vizualização e análise de dados como Pandas, NumPy, Matplotlib, Plotly, e ferramentas específicas para clusterização e redução de dimensionalidade, incluindo Scikit-learn e as implementações do HDBSCAN e UMAP em Python.

Para garantir reprodutibilidade e consistência nos resultados, todos os métodos foram configurados para serem determinísticos. O número de *clusters* foi definido em 11 para o K-Means e o GMM, refletindo a quantidade de setores da economia identificados no conjunto de dados. Esses setores incluem Energia, Imobiliário, Saúde, Consumo Cíclico, Serviços Públicos, Tecnologia, Serviços Financeiros, Comunicação, Consumo Básico, Materiais Básicos e Industrias. Para o HDBSCAN, o parâmetro de número mínimo de *clusters* foi definido em 5. Essa configuração foi escolhida para equilibrar a qualidade dos agrupamentos e evitar a formação de *clusters* excessivamente pequenos ou *outliers*.

5 Experimentos e Resultados

Este capítulo apresenta os resultados dos experimentos realizados conforme a metodologia descrita no Capítulo 4. A análise foi dividida em duas seções, uma para o Ibovespa e outra para o S&P 500, considerando horizontes de tempo trimestral, anual e 10 anos.

5.1 Ibovespa

5.1.1 Janela Trimestral (08/09/2024 - 08/12/2024)

Para a análise do Ibovespa na janela trimestral, abrangendo o período de 8 de setembro de 2024 a 8 de dezembro de 2024, foram identificadas 85 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 1.

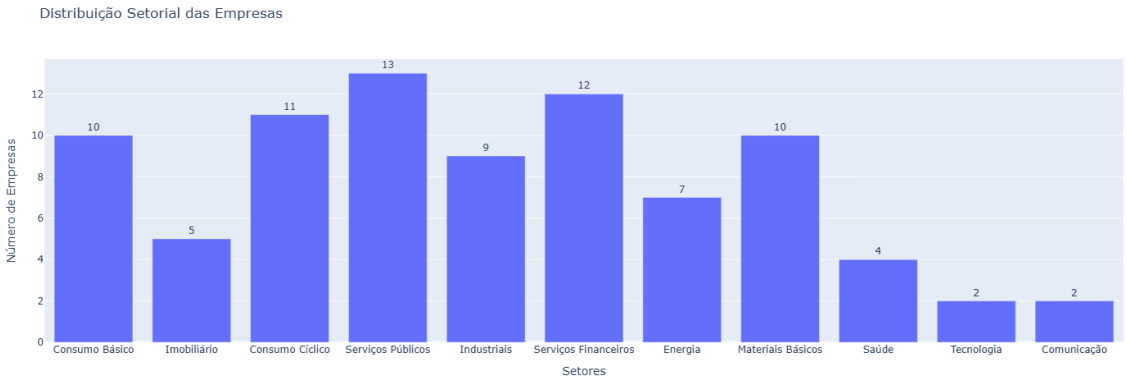

Figura 1 – Distribuição Setorial das Empresas do Ibovespa na Janela Trimestral

Price	Adj Close										...	Volume					
Ticker	ABEV3.SA	ALOS3.SA	ALPA4.SA	ASAI3.SA	AURE3.SA	AZUL4.SA	AZZA3.SA	B3SA3.SA	BBAS3.SA	BBDC3.SA	...	TOTS3.SA	UGPA3.SA	USIM5.SA	VALE3.SA	VAMO3.SA	VBBR3.SA
count	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	...	6.300000e+01	6.300000e+01	6.300000e+01	6.300000e+01	6.300000e+01	6.300000e
mean	12.919683	21.771110	6.960317	7.429682	10.293809	5.374921	40.956474	10.700169	25.926220	12.679233	...	3.898043e+06	5.562368e+06	1.363673e+07	2.211848e+07	1.024700e+07	7.958370e
std	0.419812	0.913207	0.348897	0.729129	0.415135	0.580100	3.126347	0.738941	0.941952	0.791742	...	2.227063e+06	2.106549e+06	7.070531e+06	1.009674e+07	6.678019e+06	3.371502e
min	12.330000	19.260000	6.000000	6.160000	9.460000	4.040000	33.414856	9.150000	24.309633	10.962749	...	1.084500e+06	2.076500e+06	5.889500e+06	1.020440e+07	4.229100e+06	3.564700e
25%	12.675000	21.657154	6.860000	6.990000	10.075000	5.020000	39.711475	10.250000	25.313692	12.085982	...	2.542200e+06	4.016200e+06	9.447600e+06	1.569300e+07	6.913850e+06	5.853350e
50%	12.860000	21.988001	7.030000	7.230000	10.260000	5.360000	40.650314	10.710000	25.685240	12.951722	...	3.450400e+06	4.833400e+06	1.202020e+07	1.904490e+07	8.432300e+06	7.502700e
75%	13.060000	22.292941	7.180000	7.710000	10.440000	5.880000	41.657970	10.885000	26.469633	13.244562	...	4.342400e+06	6.938700e+06	1.505615e+07	2.533425e+07	1.046850e+07	9.157650e
max	14.420000	23.009085	7.430000	9.250000	11.170000	6.250000	47.522038	12.334406	28.199848	13.994479	...	1.164110e+07	1.303430e+07	4.662200e+07	6.144930e+07	4.842620e+07	1.900890e

8 rows × 510 columns

Figura 2 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 2, o conjunto de dados é composto por 251 registros e 510 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 421 variáveis são do tipo *float*, enquanto 83 são do tipo inteiro, totalizando uma estrutura

de aproximadamente 990.3 KB em memória. Além disso, o conjunto resultante não possui dados faltantes ou linhas duplicadas.

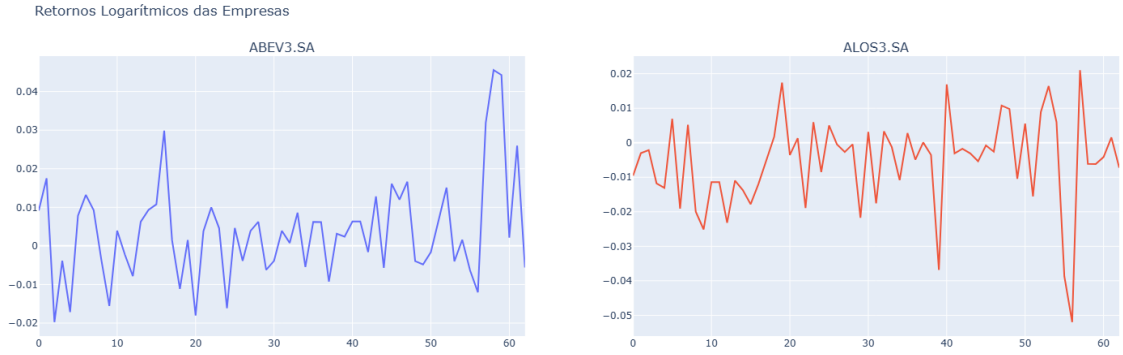

Figura 3 – Série Temporal do Retorno Logaritmo da Ambev e do Grupo Allos na Janela Trimestral

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.453	305.366
PCA	HDBSCAN	0.073	11.289
PCA	GMM	0.443	286.715
t-SNE	KMeans	0.339	91.830
t-SNE	HDBSCAN	0.074	17.230
t-SNE	GMM	0.261	65.500
UMAP	KMeans	0.365	165.232
UMAP	HDBSCAN	0.254	84.257
UMAP	GMM	0.363	166.069

Figura 4 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela Trimestral

Assim, é possível analisar que o PCA em conjunto com o KMeans resultou nos melhores resultados, conforme Figura 4.

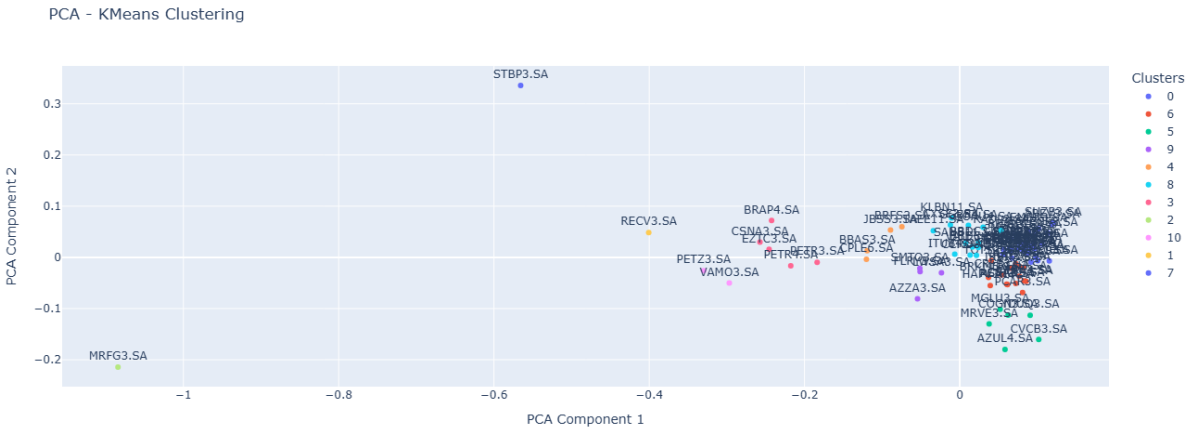

Figura 5 – Clusterização Resultante do PCA + KMeans

Dentro do período analisado, correspondente ao trimestre em questão, conforme ilustrado na Figura 5, comportamentos atípicos foram observados em algumas ações, como as de Santos Brasil (STBP3) e Marfrig (MRFG3), que se destacaram como *outliers*. No caso de Santos Brasil, a transação de venda da empresa para o grupo CMA CGM gerou uma distorção significativa em seus preços, refletindo um evento corporativo de grande impacto. Já as ações da Marfrig sofreram uma pressão externa devido à valorização do dólar, o que levou a flutuações inesperadas no seu desempenho. Esses eventos excepcionais destacaram essas ações como desvios notáveis, com comportamentos que se afastaram consideravelmente das tendências gerais do mercado durante o trimestre.

Adicionalmente, observou-se que empresas pertencentes a setores semelhantes foram agrupadas de forma coesa, revelando a eficácia dos métodos de clusterização aplicados na análise. Por exemplo, as empresas Azul (AZUL4) e CVC (CVCB3), ambas atuantes no setor de turismo, apresentaram perfis de comportamento de mercado muito semelhantes, sendo alocadas no mesmo *cluster*. O mesmo ocorreu com as empresas Engie (EGIE3), Copel (CPLE6) e Eletrobras (ELET3), que fazem parte do setor elétrico. Esses agrupamentos evidenciam a capacidade dos métodos de redução dimensional e de clusterização em identificar padrões estruturais no comportamento de ativos, com ênfase nas dinâmicas setoriais, que são cruciais para a construção de estratégias de investimento mais informadas e precisas.

5.1.2 Janela Anual (08/12/2023 - 08/12/2024)

Para a análise do Ibovespa na janela anual, abrangendo o período de 8 de dezembro de 2023 a 8 de dezembro de 2024, foram identificadas 84 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 6.

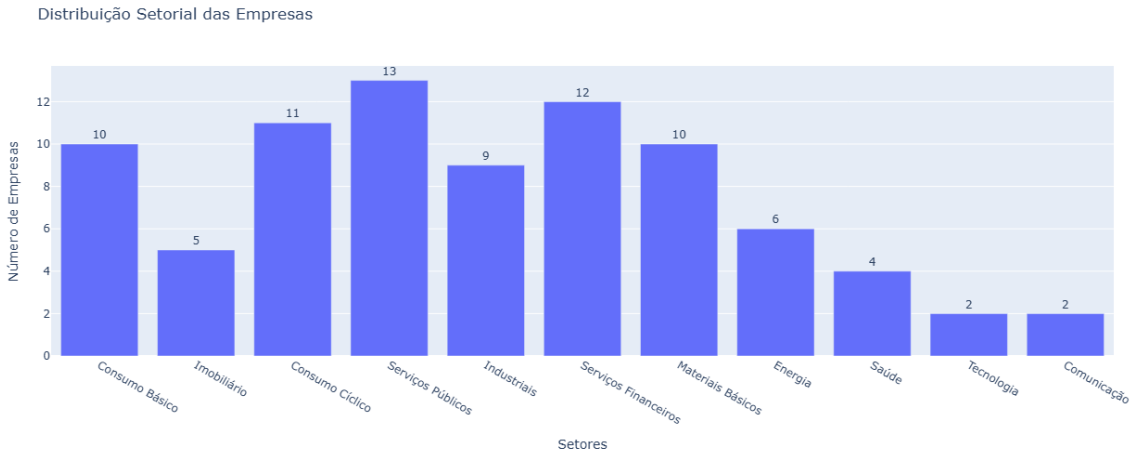

Figura 6 – Distribuição Setorial das Empresas do Ibovespa na Janela Anual

Price	Adj Close										...	Volume		
Ticker	ABEV3.SA	ALOS3.SA	ALPA4.SA	ASAI3.SA	AURE3.SA	AZUL4.SA	AZZA3.SA	B3SA3.SA	BBAS3.SA	BBDC3.SA	...	TOTS3.SA	UGPA3.SA	USIM5.SA
count	251.000000	250.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	...	2.510000e+02	2.510000e+02	2.510000e+02
mean	12.584059	22.151856	8.583147	11.279602	11.711204	9.487888	50.749318	11.590242	25.962835	12.272162	...	3.726074e+06	5.108423e+06	1.326654e+07
std	0.714738	1.188089	1.112014	2.684576	0.940642	3.470037	7.447520	1.168984	0.839633	1.062546	...	2.212141e+06	2.834167e+06	1.025392e+07
min	11.090000	19.260000	6.000000	6.160000	9.460000	4.040000	33.414856	9.150000	24.218946	10.683255	...	1.084500e+06	1.264600e+06	3.432900e+06
25%	12.030000	21.433810	7.425000	9.305000	11.245000	6.095000	46.720827	10.710000	25.332639	11.338618	...	2.311650e+06	3.328200e+06	8.477750e+06
50%	12.670000	22.148986	8.910000	11.800000	11.950000	9.180000	50.235332	11.290200	25.753448	12.014587	...	3.167500e+06	4.500300e+06	1.099760e+07
75%	13.030000	22.889052	9.460000	13.585000	12.425311	12.455000	57.101858	12.438515	26.520856	13.176059	...	4.339650e+06	5.914700e+06	1.467105e+07
max	14.420000	25.181198	10.320000	14.930000	13.099144	16.840000	66.848343	14.344547	28.316015	14.355443	...	1.828540e+07	2.022220e+07	1.121265e+08

8 rows × 504 columns

Figura 7 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 7, o conjunto de dados é composto por 251 registros e 504 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 421 variáveis são do tipo *float*, enquanto 83 são do tipo inteiro, totalizando uma estrutura de aproximadamente 990.3 KB em memória. Além disso, o conjunto resultante não possui linhas duplicadas e os valores faltantes foram preenchidos com a média dos valores adjacentes.

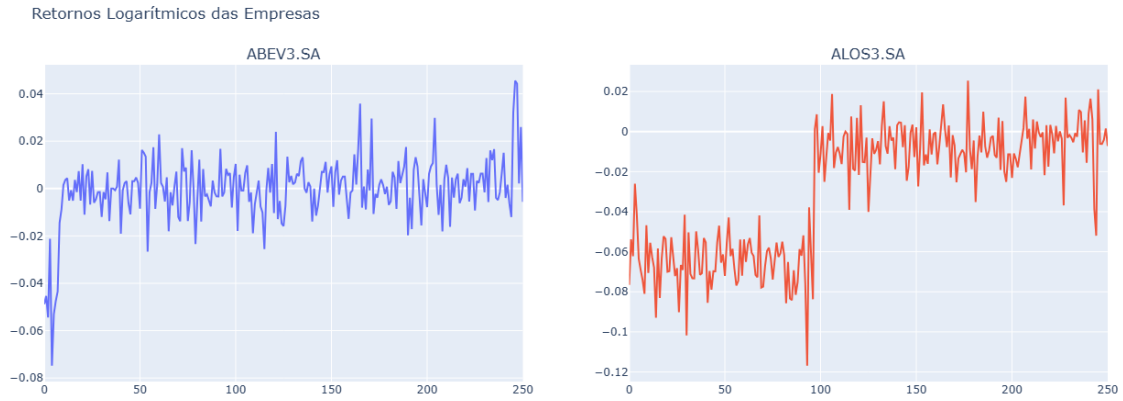

Figura 8 – Série Temporal do Retorno Logaritmo da Ambev e do Grupo Allos na Janela Anual

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.393	218.172
PCA	HDBSCAN	0.220	35.812
PCA	GMM	0.400	212.476
t-SNE	KMeans	0.364	171.634
t-SNE	HDBSCAN	0.242	39.548
t-SNE	GMM	0.340	148.512
UMAP	KMeans	0.378	345.393
UMAP	HDBSCAN	0.502	137.784
UMAP	GMM	0.360	329.663

Figura 9 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela Anual

Assim, é possível analisar que a combinação UMAP + HDBSCAN apresentou os melhores resultados no Índice Silhouette, e a combinação UMAP + KMeans obteve os melhores resultados no Índice Calinski-Harabasz, conforme Figura 9.

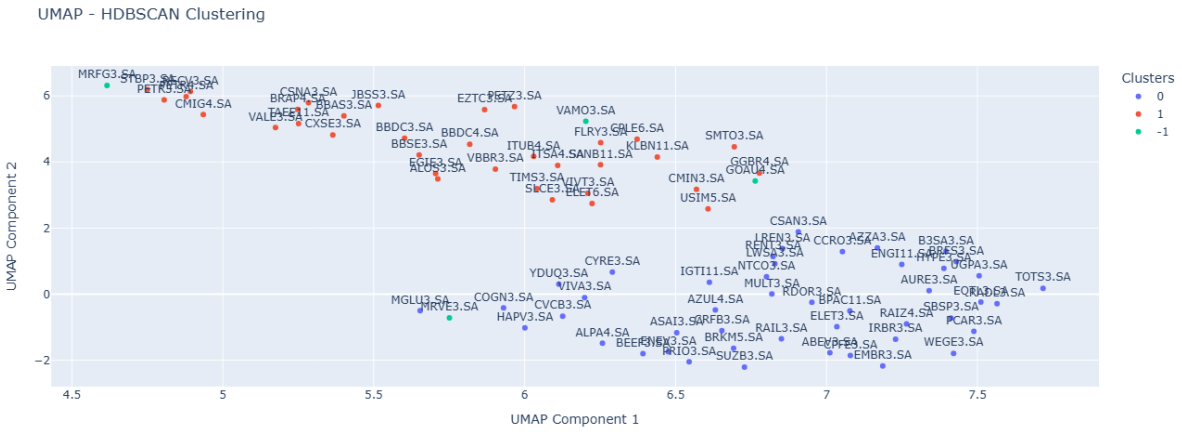

Figura 10 – Clusterização resultante do UMAP + HDBSCAN

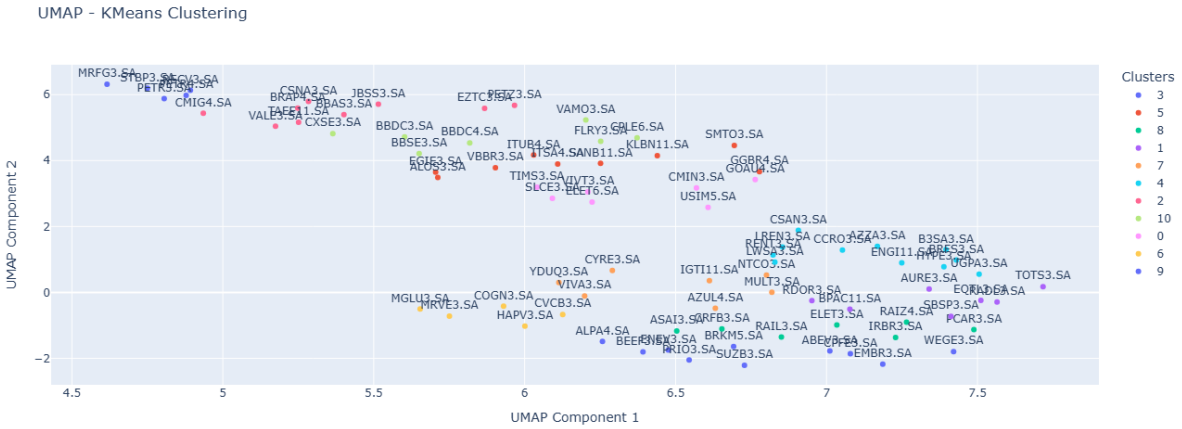

Figura 11 – Clusterização resultante do UMAP + KMeans

Dentro do período anual analisado, conforme ilustrado na Figura 11, os comportamentos atípicos observados no espectro trimestral foram, em grande parte, normalizados. Empresas como Santos Brasil (STBP3) e Marfrig (MRFG3), que anteriormente se destacaram como *outliers*, apresentaram dinâmicas mais próximas das observadas em outras empresas do mercado, o que facilitou sua inclusão em *clusters*. No caso de Santos Brasil, a transação de venda para o grupo CMA CGM, que antes causava grandes distorções nos preços, teve seu impacto suavizado ao longo do tempo, enquanto Marfrig, que era fortemente influenciada pela alta do dólar, passou a mostrar comportamentos mais alinhados com as tendências de outras empresas do setor.

Além disso, a análise revelou que empresas pertencentes a grupos com características estruturais e econômicas semelhantes continuaram a ser agrupadas de forma coesa dentro dos mesmos *clusters*. Exemplos claros disso incluem Suzano (SUZB3), Minerva (BEEF3) e Braskem (BRKM5), cujos comportamentos de mercado, em grande parte, refletiram as influências de fatores macroeconômicos que afetam diretamente o desempenho de empresas

desses setores. A forte exposição dessas companhias à variação cambial, bem como sua dependência das flutuações na demanda por *commodities*, foi um fator determinante para suas performances correlacionadas ao longo do ano. A demanda global por *commodities* e as oscilações cambiais impactaram não só o desempenho financeiro, mas também as projeções futuras dessas empresas, o que resultou em um comportamento de mercado interligado.

Esses agrupamentos podem sugerir relações estruturais entre empresas com perfis de risco semelhantes, mas também indicam que os resultados podem ser influenciados pelas características intrínsecas dos dados, refletindo as negociações de mercado e as similaridades qualitativas entre as empresas. Dependendo do período analisado, os resultados podem convergir para as características do mercado, com *clusters* mais dependentes das condições externas, como as forças macroeconômicas. Embora os métodos de *clustering* e redução de dimensionalidade mostrem certa eficácia na identificação de padrões, é evidente que as conclusões podem variar de acordo com a janela temporal analisada, revelando a complexidade das interações no mercado.

5.1.3 Janela de 10 Anos (08/12/2014 - 08/12/2024)

Para a análise do Ibovespa na janela de 10 anos, abrangendo o período de 8 de dezembro de 2014 a 8 de dezembro de 2024, foram identificadas 63 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 12.

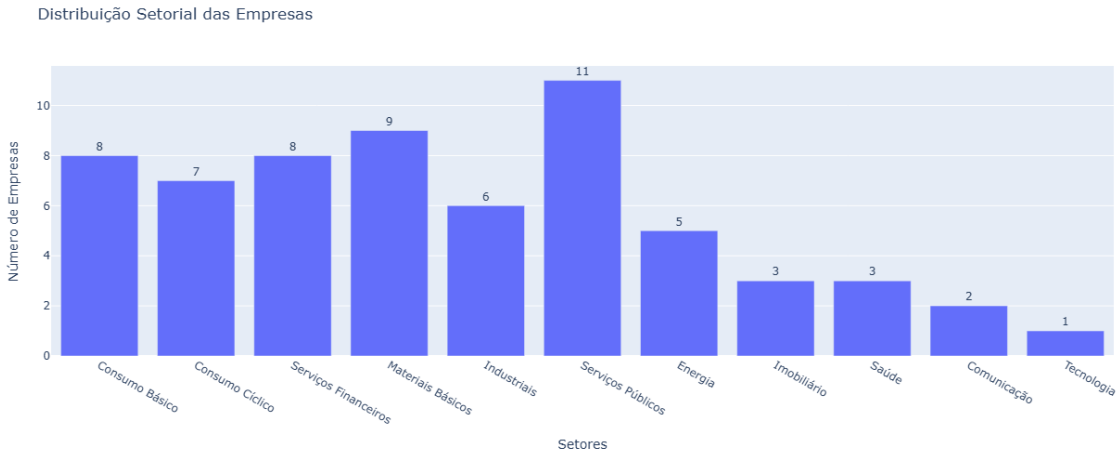

Figura 12 – Distribuição Setorial das Empresas do Ibovespa na Janela de 10 Anos

Price	Adj Close										...	Volume	
Ticker	ABEV3.SA	ALPA4.SA	AZZA3.SA	B3SA3.SA	BBAS3.SA	BBDC3.SA	BBDC4.SA	BBSE3.SA	BEEF3.SA	BRAP4.SA	...	SUZB3.SA	TAEE11.SA
count	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	2488.000000	...	2.488000e+03	2.488000e+03
mean	13.856963	17.198366	47.724200	9.162444	13.485448	12.465054	13.627547	20.240693	8.270028	10.554326	...	4.694467e+06	1.731411e+06
std	1.596331	12.757255	21.051597	4.310627	6.120485	2.917954	3.711730	6.133680	2.136749	7.594453	...	4.538382e+06	1.521362e+06
min	9.292284	4.322746	13.587317	2.021479	3.692131	5.580945	5.107516	10.452435	3.666958	0.513559	...	0.000000e+00	0.000000e+00
25%	12.922523	7.878472	28.028420	4.958416	9.321460	10.746604	11.321959	15.706041	6.784414	3.760508	...	0.000000e+00	1.002550e+06
50%	13.800894	11.267827	46.334520	10.314145	12.299796	12.337315	13.723829	17.857630	8.137910	7.193716	...	4.497000e+06	1.509050e+06
75%	14.673638	23.631275	64.131996	12.434538	16.989188	14.077464	16.139116	23.613381	9.445866	18.266638	...	6.946900e+06	2.135225e+06
max	19.354807	60.498562	97.764854	18.666693	28.316017	20.541697	21.942633	37.419998	15.145242	26.675676	...	5.432910e+07	4.019060e+07

8 rows × 378 columns

Figura 13 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 13, o conjunto de dados é composto por 2.488 registros e 378 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 315 variáveis são do tipo `float64` e 63 são do tipo `int64`, totalizando uma estrutura de aproximadamente 7.2 MB em memória. Além disso, o conjunto resultante não possui dados faltantes ou linhas duplicadas.

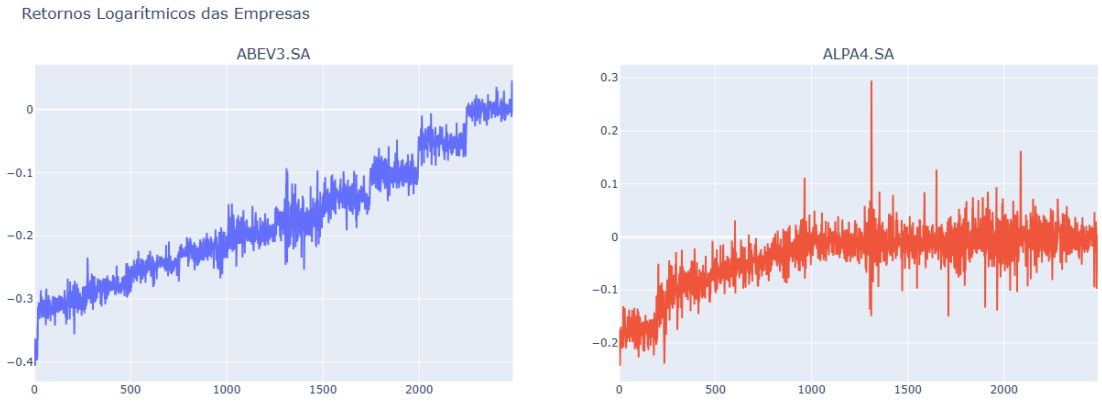

Figura 14 – Série Temporal do Retorno Logaritmo da Ambev e do Grupo Allos na Janela de 10 Anos

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.371	186.214
PCA	HDBSCAN	0.170	16.999
PCA	GMM	0.336	160.808
t-SNE	KMeans	0.432	266.976
t-SNE	HDBSCAN	0.208	44.506
t-SNE	GMM	0.392	217.492
UMAP	KMeans	0.442	494.683
UMAP	HDBSCAN	0.428	75.536
UMAP	GMM	0.407	447.971

Figura 15 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela de 10 Anos

Assim, é possível analisar que o UMAP em conjunto com o KMeans resultou nos melhores resultados, conforme Figura 15.

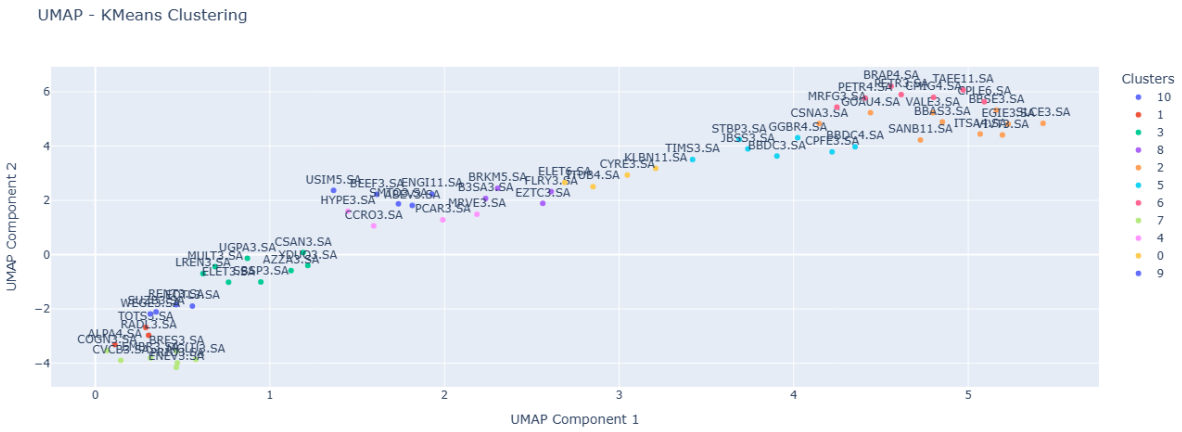

Figura 16 – Clusterização Resultante do UMAP + KMeans

Com o aumento do horizonte temporal para 10 anos, os comportamentos atípicos observados em períodos mais curtos foram, em sua completude suavizados, uma vez que as variações de curto prazo se diluíram ao longo de um intervalo mais amplo de tempo. Essa dinâmica pode ter permitido que as análises dos *clusters* se tornassem mais conclusivas, ao capturar diferentes ciclos de mercado e evidenciar padrões estruturais mais robustos nos dados. A maior estabilidade proporcionada pelo período mais longo favoreceu a identificação de agrupamentos consistentes, refletindo as tendências subjacentes que guiam o comportamento dos ativos ao longo do tempo. Nesse contexto, o UMAP novamente se destacou pela sua capacidade de lidar com grandes volumes de dados, preservando tanto

as estruturas locais quanto globais e facilitando a identificação de padrões latentes. Isso sugere sua eficiência em cenários complexos e de alta dimensionalidade, conforme mostrado na Figura 16.

5.2 S&P 500

5.2.1 Janela Trimestral (08/09/2024 - 08/12/2024)

Para a análise do S&P 500 na janela trimestral, abrangendo o período de 8 de setembro de 2024 a 8 de dezembro de 2024, foram identificadas 500 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 17.

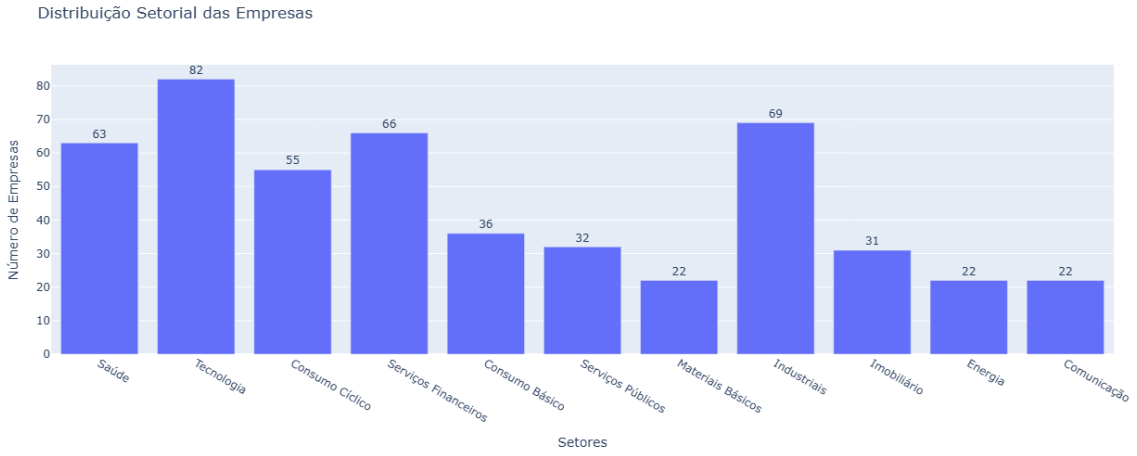

Figura 17 – Distribuição Setorial das Empresas do S&P 500 na Janela Trimestral

Price	Adj Close										...	Volume		
Ticker	A	AAPL	ABBV	ABNB	ABT	ACGL	ACN	ADBE	ADI	ADM	...	WTW	WY	WYNN
count	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	...	6.400000e+01	6.400000e+01	6.400000e+01
mean	137.845044	228.687219	187.929467	132.204219	115.659211	101.853767	355.804681	511.538749	222.917847	55.902134	...	6.514812e+05	3.619803e+06	2.967791e+06
std	5.443151	6.094019	10.034196	6.694133	1.946499	5.063342	10.649825	23.751611	6.661302	2.978294	...	2.331237e+05	1.394126e+06	2.159319e+06
min	125.690002	216.082275	164.990005	115.120003	111.703552	90.335358	333.880951	478.079987	205.479965	50.949524	...	3.260000e+05	1.639000e+06	1.026500e+06
25%	133.855000	225.089996	181.702503	129.700005	114.188072	98.317503	348.424286	496.034996	218.044266	53.102501	...	5.079750e+05	2.593425e+06	1.680850e+06
50%	138.110138	227.984604	191.446190	133.834999	115.950001	101.988625	357.714996	506.514999	224.499428	55.922628	...	5.796500e+05	3.292450e+06	2.237000e+06
75%	140.846127	232.107136	193.410503	136.470005	117.027056	106.650402	362.092506	521.699997	226.924171	58.461186	...	7.881250e+05	4.230300e+06	3.452200e+06
max	148.244003	243.039993	203.869995	147.369995	119.389999	109.220207	376.859985	586.549988	235.433136	61.866573	...	1.293100e+06	7.981300e+06	1.125310e+07

8 rows × 3000 columns

Figura 18 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 18, o conjunto de dados é composto por 64 registros e 3.000 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 2.500 variáveis são do tipo `float64` e 500 variáveis são do tipo `int64`, totalizando aproximadamente 1.5 MB de memória. Além disso, o conjunto de dados não possui valores faltantes ou duplicados, garantindo sua integridade para as análises realizadas.

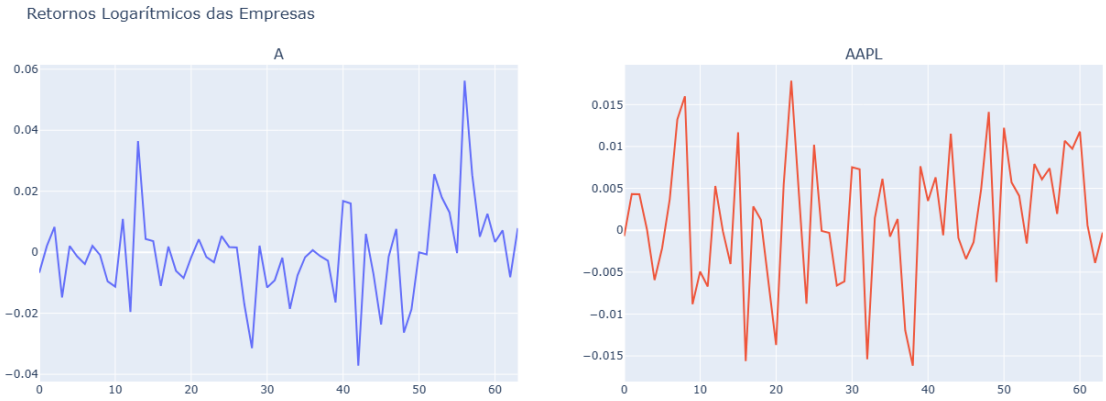

Figura 19 – Série Temporal do Retorno Logaritmo da Agilent e da Apple na Janela Trimestral

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.324	254.804
PCA	HDBSCAN	0.225	19.230
PCA	GMM	0.286	206.055
t-SNE	KMeans	0.350	416.215
t-SNE	HDBSCAN	-0.065	21.610
t-SNE	GMM	0.315	350.523
UMAP	KMeans	0.415	616.306
UMAP	HDBSCAN	0.067	37.625
UMAP	GMM	0.378	515.180

Figura 20 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela Trimestral

Assim, é possível analisar que o UMAP em conjunto com o KMeans resultou nos melhores resultados, conforme Figura 20.

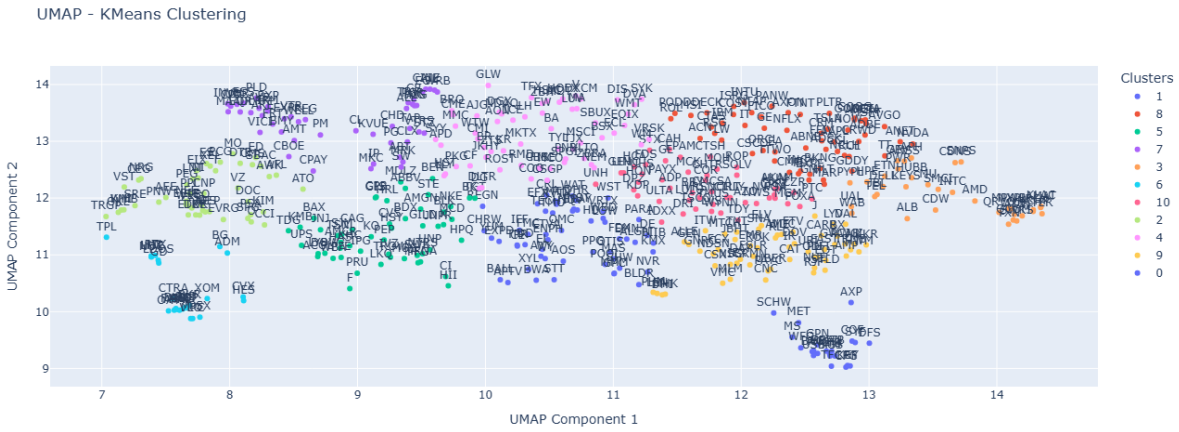

Figura 21 – Clusterização Resultante do UMAP + KMeans

Ao expandir a análise para o S&P 500, as deduções práticas dos agrupamentos se tornaram menos evidentes devido ao maior volume de ações e a complexidade dos dados. Nesse cenário, conforme Figura 21, o UMAP destacou-se como o algoritmo predominante desde a primeira janela analisada, mostrando sua eficiência em lidar com grandes volumes de dados e sua capacidade de preservar estruturas locais e globais.

5.2.2 Janela Anual (08/12/2023 - 08/12/2024)

Para a análise do S&P 500 na janela anual, abrangendo o período de 8 de dezembro de 2023 a 8 de dezembro de 2024, foram identificadas 497 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 22.

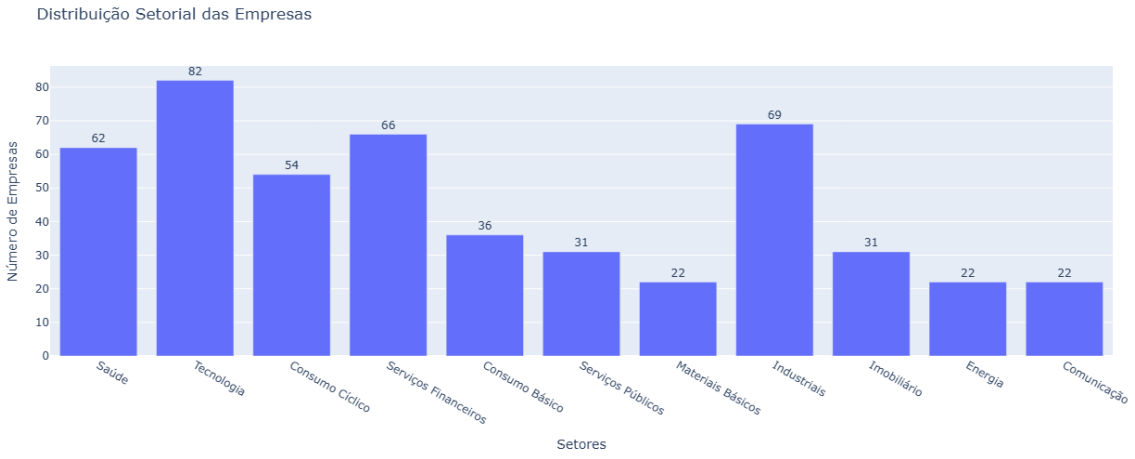

Figura 22 – Distribuição Setorial das Empresas do S&P 500 na Janela Anual

Price	Adj Close										...	Volume		
Ticker	A	AAPL	ABBV	ABNB	ABT	ACGL	ACN	ADBE	ADI	ADM	...	WTW	WY	WYNN
count	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	251.000000	...	2.510000e+02	2.510000e+02	2.510000e+02
mean	137.243974	203.157913	173.191135	142.512052	110.097535	91.407919	336.319109	533.061712	210.043941	58.900856	...	5.030693e+05	3.619955e+06	2.190146e+06
std	6.415439	23.052495	14.684111	13.897119	5.312256	10.378615	26.581426	47.028957	17.712810	5.106858	...	2.208321e+05	1.246065e+06	1.404476e+06
min	125.220650	164.405121	144.015503	113.010002	99.600586	69.814972	279.392487	439.019989	181.151031	49.590546	...	1.944000e+05	1.497800e+06	8.475000e+05
25%	132.089355	183.418915	161.273666	134.595001	105.171192	85.457253	314.192596	493.834991	192.472977	55.362904	...	3.630000e+05	2.808800e+06	1.403750e+06
50%	137.030807	197.144180	170.954834	144.229996	110.427788	93.064438	337.834869	526.880005	213.150833	59.130257	...	4.409000e+05	3.364600e+06	1.832700e+06
75%	140.864838	225.651741	188.316185	151.735001	114.395557	97.791187	360.659988	567.765015	226.376480	60.936056	...	5.770000e+05	4.128050e+06	2.370150e+06
max	154.108200	243.039993	203.869995	168.179993	119.389999	109.220207	382.179138	634.760010	241.354797	73.758446	...	1.476100e+06	1.162730e+07	1.125310e+07

8 rows × 2982 columns

Figura 23 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 23, o conjunto de dados é composto por 64 registros e 3.000 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 2.500 variáveis são do tipo `float64` e 500 variáveis são do tipo `int64`, totalizando aproximadamente 1.5 MB de memória. Além disso, o conjunto de dados não possui valores faltantes ou duplicados, garantindo sua integridade para as análises realizadas.

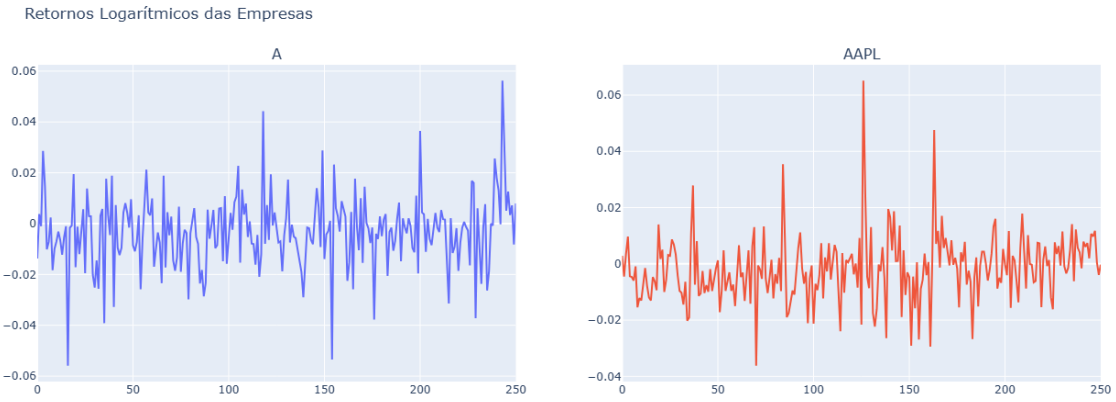

Figura 24 – Série Temporal do Retorno Logaritmo da Agilent e da Apple na Janela Anual

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.351	554.166
PCA	HDBSCAN	0.231	54.344
PCA	GMM	0.294	384.613
t-SNE	KMeans	0.350	662.849
t-SNE	HDBSCAN	-0.104	13.224
t-SNE	GMM	0.307	563.086
UMAP	KMeans	0.396	924.293
UMAP	HDBSCAN	0.115	65.979
UMAP	GMM	0.362	787.605

Figura 25 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela Anual

Assim, é possível analisar que o UMAP em conjunto com o KMeans resultou nos melhores resultados, conforme Figura 25.

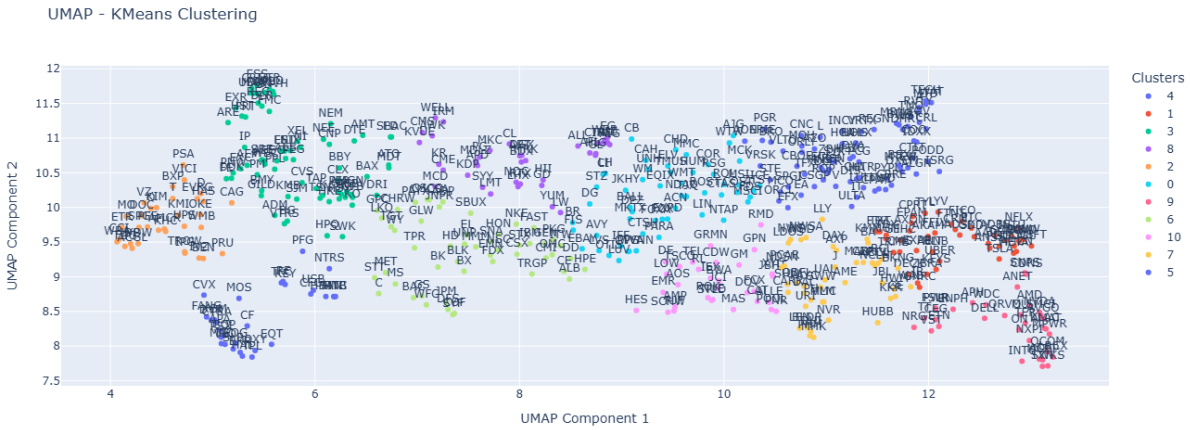

Figura 26 – Clusterização Resultante do UMAP + KMeans

Na análise da janela de um ano para o S&P 500, conforme Figura 26, o UMAP manteve-se como o algoritmo predominante, com a formação de *clusters* se tornando mais evidente. Isso se deve ao maior volume de dados processados, o que contribui para uma representação mais precisa das relações complexas entre os pontos. De forma similar ao Ibovespa, as características intrínsecas dos dados e o *timeframe* selecionado permanecem como variáveis de suma importância, influenciando diretamente o *output* do resultado e a capacidade do modelo em capturar padrões significativos ao longo do tempo.

5.2.3 Janela de 10 Anos (08/12/2014 - 08/12/2024)

Para a análise do S&P 500 na janela de 10 anos, abrangendo o período de 8 de dezembro de 2014 a 8 de dezembro de 2024, foram identificadas 468 ações distribuídas entre 11 setores da economia, conforme mostrado na Figura 27.

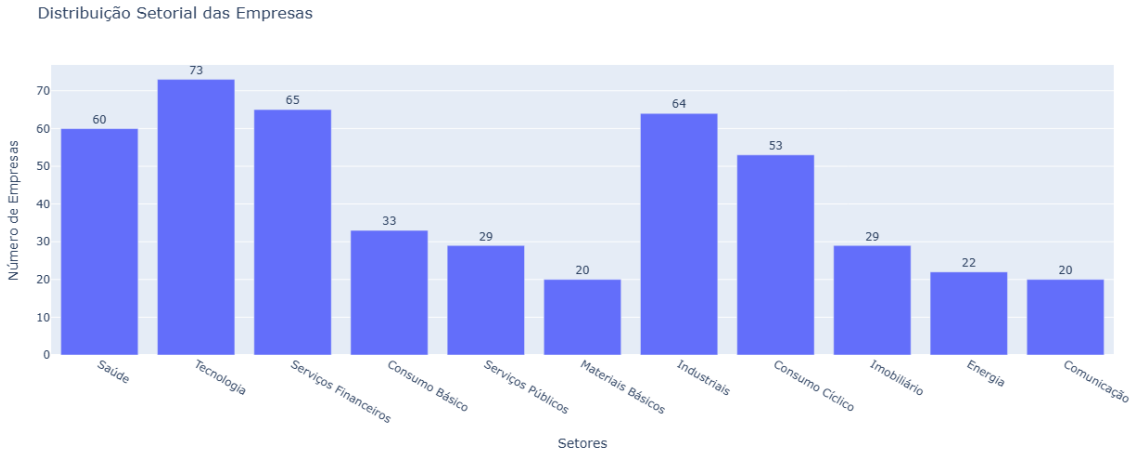

Figura 27 – Distribuição Setorial das Empresas do S&P 500 na Janela de 10 Anos

Price	Adj Close										...	Volume	
Ticker	A	AAPL	ABBV	ABT	ACGL	ACN	ADBE	ADI	ADM	ADP	...	WTW	WY
count	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	...	2.517000e+03	2.517000e+03
mean	88.956313	92.814823	89.056427	76.371300	40.971640	200.616935	317.359805	114.197628	47.940943	149.711215	...	7.370609e+05	4.167648e+06
std	40.294297	64.751266	43.941786	30.540976	21.961279	91.701147	176.882681	53.341517	17.234620	64.210422	...	6.865867e+05	1.913351e+06
min	31.031206	20.697262	32.962009	30.864792	18.235064	69.412506	69.739998	40.739754	23.524809	59.849777	...	0.000000e+00	7.096000e+05
25%	56.039059	34.503033	49.751289	42.359497	26.216272	110.629539	141.119995	69.690857	34.918850	87.796425	...	4.330000e+05	2.933400e+06
50%	76.060036	63.438080	72.998558	77.500435	31.408176	178.413239	299.869995	102.393425	38.974232	141.229050	...	5.926000e+05	3.738800e+06
75%	129.123932	149.497910	130.420013	104.589584	44.254818	285.336853	479.119995	158.009491	60.085068	208.333328	...	8.558000e+05	4.881400e+06
max	175.479553	243.039993	203.869995	133.728119	109.220207	397.036926	688.369995	241.354797	92.135544	306.382477	...	2.204698e+07	2.339580e+07

8 rows × 2808 columns

Figura 28 – Estatística Descritiva do *Dataset* Resultante

Conforme mostrado na Figura 28, o conjunto de dados é composto por 2.517 registros e 2.808 colunas, representando diferentes atributos dos ativos analisados. Entre essas colunas, 2.340 variáveis são do tipo `float64` e 468 variáveis são do tipo `int64`, totalizando aproximadamente 53.9 MB de memória. O conjunto de dados não possui valores faltantes ou duplicados

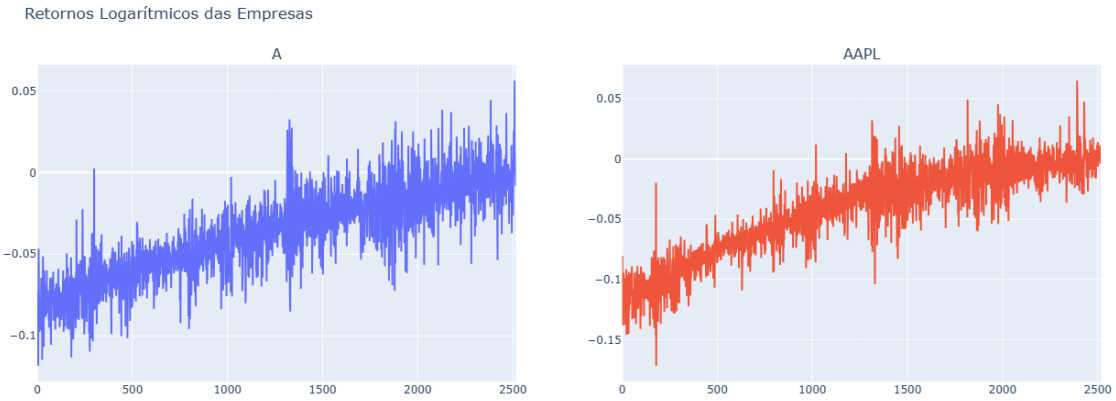

Figura 29 – Série Temporal do Retorno Logaritmo da Agilent e da Apple na Janela de 10 Anos

Ao aplicar os algoritmos de redução de dimensionalidade e os métodos de clusterização, foi possível analisar as métricas de avaliação previamente selecionadas para comparar a qualidade dos agrupamentos e a preservação da estrutura dos dados.

Resultados de Métricas por Técnica de Redução e Agrupamento

Reduction Technique	Clustering Method	Silhouette Score	Calinski-Harabasz Score
PCA	KMeans	0.500	2155.633
PCA	HDBSCAN	0.109	90.959
PCA	GMM	0.397	1050.210
t-SNE	KMeans	0.480	2574.834
t-SNE	HDBSCAN	0.385	472.306
t-SNE	GMM	0.461	2319.132
UMAP	KMeans	0.663	8391.415
UMAP	HDBSCAN	0.555	962.800
UMAP	GMM	0.612	6578.339

Figura 30 – Resultados das Métricas por Técnica de Redução e Agrupamento na Janela de 10 Anos

Assim, é possível analisar que o UMAP em conjunto com o KMeans resultou nos melhores resultados, conforme mostrado na Figura 30.

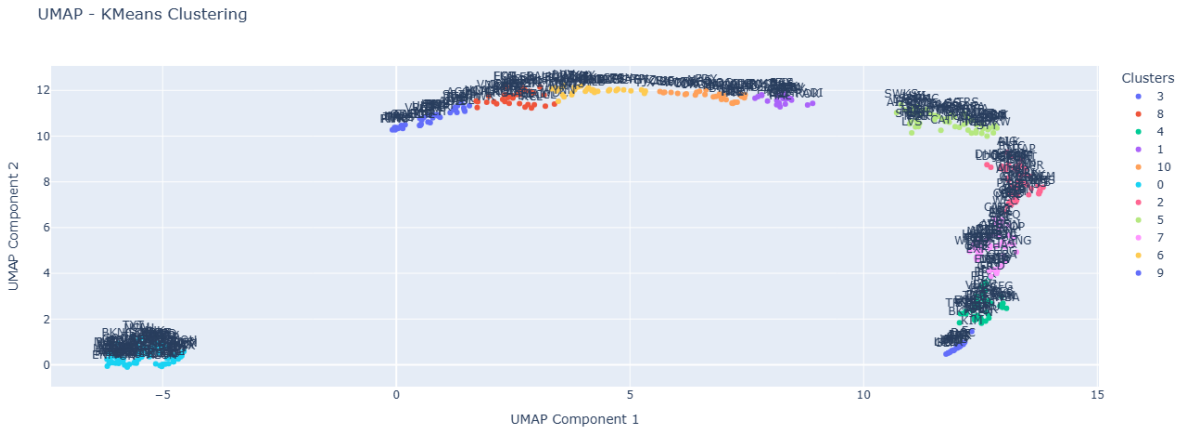

Figura 31 – Clusterização Resultante do UMAP + KMeans

Por fim, na análise da janela de 10 anos para o S&P 500, conforme mostrado na Figura 31, o UMAP manteve-se como o algoritmo predominante, consolidando a sua eficiência em lidar com o grande volume de dados e a identificação de padrões relevantes. Além disso, os *clusters* emergiram de maneira mais clara, dada a dinâmica mais consolidada do mercado americano, que é caracterizado por uma maior estabilidade e maturidade em comparação a outros mercados.

6 Conclusão

Este trabalho analisou a aplicação de técnicas de redução de dimensionalidade e algoritmos de clusterização em diferentes horizontes temporais, com o objetivo de setorizar ativos e identificar padrões nos índices Ibovespa e S&P 500. Ao longo das análises, ficou evidente que o resultado das metodologias aplicadas varia dependendo do período e do índice em questão, refletindo a interação dinâmica entre características intrínsecas dos dados e a natureza dos métodos utilizados.

Para *timeframes* curtos - até janela trimestrais - os resultados indicaram que não há um padrão claro entre os algoritmos. Essa ausência de consistência pode ser atribuída a qualidade limitada das informações disponíveis, uma vez que séries temporais curtas capturam mais ruídos e menos tendências estruturais. Os *clusters* formados diferiram significativamente entre os métodos, destacando que, em cenários de alta volatilidade e menor estabilidade, os algoritmos refletem nuances momentâneas das séries temporais. Nesse contexto, os métodos como PCA mostraram-se úteis para captar variações locais ou globais, respectivamente, mas não apresentaram resultados convergentes em termos de setorização.

Quando analisados horizontes temporais mais longos, como janelas anuais e de 10 anos, padrões mais claros emergiram, e o UMAP destacou-se consistentemente como a técnica mais eficaz. Isso ocorre porque séries temporais longas proporcionam maior estabilidade e capturam tendências de mercado mais estruturadas, reduzindo o impacto de flutuações de curto prazo. O UMAP demonstrou ser particularmente eficaz para identificar padrões não lineares e relações complexas entre ativos. No S&P 500, sua eficácia foi ainda mais evidente, devido ao maior número de ações e a riqueza de informações, que possibilitaram uma análise mais robusta. Em contrapartida, o PCA, embora eficiente para identificar *clusters* amplamente separados em alguns casos, apresentou limitações em cenários com dinâmicas mais intrincadas, enquanto o t-SNE sofreu com baixa escalabilidade e dificuldades em preservar distâncias globais ao longo dos diferentes *timeframes*.

Outro importante aspecto observado foi a ausência de padrões intrínsecos fixos nos índices analisados. Tanto o Ibovespa quanto o S&P 500 são dinâmicos, refletindo o comportamento dos investidores frente a mudanças econômicas, políticas e setoriais. Essa característica mutável ressalta a importância de utilizar técnicas de RD e algoritmos de clusterização como ferramentas para revelar padrões específicos em situações particulares de mercado. Assim, a escolha das metodologias deve ser orientada pelos objetivos da análise e pelas características do conjunto de dados.

Em resumo, este estudo destacou que a eficácia das técnicas de redução de dimensi-

onalidade e dos algoritmos de clusterização dependem do contexto local e temporal em que são aplicados. Enquanto o UMAP emergiu como a técnica mais robusta para setorização em horizontes temporais de média e longa duração, os outros métodos também possuem seus méritos em cenários específicos. Essas descobertas oferecem valiosas contribuições para a análise de dados financeiros, e consolidam uma importante ferramenta de suporte para auxiliar gestores e investidores a tomarem decisões mais informadas sobre a alocação de ativos e a interpretação de dinâmicas de mercado.

Referências

AGGARWAL, C. C. Charu c. aggarwal. *Machine Learning for Text*, v. 7, p. 9, 2018. Citado na página 15.

AGHABOZORGI, S.; Seyed Shirkhorshidi, A.; Ying Wah, T. Time-series clustering – a decade review. *Information Systems*, v. 53, p. 16–38, 2015. ISSN 0306-4379. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306437915000733>. Citado na página 20.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 14.

BISHOP, C. M. Pattern recognition and machine learning. *Springer google schola*, v. 2, p. 1122–1128, 2006. Citado na página 17.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. *Density-based clustering based on hierarchical density estimates*. [S.l.]: Springer-Verlag, 2013. Citado na página 16.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2021. Citado na página 14.

GHOJOGH, B. et al. *Elements of Dimensionality Reduction and Manifold Learning*. [S.l.]: Springer, 2023. Citado na página 20.

GHOJOGH, B. et al. *Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey*. 2021. Disponível em: <https://arxiv.org/abs/2109.02508>. Citado na página 23.

HARDT, M.; RECHT, B. Patterns, predictions, and actions: A story about machine learning. *CoRR*, abs/2102.05242, 2021. Disponível em: <https://arxiv.org/abs/2102.05242>. Citado na página 14.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. USA: Prentice-Hall, Inc., 1988. ISBN 013022278X. Citado na página 18.

JAMES, G. *An introduction to statistical learning*. [S.l.]: springer, 2013. Citado na página 15.

LANDALUCE-CALVO, M. I.; MODROÑO-HERRÁN, J. I. Classification for time series data. an unsupervised approach based on reduction of dimensionality. *Journal of Classification*, Springer, v. 37, n. 2, p. 380–398, 2020. Citado na página 20.

LU, Z. Q. J. The elements of statistical learning: Data mining, inference, and prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*, v. 173, n. 3, p. 693–694, 06 2010. ISSN 0964-1998. Disponível em: <https://doi.org/10.1111/j.1467-985X.2010.00646_6.x>. Citado na página 15.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. 11, 2008. Citado na página 22.

MCINNES, L. et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, v. 2, n. 11, p. 205, 2017. Citado na página 16.

MCINNES, L.; HEALY, J.; MELVILLE, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. Disponível em: <https://arxiv.org/abs/1802.03426>. Citado na página 23.

MURPHY, K. P. *Machine learning: a probabilistic perspective*. [S.l.]: MIT press, 2012. Citado na página 17.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Pearson, 2016. Citado na página 14.

WANG, Y. et al. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, v. 22, n. 201, p. 1–73, 2021. Disponível em: <http://jmlr.org/papers/v22/20-1061.html>. Citado na página 21.