Análise de Agrupamento de Ações com K-Means: Aplicação de Técnicas de Aprendizado Não Supervisionado para Identificação de Padrões no Mercado de Ações: Utilização de Clustering para Investigações de Padrões em Preço da Ação, Quantidade de Cotas e Valor de Mercado

Stock Clustering with K-Means: Application of Unsupervised Learning Techniques for Pattern Identification in the Stock Market: Using Clustering to Investigate Patterns in Stock Price, Share Quantity, and Market Value

Data da versão final: 02 de dezembro de 2024.

RESUMO

Este estudo tem como finalidade explorar a aplicação de técnicas de aprendizado não supervisionado para categorizar ações do mercado financeiro com base em atributos como preço da ação, volume de cotas e capitalização de mercado. A investigação foi conduzida utilizando o algoritmo K-means para identificar grupos, sem a necessidade de rótulos ou categorização prévia das ações. Inicialmente, os dados das ações foram limpos e preparados, transformando variáveis categóricas em numéricas e realizando uma normalização para assegurar a consistência nas análises. O trabalho apresentou a execução de diferentes agrupamentos com 4, 5 e 8 clusters, fundamentando-se em métricas como o índice de silhueta e o gráfico do cotovelo para determinar o número ideal de clusters. Os resultados evidenciaram a eficácia do K-means em segmentar as ações em grupos com características financeiras semelhantes, oferecendo insights sobre o comportamento do mercado.

Palavras-chave: aprendizado não supervisionado; K-means; agrupamento; análise de dados financeiros; mercado de ações.

ABSTRACT

This study aims to explore the application of unsupervised learning techniques to categorize financial market shares based on attributes such as share price, share volume and market capitalization. The investigation was conducted using the K-means algorithm to identify groups, without the need for labels or prior categorization of actions. Initially, the action data was cleaned and prepared, transforming categorical variables into numerical variables and performing normalization to ensure consistency in the analyses. The work presented the execution of different groupings with 4, 5 and 8 clusters, based on metrics such as the silhouette index and the elbow graph to determine the ideal number of clusters. The results demonstrated the effectiveness of K-means in segmenting shares into groups with similar financial characteristics, offering insights into market behavior.

Keywords: unsupervised learning; K-means; grouping; financial data analysis; stock market.

1 INTRODUÇÃO

A aplicação de análise de dados no mercado financeiro é fundamental para identificar padrões e desenvolver decisões estratégicas. Este estudo aborda a utilização do K-Means como uma solução eficaz para agrupar ações de forma automática, superando os desafios relacionados à complexidade e ao volume de informações nesse setor.

Uma das abordagens mais úteis é a segmentação de ações, que permite agrupar empresas com características similares, facilitando análises e orientando estratégias de investimento. Neste trabalho, utilizamos o algoritmo de clustering K-Means para analisar um conjunto de dados financeiros de empresas, com o objetivo de identificar padrões relevantes para investidores. O K-Means é uma técnica não supervisionada amplamente empregada, projetada para agrupar dados semelhantes e reduzir a variabilidade dentro de cada grupo. Essa abordagem pode oferecer insights valiosos sobre o perfil das ações, contribuindo para o desenvolvimento de estratégias de investimento mais eficazes.

Para avaliar o desempenho do algoritmo, aplicamos duas métricas de análise: o Método do Cotovelo e a Análise de Silhueta. Essas técnicas são essenciais para determinar a quantidade ideal de clusters e avaliar a qualidade dos agrupamentos gerados. Além disso, a redução da dimensionalidade dos dados foi realizada utilizando o método Principal Component Analysis (PCA), permitindo a criação de visualizações em 2D e 3D para uma melhor interpretação dos resultados.

2 REVISÃO DE LITERATURA

A escolha do número ideal de clusters é um dos aspectos mais críticos em qualquer análise de clustering. Métodos como o do Cotovelo avaliam a variância explicada pelos clusters formados, enquanto a Análise de Silhueta mede a coesão interna e a separação entre os clusters (Melo, 2019).

Além disso, a visualização dos resultados desempenha um papel importante na interpretação dos agrupamentos. Ferramentas como a Análise de Componentes Principais (PCA) são amplamente utilizadas para reduzir a dimensionalidade dos dados, permitindo a criação de gráficos mais claros e intuitivos, que ajudam na análise dos resultados (Almeida, 2022).

O K-Means é um dos algoritmos mais amplamente utilizados para clustering devido à sua simplicidade e eficiência computacional. No contexto financeiro, sua aplicação é particularmente útil na análise de ações, pois permite identificar padrões ocultos em grandes volumes de dados financeiros (Ferreira, 2021).

Pesquisas recentes demonstram que o K-Means pode ser empregado para agrupar ações com base em indicadores financeiros, como retorno sobre o patrimônio, crescimento da receita, volatilidade, e liquidez. Essa abordagem ajuda investidores a identificar setores ou

empresas com características semelhantes, promovendo uma melhor diversificação de portfólio (Souza, 2020).

3 METODOLOGIA

A metodologia empregada neste estudo abrange a coleta, limpeza e preparação dos dados financeiros das ações. Antes de aplicar o algoritmo de agrupamento, realizamos uma análise exploratória dos dados. A visualização das distribuições das variáveis e a análise estatística básica são essenciais para compreender as características dos dados e identificar possíveis problemas que possam impactar o modelo. Os dados foram normalizados para evitar que variáveis com escalas distintas influenciassem os resultados do agrupamento.

O algoritmo K-means foi implementado com diferentes quantidades de clusters (4, 5 e 8), utilizando como variáveis o preço da ação, a quantidade de cotas e o valor de mercado. Os resultados foram avaliados com base no índice de silhueta e no método do cotovelo. Além disso, gráficos de dispersão em 3D foram utilizados para visualizar os clusters formados, permitindo uma análise mais aprofundada das características financeiras de cada grupo.

Para determinar o número ideal de clusters, empregamos duas técnicas populares:

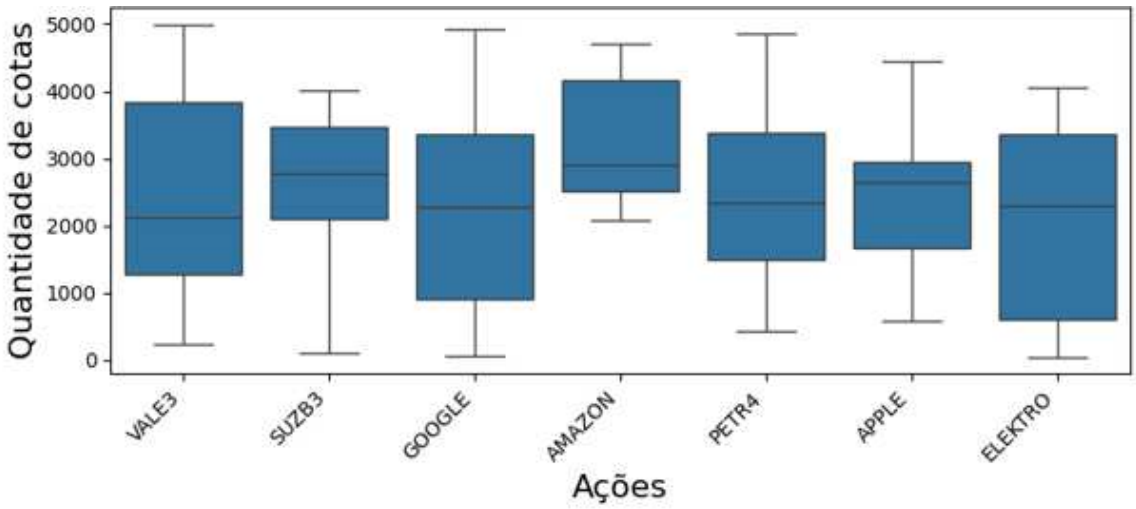
Método do Cotovelo: Este método consiste em calcular a inércia (ou soma das distâncias quadradas das amostras em relação ao centro de cada cluster) para diferentes valores de k (número de clusters). Ao traçar a inércia em função de k, podemos identificar o ponto em que a diminuição da inércia começa a se tornar menos significativa. Esse ponto é considerado o número ideal de clusters.

Análise de Silhueta: Essa análise avalia a qualidade dos clusters formados. Para cada ponto, calcula-se a distância média entre ele e outros pontos dentro do mesmo cluster, assim como a distância média até os pontos do cluster mais próximo. A pontuação de silhueta varia de -1 a 1, onde valores próximos a 1 indicam clusters bem definidos, enquanto valores negativos sugerem que os pontos estão mal agrupados. Utilizamos essa técnica para avaliar a qualidade dos clusters com diferentes valores de k.

4 RESULTADOS E DISCUSSÕES

Além da análise de agrupamento e da visualização em 2D, uma abordagem significativa para investigar a distribuição de dados financeiros é a utilização do boxplot, que oferece uma visão clara sobre a dispersão e a presença de outliers nas variáveis analisadas. No contexto deste estudo, elaboramos um boxplot para ilustrar a distribuição do valor de mercado das ações de diferentes empresas. O valor de mercado foi selecionado como variável de interesse devido à sua importância na avaliação do porte e da posição de mercado das empresas.

Figura 1: Boxplot do Valor de Mercado



Fonte: Elaborado pelo Autor

Neste gráfico, o eixo X representa as diferentes ações (ou empresas), enquanto o eixo Y exibe o valor de mercado de cada uma, expresso em bilhões de reais. O boxplot facilita a compreensão da distribuição do valor de mercado para cada ação, permitindo visualizar as medidas de tendência central (mediana), a dispersão (intervalo interquartil) e os outliers (valores extremos).

Caixa (Box): Representa o intervalo interquartil, que abrange 50% dos dados. A linha central na caixa indica a mediana dos valores.

Bigodes (Whiskers): Indicam a extensão dos dados dentro de 1,5 vezes o intervalo interquartil. Os pontos que estão fora dessa extensão são considerados outliers.

Outliers: Valores extremos que se situam fora dos bigodes e podem indicar empresas com características atípicas em relação ao valor de mercado.

Esse tipo de gráfico é útil para identificar empresas com valores de mercado que se distanciam significativamente dos demais, o que pode sinalizar ativos com alta volatilidade ou comportamentos que fogem da norma do grupo.

Ao visualizar os primeiros registros com a coluna correspondente a 4 clusters, verificamos que a ação VALE3 foi atribuída ao cluster 3, a ação SUZB3 ao cluster 1, e as demais ações ao cluster 0. Esses resultados indicam que as ações foram agrupadas com base em características financeiras similares, refletindo diferenças significativas entre os grupos identificados.

Na coluna correspondente a 5 clusters, observamos que as ações foram distribuídas de forma um pouco mais granular: VALE3 permaneceu no cluster 3, enquanto SUZB3 foi atribuída ao cluster 4. Isso sugere que o aumento no número de clusters permitiu uma separação mais específica entre as ações analisadas, destacando particularidades antes agrupadas.

Com 8 clusters, os resultados foram ainda mais detalhados: VALE3 foi atribuída ao cluster 4, SUZB3 ao cluster 2, com outras ocorrências de SUZB3 nos clusters 0 e 6, e a ação GOOGLE ao cluster 6. Essa segmentação mais refinada indica que o modelo conseguiu identificar padrões mais sutis nas características das ações, separando-as com base em variações adicionais como preço, quantidade de cotas e valor de mercado.

Esses agrupamentos sugerem que o modelo K-means conseguiu capturar diferentes padrões entre as ações analisadas. Por exemplo, VALE3 pode ter sido associada a clusters relacionados a altos valores de mercado e volume de cotas, enquanto SUZB3 e GOOGLE foram agrupadas com base em características específicas que as diferenciam de outras ações. O aumento no número de clusters revelou mais detalhes sobre essas diferenças, mas também destacou a importância de interpretar os clusters dentro do contexto financeiro para garantir que os insights sejam significativos e aplicáveis. Isso demonstra a eficácia do aprendizado não supervisionado na descoberta de padrões ocultos em dados financeiros complexos.

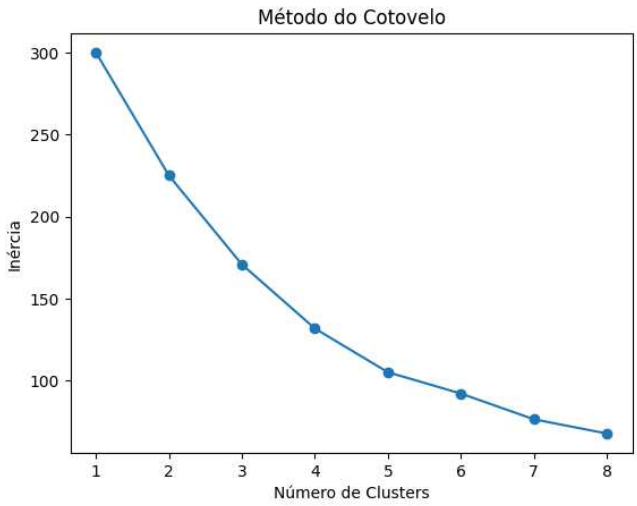
Figura 2: Clusters

	Cluster	Cluster_5	Cluster_8
0	3	3	4
1	1	4	2
2	0	0	6
3	0	0	0
...			
1	1	4	2
2	0	0	6
3	0	0	0
4	0	0	6

Fonte: Elaborado pelo Autor

Plotando o gráfico do cotovelo, identificamos que o valor ideal para utilização de clusters se encontra em 8, no qual a inércia identificada atinge um ponto de inflexão. Esse resultado indica que, a partir de 8 clusters, a redução na inércia começa a ser marginal, tornando esse o número mais apropriado para capturar as variações principais nos dados sem sobrecarregar a análise com informações redundantes

Figura 3: Método do Cotovelo

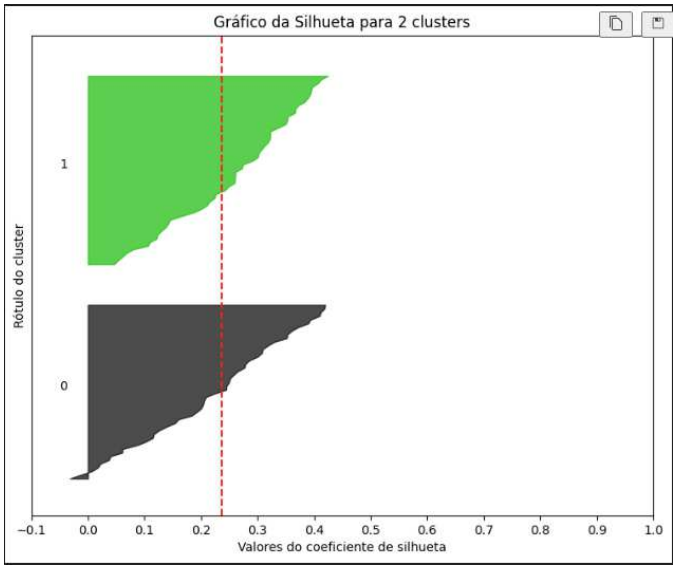


Fonte: Elaborado pelo autor

Para diferentes números de clusters, foram obtidos os seguintes resultados médios do coeficiente de silhueta:

- 2 clusters: 0.237

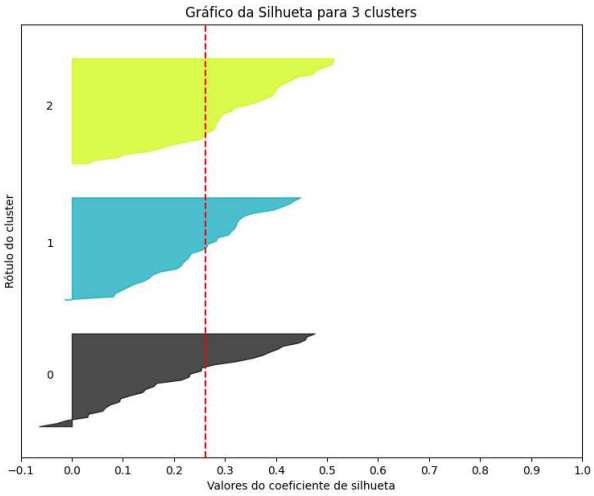
Figura 4: Gráfico Silhueta 2 Clusters



Fonte: Autoria Própria

- 3 clusters: 0.262

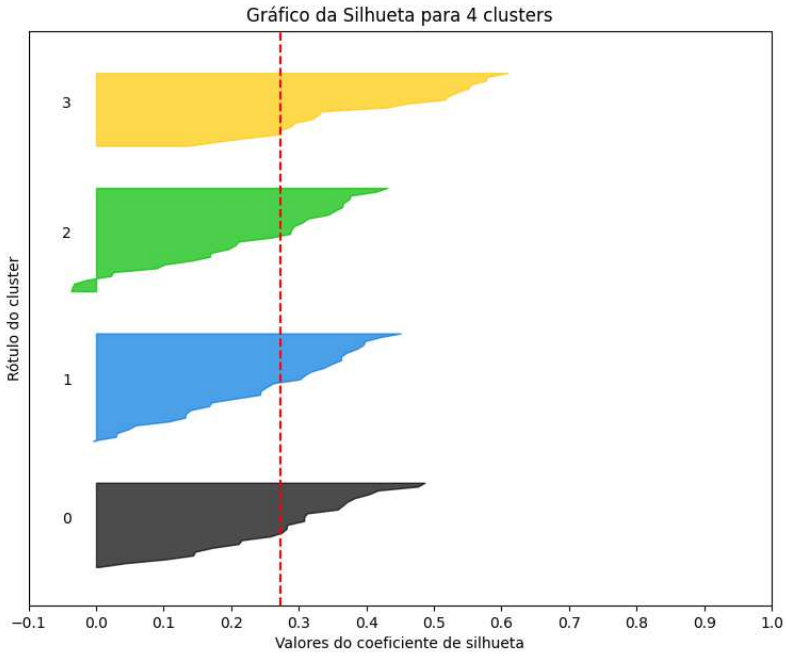
Figura 5: Gráfico Silhueta 3 Clusters



Fonte: Autoria Própria

- 4 clusters: 0.273

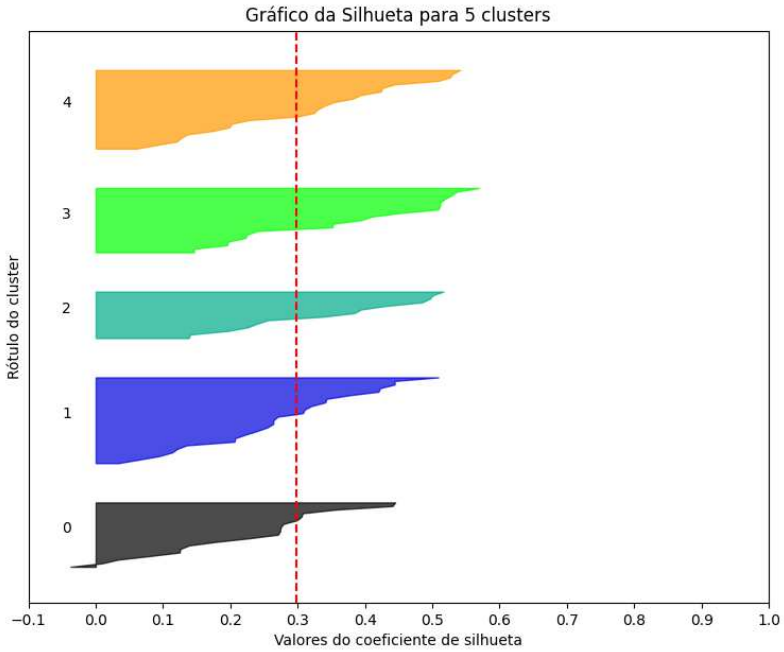
Figura 6: Gráfico Silhueta 4 clusters



Fonte: Autoria Própria

- 5 clusters: 0.298

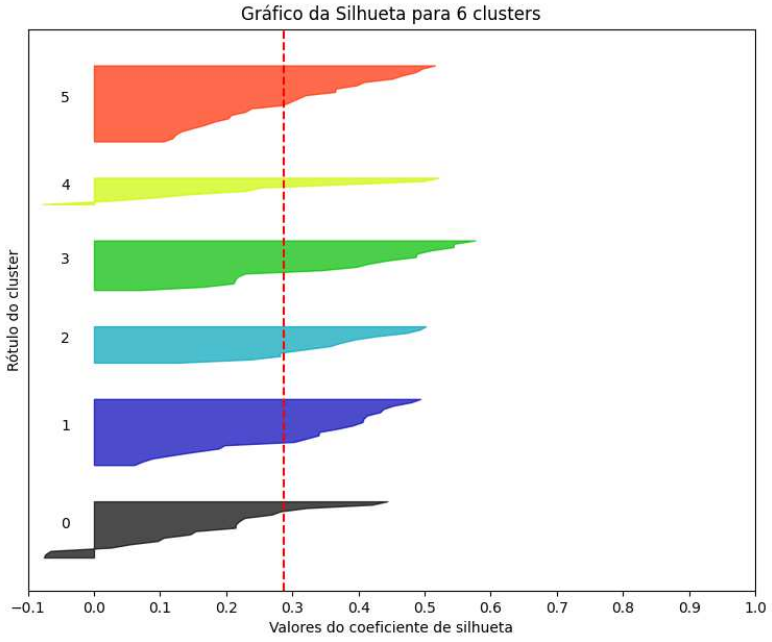
Figura 7: Gráfico Silhueta 5 clusters



Fonte: Autoria Própria

- 6 clusters: 0.286

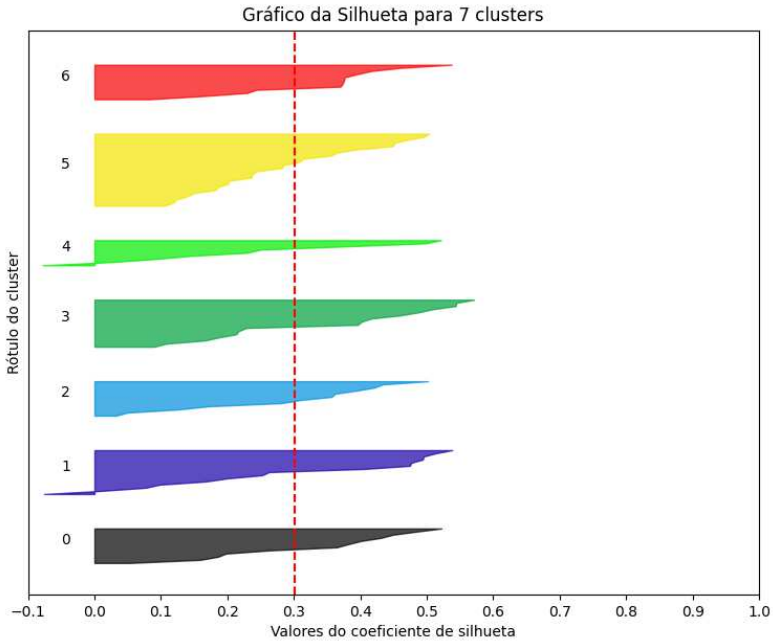
Figura 8: Gráfico Silhueta 6 clusters



Fonte: Autoria Própria

- 7 clusters: 0.301

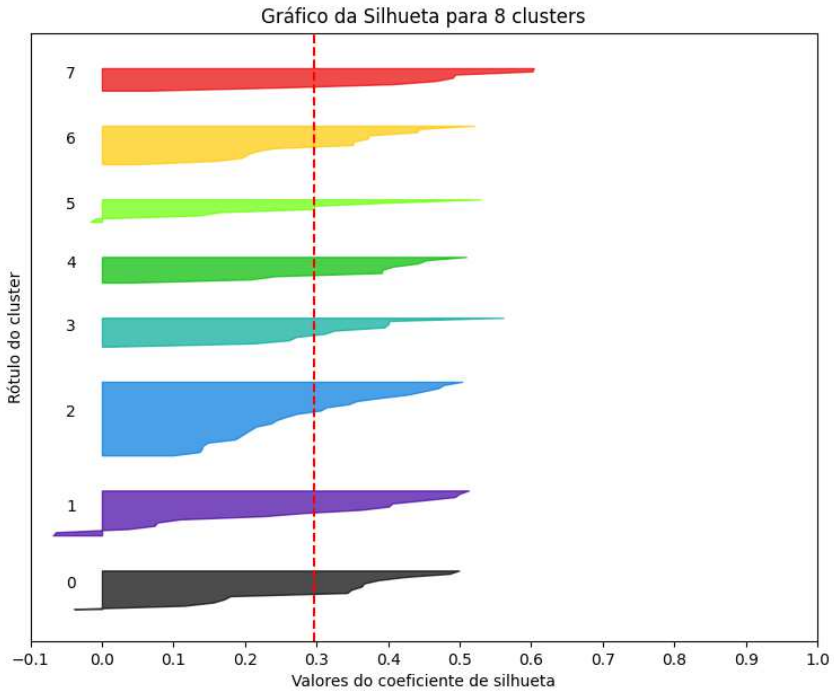
Figura 9: Gráfico Silhueta 7 clusters



Fonte: Autoria Própria

- 8 clusters: 0.296

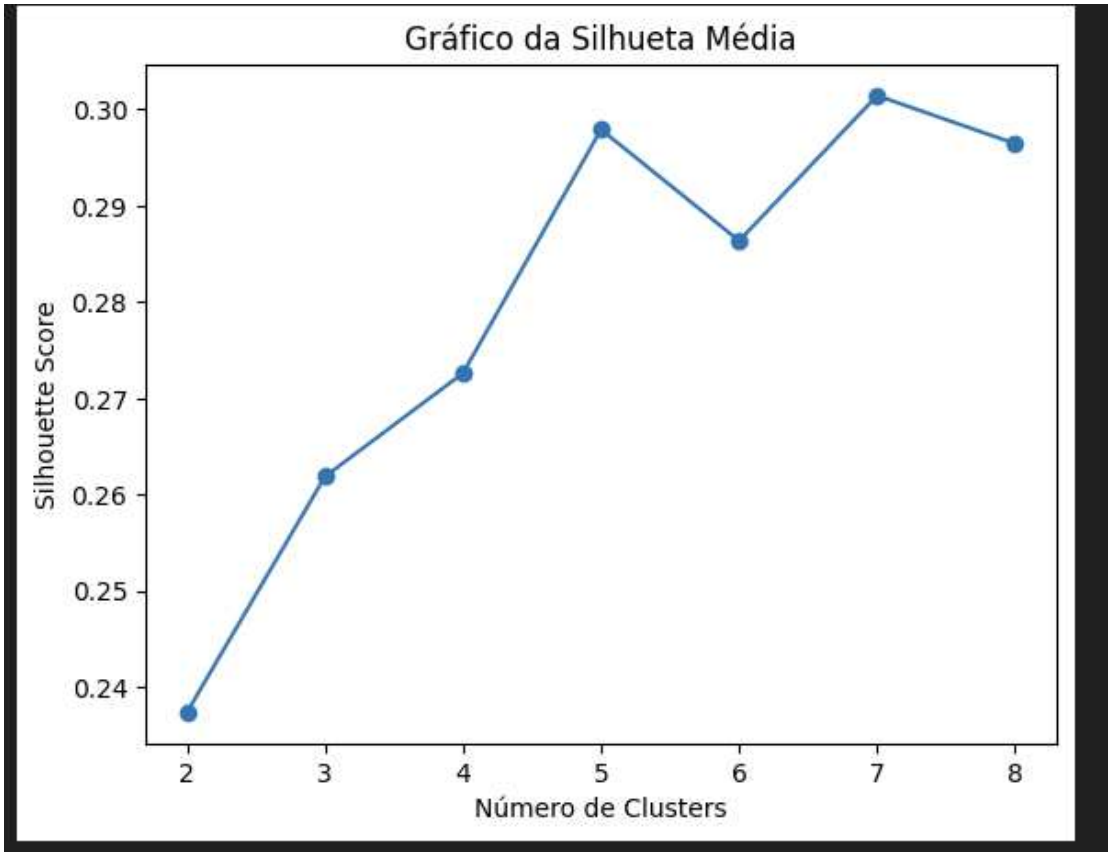
Figura 10: Gráfico Silhueta 8 clusters



Fonte: Autoria Própria

Com base nesses dados, constatamos que o desempenho mais eficaz (com a maior silhueta média) foi obtido com 7 clusters. No entanto, ao examinarmos o gráfico do cotovelo, percebemos que o número ideal de clusters é 8, onde a inércia atinge um ponto de inflexão. Esse achado indica que, a partir de 8 clusters, a diminuição da inércia se torna marginal, tornando essa quantidade a mais adequada para capturar as principais variações nos dados, sem sobrecarregar a análise com informações desnecessárias. A seguir, apresentamos o gráfico da silhueta média.

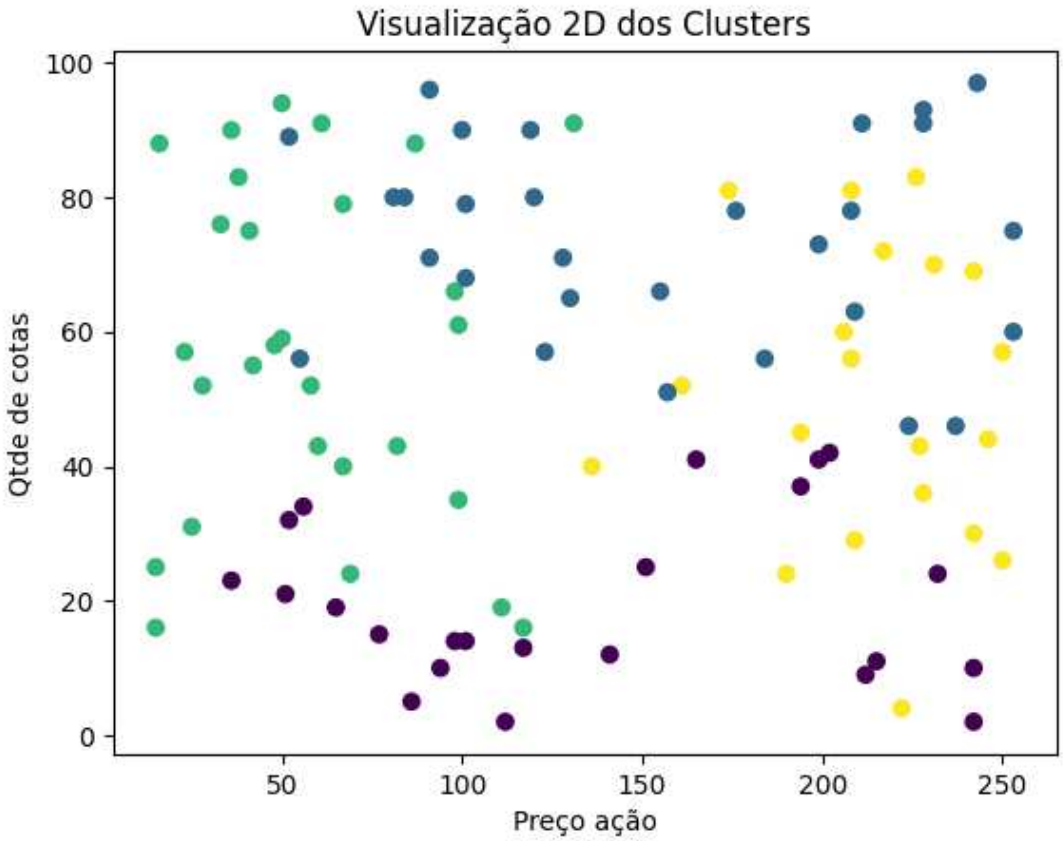
Figura 11: Gráfico Silhueta 8 clusters



Fonte: Autoria Própria

A representação em duas dimensões é uma ferramenta fundamental para analisar graficamente a distribuição dos clusters criados. No gráfico produzido, utilizamos as variáveis "Preço da ação" e "Quantidade de cotas" para ilustrar os clusters, com cores distintas para diferenciar os grupos identificados pelo modelo.

Figura 12: Gráfico Silhueta 8 clusters

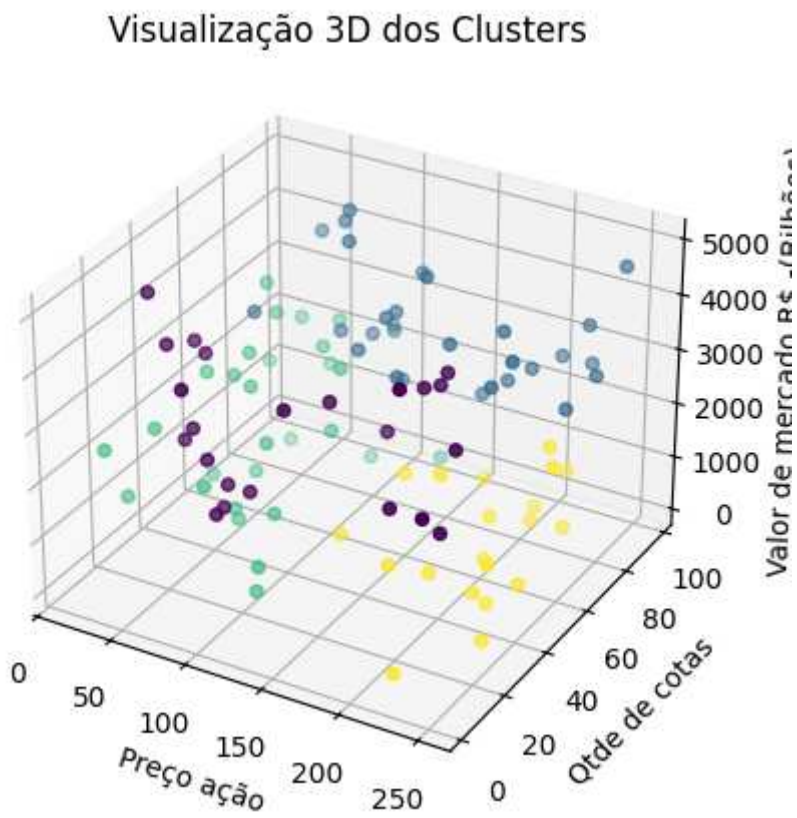


Fonte: Autoria Própria

Essa visualização possibilita analisar como os clusters estão distribuídos no espaço bidimensional e examinar a proximidade entre os pontos de cada grupo. A seleção das variáveis tem um impacto direto na interpretação dos agrupamentos, permitindo a identificação de tendências ou padrões relevantes nos dados. No caso em questão, os clusters evidenciaram diferenças significativas na relação entre o preço das ações e o volume de cotas, refletindo as características financeiras que influenciaram os agrupamentos.

Para uma análise mais completa, foi realizada uma visualização em três dimensões, incorporando uma terceira variável: "Valor de mercado (R$ - Bilhões)". O gráfico 3D é especialmente útil para identificar padrões e clusters que não são perceptíveis em duas dimensões.

Figura 13: Visualização 3D dos Clusters



.

Fonte: Autoria Própria

Neste gráfico, é possível observar a organização dos clusters levando em conta três dimensões ao mesmo tempo. Os diferentes grupos se apresentam de forma bem delimitada, com variações em relação às três variáveis analisadas, evidenciando a eficácia do modelo K-means na identificação de padrões em um espaço tridimensional. Esse tipo de visualização enriquece a análise ao oferecer uma compreensão mais aprofundada sobre as distinções entre os grupos, sendo especialmente relevante em estudos financeiros onde diversas variáveis interagem.

5 CONCLUSÃO

O objetivo deste estudo foi investigar o agrupamento de ações financeiras utilizando o método K-means e analisar como diferentes configurações de clusters poderiam revelar padrões ocultos nos dados. Esse propósito foi alcançado com sucesso, possibilitando a identificação de características específicas que diferenciam ações como VALE3, SUZB3 e GOOGLE, com base em variáveis como preço, quantidade de cotas e valor de mercado.

Os resultados obtidos mostraram que o modelo K-means é eficaz na identificação de padrões em dados financeiros. A análise inicial com 4 clusters proporcionou uma visão geral dos agrupamentos, enquanto a configuração com 8 clusters revelou nuances mais sutis nas diferenças entre as ações. O valor médio do coeficiente de silhueta validou a qualidade

satisfatória dos agrupamentos, especialmente na configuração com um maior número de clusters.

As visualizações em 2D permitiram observar as relações entre variáveis como preço e quantidade de cotas, enquanto a visualização em 3D adicionou uma dimensão extra à análise, destacando o valor de mercado como uma variável essencial na formação dos clusters. Essas representações gráficas foram cruciais para a interpretação dos resultados, evidenciando a versatilidade e o potencial do aprendizado não supervisionado.

REFERÊNCIAS

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.

MURPHY, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge: The MIT Press, 2012.

RUSSELL, Stuart; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 3. ed. Upper Saddle River: Pearson, 2010.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645-678, 2005.

BISHOP, Christopher M. Pattern recognition and machine learning. *Journal of Machine Learning Research*, v. 8, n. 1, p. 1123-1128, 2007.