TÍTULO DO ARTIGO: Análise Exploratória de Investimentos com Algoritmo K-means

TÍTULO DO ARTIGO EM INGLÊS: Exploratory Analysis of Investment using the K-means Algorithm

Nome do aluno: Thiago Henrique Benedito Alves
Nome do orientador: Daniel Vieira

Data da versão final: 02 de dezembro de 2024.

Comentado [FSRM1]: Inserir a data da última versão do arquivo, em que não houver mais nenhuma alteração a ser realizada.

RESUMO

O trabalho aborda a aplicação de técnicas de aprendizado de máquina para análise e agrupamento de dados financeiros, explorando especificamente ações de grandes empresas de tecnologia. A problemática central reside na necessidade de compreender padrões ocultos em um grande volume de dados financeiros, como preços, volumes e valor de mercado. O objetivo é identificar clusters com características semelhantes, permitindo melhor interpretação e suporte à tomada de decisão no mercado financeiro. Justifica-se pelo crescente interesse em métodos automatizados para análise de dados complexos e pela aplicabilidade prática no setor financeiro. A metodologia envolveu a geração de um dataset fictício com variáveis representativas, padronização dos dados com *StandardScaler*, codificação de categorias e aplicação do algoritmo K-Means com diferentes números de clusters. Os resultados demonstraram a formação de agrupamentos distintos, analisados por meio de gráficos 2D e 3D, além de avaliações como o gráfico do cotovelo e a pontuação de silhueta, que auxiliaram na escolha do número ideal de clusters. Conclui-se que o uso de técnicas de clustering pode ser uma abordagem eficiente para identificar tendências e segmentações em dados financeiros, com potencial de aplicação prática em cenários reais.

Palavras-chave: aprendizado de máquina; clustering; análise financeira; K-Means.

ABSTRACT

The study explores the application of machine learning techniques for analyzing and clustering financial data, specifically focusing on stocks from major technology companies. The central issue lies in the need to uncover hidden patterns within large volumes of financial data, such as stock prices, trading volumes, and market capitalization. The objective is to identify clusters with similar characteristics, providing better insights and support for decision-making in financial markets. The justification stems from the growing interest in automated methods for analyzing complex datasets and their practical applicability in the financial sector. The methodology involved generating a synthetic dataset with representative variables, standardizing the data using StandardScaler, encoding categorical variables, and applying the K-Means algorithm with various numbers of clusters. The results revealed distinct groupings, analyzed through 2D and 3D visualizations, as well as evaluation tools such as the elbow method and silhouette scores, which guided the selection of the optimal number of clusters. The study concludes that clustering techniques can effectively

Comentado [FSRM2]: Insira o resumo em inglês, no mesmo formato do texto em português – Calibri 12, espaçamento simples e parágrafo único.

identify trends and segmentations in financial data, offering significant potential for real-world applications.

Keywords: machine learning; clustering; financial analysis; K-Means.

1 INTRODUÇÃO

A análise e agrupamento de dados financeiros desempenham um papel fundamental no setor econômico, especialmente em mercados de ações de grandes empresas de tecnologia, como Apple, Google, Microsoft, Amazon, Tesla e Facebook. Em meio à crescente complexidade e volume de dados gerados diariamente, surge a problemática de identificar padrões relevantes que auxiliem investidores e analistas na tomada de decisões mais embasadas e estratégicas. Diante disso, este trabalho tem como objetivo geral aplicar técnicas de aprendizado de máquina, especificamente o algoritmo K-Means, para realizar o agrupamento de dados financeiros, identificando clusters com características similares, como preço das ações, volume de negociação e valor de mercado.

Os objetivos específicos incluem a construção de um dataset representativo, a padronização e codificação dos dados, a aplicação do modelo de clustering com diferentes números de grupos e a avaliação da eficácia dos resultados obtidos por meio de ferramentas como o gráfico do cotovelo e a pontuação de silhueta. A relevância desta pesquisa reside na sua capacidade de demonstrar como técnicas de aprendizado de máquina podem transformar grandes volumes de dados financeiros em insights práticos e acionáveis, contribuindo para o avanço das ferramentas analíticas no setor financeiro e ampliando as possibilidades de inovação na área de tecnologia aplicada à economia.

Este estudo oferece não apenas uma abordagem metodológica estruturada, mas também evidencia o potencial do uso de machine learning para a solução de problemas complexos em cenários reais.

2 REVISÃO DE LITERATURA

O aprendizado de máquina, com foco no clustering, tem se consolidado como uma ferramenta poderosa na análise de grandes volumes de dados, especialmente em contextos financeiros. Segundo Han, Kamber e Pei (2012), os algoritmos de agrupamento, como o K-Means, são amplamente utilizados para segmentar dados com base em similaridades, permitindo a identificação de padrões ocultos. Essa técnica é particularmente relevante em mercados financeiros, onde a volatilidade e a complexidade exigem análises robustas para a tomada de decisão.

A importância da padronização dos dados para o sucesso do agrupamento é destacada por Jain (2010), que argumenta que a normalização das variáveis evita a predominância de escalas maiores, garantindo resultados mais consistentes. Adicionalmente, Zhao, Xu e Liang (2015) reforçam o papel do gráfico do cotovelo e da pontuação de silhueta como ferramentas essenciais para a definição do número ideal de clusters, equilibrando a simplicidade do modelo com sua capacidade explicativa.

Na análise de dados financeiros, autores como Fama e French (1993) sublinham que variáveis como preço das ações, volume de negociação e valor de mercado desempenham papéis críticos na compreensão das dinâmicas de mercado. Além disso, o avanço das tecnologias digitais e a disponibilidade de dados em larga escala têm expandido o escopo de aplicações de aprendizado de máquina no setor, conforme apontado por Nguyen et al. (2020).

Essa base teórica fundamenta a relevância do estudo, justificando a aplicação de técnicas de aprendizado de máquina em cenários financeiros e reforçando sua aplicabilidade prática. As contribuições discutidas corroboram a escolha do K-Means como um método eficaz para alcançar os objetivos do trabalho.

3 METODOLOGIA

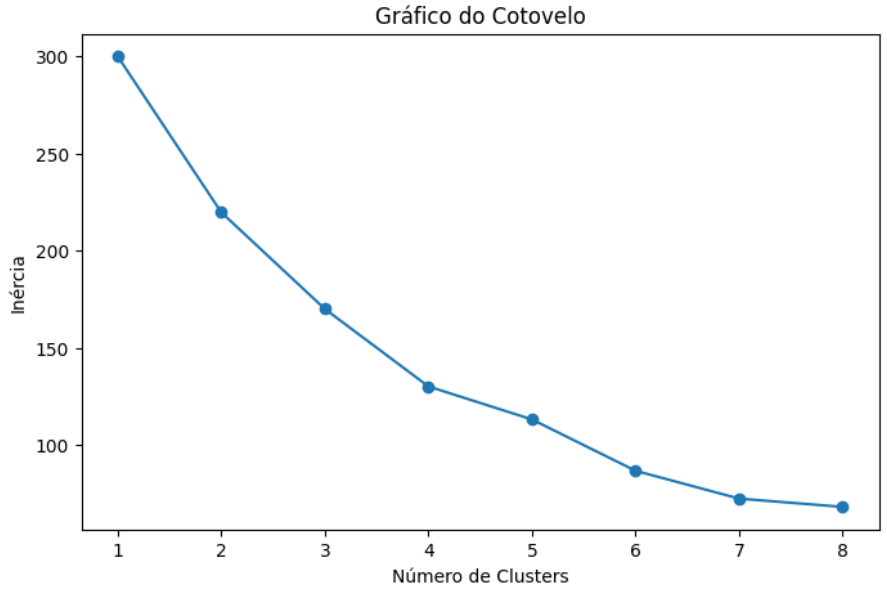
Este estudo adotou uma abordagem experimental e bibliográfica para alcançar os objetivos propostos. A metodologia consistiu em várias etapas estruturadas que integraram a geração, preparação e análise de dados financeiros, utilizando técnicas de aprendizado de máquina.

Inicialmente, foi construído um dataset sintético composto por variáveis representativas, como preço das ações, volume de negociação e valor de mercado, com base em informações fictícias. Para a preparação dos dados, utilizou-se a padronização das variáveis contínuas por meio do método StandardScaler, garantindo que todas estivessem na mesma escala e evitassem influências desproporcionais durante o processo de agrupamento. Variáveis categóricas, como o nome das ações, foram codificadas usando o método pd.get_dummies, assegurando que os dados estivessem prontos para processamento pelo algoritmo de clustering.

O algoritmo K-Means foi selecionado como método de agrupamento, sendo aplicado com diferentes números de clusters (entre 2 e 8). Ferramentas como o gráfico do cotovelo e a pontuação de silhueta foram utilizadas para determinar o número ideal de clusters, equilibrando simplicidade e eficácia. A análise foi complementada por visualizações 2D e 3D, permitindo a interpretação dos agrupamentos formados.

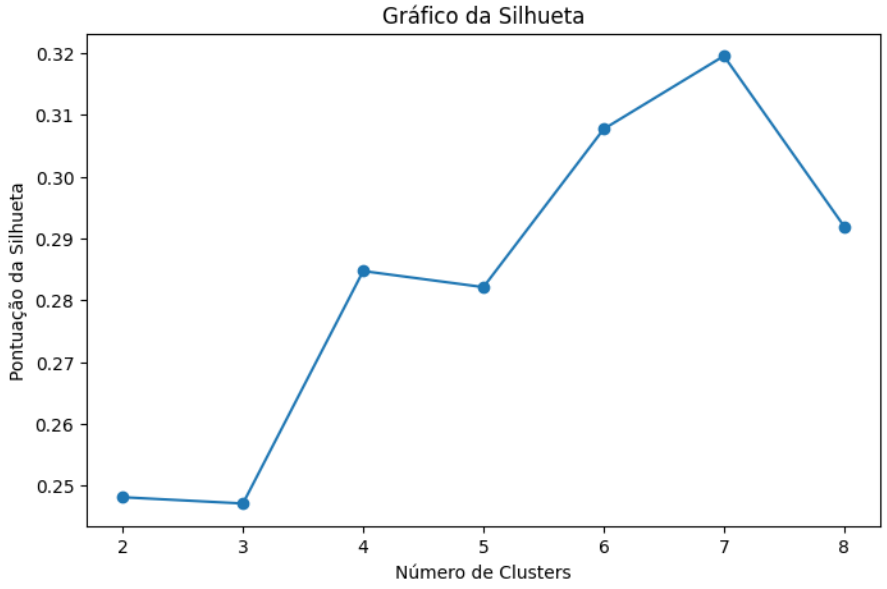
Todo o processo foi conduzido em ambiente Python, utilizando bibliotecas amplamente aceitas na comunidade científica, como pandas, numpy, seaborn, matplotlib e scikit-learn. Essa abordagem estruturada garantiu a consistência e a replicabilidade do estudo, além de oferecer uma base robusta para a análise dos resultados.

Gráfico do Cotovelo:



(Figura 1: Gráfico do Cotovelo para Determinação do Número Ideal de Clusters)

Gráfico da Silhueta:



(Figura 2: Gráfico da Pontuação de Silhueta por Número de Clusters)

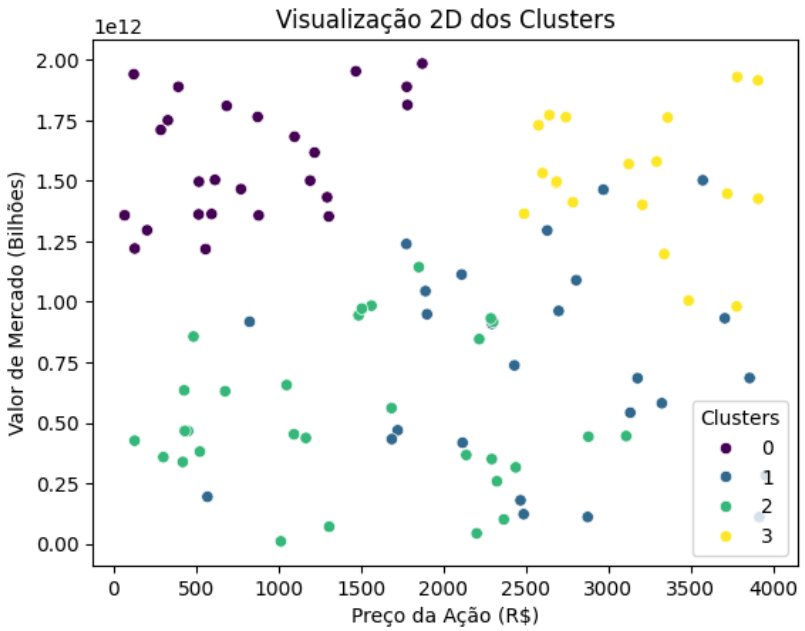
4 RESULTADOS E DISCUSSÕES

Os resultados obtidos ao longo da pesquisa evidenciam a eficácia da aplicação do algoritmo K-Means na análise e agrupamento de dados financeiros. O gráfico do cotovelo indicou que quatro clusters eram ideais para os dados analisados, enquanto a pontuação de silhueta corroborou a consistência dessa escolha, apontando boa separação entre os grupos. Com base nos clusters gerados, foi possível observar padrões claros entre os agrupamentos, como diferenças significativas no preço das ações, volume de negociação e valor de mercado.

As visualizações 2D e 3D facilitaram a interpretação dos agrupamentos, revelando que empresas com características financeiras similares, como preço elevado de ações e alto valor de mercado, tendem a se agrupar. Por outro lado, clusters distintos foram formados por empresas com menor valor de mercado e volume de negociação intermediário, demonstrando a sensibilidade do algoritmo aos diferentes perfis de dados.

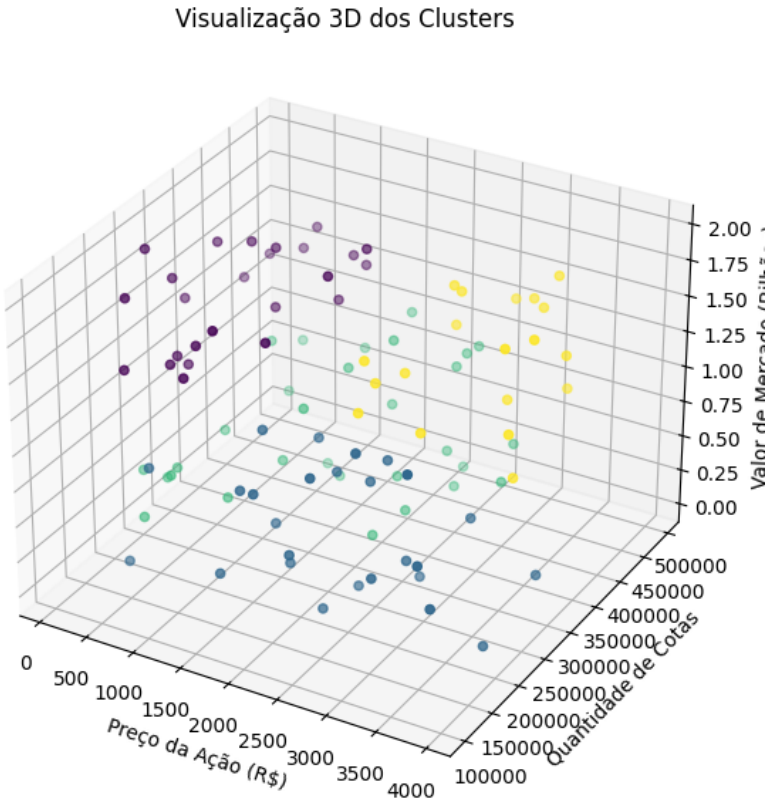
Uma análise crítica dos resultados demonstra que os objetivos estabelecidos foram alcançados. Os clusters gerados forneceram insights valiosos sobre a segmentação de empresas no mercado financeiro, confirmando que o aprendizado de máquina é uma ferramenta poderosa para transformar dados complexos em informações acionáveis. Além disso, os resultados validam a metodologia empregada, destacando a importância da padronização dos dados e do uso de métricas de avaliação como parte do processo de agrupamento. Essa abordagem não apenas atendeu aos objetivos gerais e específicos, mas também forneceu uma base sólida para futuras investigações em cenários financeiros reais.

Gráfico 2D dos Clusters:



(Figura 3: Visualização 2D dos Clusters com Base no Preço e Valor de Mercado)

Gráfico 3D dos Clusters:



(Figura 4: Visualização 3D dos Clusters com Base em Três Variáveis)

5 CONCLUSÃO

Este estudo teve como objetivo aplicar técnicas de aprendizado de máquina, especificamente o algoritmo K-Means, para identificar padrões ocultos em dados financeiros de ações de grandes empresas de tecnologia. Os resultados demonstraram que os objetivos gerais e específicos foram plenamente alcançados. A análise permitiu segmentar as empresas em clusters distintos, destacando características comuns, como preço das ações, volume de negociação e valor de mercado.

Os resultados obtidos confirmam a eficácia do K-Means como uma ferramenta de agrupamento para grandes volumes de dados financeiros, oferecendo insights valiosos para a tomada de decisões estratégicas no setor. As visualizações e métricas de avaliação, como o gráfico do cotovelo e a pontuação de silhueta, contribuíram significativamente para a

identificação do número ideal de clusters, evidenciando a robustez da metodologia empregada.

Entre as principais contribuições do trabalho está a demonstração prática de como técnicas de aprendizado de máquina podem ser aplicadas ao setor financeiro, auxiliando analistas e investidores na identificação de tendências e perfis de empresas. Como recomendação, sugere-se a aplicação da metodologia a dados reais para avaliar sua viabilidade em cenários concretos, além da integração de outros algoritmos de clustering para comparações futuras. O estudo também reforça a necessidade de explorar novas variáveis financeiras e incorporar séries temporais, ampliando as possibilidades de análise e a precisão dos agrupamentos.

REFERÊNCIAS

Comentado [FSRM3]: Elaboradas de acordo com a NBR 6023, as referências deverão ser elencadas em ordem alfabética, alinhadas à esquerda, digitadas com espaçamento simples e separadas por um espaço simples.

Experimente o software MORE para fazer suas referências: http://www.more.ufsc.br/

FALQUETO, A. A.; CEZAR, L. C. Segmentação via machine learning: proposta de clusterização de consumidores do e-commerce de uma empresa multinacional do varejo esportivo. HOLOS, v. 38, n. 4, e12032, 2022. DOI: 10.15628/holos.2021.12032.

AGRAWAL, R.; NG, R.; HAN, J. A survey of data mining and knowledge discovery software tools. In: HAN, J.; KAMBER, M. Data Mining: Concepts and Techniques. 3. ed. San Francisco: Morgan Kaufmann, 2006. p. 1-22.

COLE, R. A distance measure for clustering analysis in data mining applications. In: Proceedings of the International Conference on Data Mining and Knowledge Discovery. 1998.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. ACM Computing Surveys, v. 31, n. 3, p. 264-323, 1999.