NOME: Murilo Silva dos Santos
NOME: Bruno Ricardo Vieira

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: 07 de dezembro de 2024.

RESUMO

O mercado de ações é caracterizado por dados dinâmicos e de alta complexidade, o que torna desafiadora a identificação de padrões e tendências. Este estudo explora a aplicação do algoritmo de agrupamento K-Means para descobrir padrões no comportamento dos ativos financeiros, utilizando como variáveis principais o preço das ações, o volume negociado e o valor de mercado. Por meio do aprendizado não supervisionado, busca-se agrupar ações com características similares, facilitando uma análise mais detalhada das relações e interdependências entre os ativos.

A metodologia consistiu na aplicação do K-Means a um conjunto de dados históricos de ações, definindo um número ideal de *k* clusters para segmentar os ativos de maneira significativa. Os resultados demonstraram que o K-Means foi eficiente na identificação de agrupamentos consistentes, destacando padrões que oferecem informações valiosas para a tomada de decisões estratégicas e elaboração de planos de investimento no mercado financeiro.

Conclui-se que o uso de técnicas de agrupamento pode ser uma ferramenta robusta para investidores e analistas financeiros, permitindo a extração de insights relevantes a partir de grandes volumes de dados e facilitando a identificação de tendências de mercado sem a necessidade de intervenção manual. Além disso, a metodologia apresentada reforça o potencial do aprendizado de máquina no suporte à análise de dados financeiros e à criação de estratégias de investimento mais informadas.

Palavras-chave: aprendizado não supervisionado; K-Means; mercado de capitais; clustering; análise de dados financeiros; inteligência artificial.

ABSTRACT

The stock market is characterized by dynamic and highly complex data, making the identification of patterns and trends a challenging task. This study explores the application of the K-Means clustering algorithm to uncover patterns in the behavior of financial assets, using key variables such as stock prices, trading volume, and market capitalization. Through unsupervised learning, the goal is to group stocks with similar characteristics, enabling a more detailed analysis of the relationships and interdependencies among assets.

The methodology involved applying the K-Means algorithm to a dataset of historical stock data, defining an optimal number of *k* clusters to segment the assets meaningfully. The results demonstrated that K-Means was effective in identifying consistent groupings, highlighting patterns that provide valuable insights for strategic decision-making and investment planning in the financial market.

It is concluded that the use of clustering techniques can be a robust tool for investors and financial analysts, enabling the extraction of relevant insights from large volumes of data and facilitating the identification of market trends without the need for manual intervention. Furthermore, the presented methodology underscores the potential of machine learning in supporting financial data analysis and developing more informed investment strategies.

Keywords: unsupervised learning; K-Means; capital markets; clustering; financial data analysis; artificial intelligence.

1 INTRODUÇÃO

O mercado de ações é um dos pilares fundamentais da economia global, caracterizado por sua natureza altamente dinâmica e volátil, o que torna a análise e a previsão de seu comportamento um desafio constante. Com o crescimento exponencial do volume de dados financeiros e o avanço de tecnologias inovadoras, diversas técnicas analíticas têm surgido para apoiar a tomada de decisões estratégicas. Entre essas, o aprendizado de máquina, especialmente o aprendizado não supervisionado, destaca-se por sua capacidade de identificar padrões e organizar dados de forma eficiente, sem a necessidade de rótulos pré-definidos.

Nesse cenário, o uso de algoritmos de agrupamento, como o K-Means, apresenta-se como uma ferramenta promissora para identificar grupos de ações com características e comportamentos semelhantes. Essa abordagem pode gerar insights valiosos para investidores e analistas financeiros, permitindo uma compreensão mais profunda das inter-relações entre ativos.

Além disso, o mercado de ações, com seu vasto volume de informações — incluindo preços, volumes de negociação e valores de mercado —, beneficia-se significativamente de técnicas de clustering. Essas ferramentas possibilitam uma análise mais eficiente e detalhada, revelando padrões que frequentemente escapam às metodologias tradicionais, promovendo decisões mais embasadas e estratégias otimizadas.

2 REVISÃO DE LITERATURA

O aprendizado de máquina, especialmente suas técnicas não supervisionadas, tem sido amplamente estudado na literatura por sua capacidade de analisar dados sem a necessidade de rótulos ou supervisão direta. Entre essas técnicas, o algoritmo K-Means destaca-se como uma das ferramentas mais utilizadas no agrupamento de dados (clustering). Introduzido por MacQueen (1967), o K-Means organiza os dados em clusters com base em similaridades, minimizando as distâncias entre os pontos de cada cluster e seu centróide.

Na análise de mercados financeiros, essa abordagem permite identificar padrões e agrupamentos de ações ou ativos com características semelhantes. Estudos como os de Bishop (2006) e Aggarwal (2015) enfatizam o papel do aprendizado não supervisionado na descoberta de relações complexas em grandes volumes de dados, tornando-o particularmente útil em áreas como segmentação de clientes, detecção de anomalias e análise de séries temporais.

O pré-processamento dos dados é outro ponto crítico destacado na literatura. Técnicas como normalização, tratamento de valores ausentes e remoção de outliers são fundamentais para garantir a qualidade e a precisão dos resultados dos algoritmos de aprendizado. Além disso, pesquisas recentes, como as de Han et al. (2011) e Lin et al. (2018), mostram que a aplicação de algoritmos de clustering em dados financeiros pode oferecer insights valiosos, como agrupamentos de ações com comportamentos semelhantes ou identificação de tendências emergentes.

Portanto, a literatura confirma que o aprendizado não supervisionado, aliado a uma boa preparação dos dados, oferece um potencial significativo para a análise e exploração de padrões ocultos em dados financeiros.

3 METODOLOGIA

A pesquisa utilizou um método exploratório e quantitativo, fundamentado em uma investigação documental, utilizando informações financeiras de ações. O estudo foi realizado por meio do uso do algoritmo de agrupamento K-Means para detectar padrões no comportamento das ações, agrupando-as em agrupamentos com base em atributos como o preço da ação, número de ações e valor de mercado das companhias. Seguem os passos realizados na pesquisa:

3.1 - Preparação dos Dados

A etapa inicial consistiu na importação dos dados, que abrangiam informações detalhadas sobre as ações, incluindo:

- **Título da ação**: nome ou identificação do ativo.
- **Valor da ação**: preço unitário de aquisição do ativo.
- **Quantidade de cotas**: número total de cotas emitidas pela empresa.
- **Valor de mercado**: avaliação comercial total da empresa.

Esses dados foram coletados de bases financeiras públicas e organizados em um *DataFrame* utilizando a biblioteca Pandas, permitindo o manuseio e a análise estruturada das informações de forma eficiente.

3.2 - Exploração dos Dados

Na fase de exploração, os dados foram analisados para compreender sua distribuição e identificar possíveis outliers:

Foram gerados **box plots** para examinar a distribuição do preço das ações e do valor de mercado das empresas. Para isso, utilizou-se a biblioteca **Seaborn** (com o método sns.boxplot).

O método dados.info() foi empregado para obter uma visão geral do **DataFrame**, permitindo verificar os tipos de dados, a presença de valores ausentes e outras características importantes do conjunto de dados.

Além disso, o método dados.describe() foi utilizado para calcular estatísticas descritivas, como média, desvio padrão, valores mínimo e máximo, entre outros, fornecendo uma análise detalhada de cada variável no conjunto de dados.

3.3 - Pré-processamento dos Dados

O pré-processamento dos dados foi realizado com o objetivo de garantir sua qualidade antes de aplicar o algoritmo K-Means.

Primeiramente, os valores ausentes foram tratados de maneira adequada, utilizando técnicas apropriadas para preenchê-los ou removê-los, conforme necessário, garantindo que os dados estivessem completos e prontos para análise.

Além disso, a coluna de variáveis categóricas (quando presente) foi codificada por meio da função pd.get_dummies(), que converteu as variáveis em formato binário (True ou False). A opção drop_first=True foi utilizada para evitar a multicolinearidade, eliminando a primeira categoria de cada variável e assegurando que não houvesse redundância nas informações. Dessa forma, os dados foram transformados em um formato adequado para o processamento pelo algoritmo de clustering.

3.4 - Aplicação do Algoritmo K-Means

Após a preparação adequada dos dados, o algoritmo K-Means foi empregado para agrupar as ações em agrupamentos com base nas variáveis escolhidas.

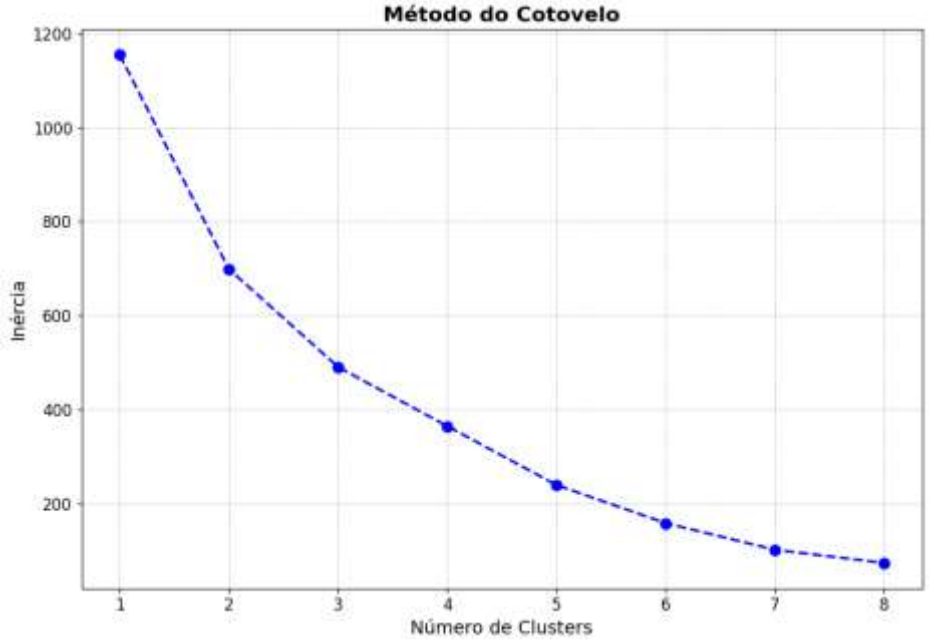
Primeiramente, estabeleceu-se n_clusters = 4 para dividir as ações em 4 grupos, com a finalidade de identificar padrões e segmentações dentro desses grupos.

Depois, ajustamos o número de clusters para 5 e 8, para examinar como as alterações no número de clusters afetam a configuração dos grupos. Esta estratégia possibilitou uma análise mais detalhada da estrutura dos dados e a detecção de potenciais aprimoramentos na segmentação.

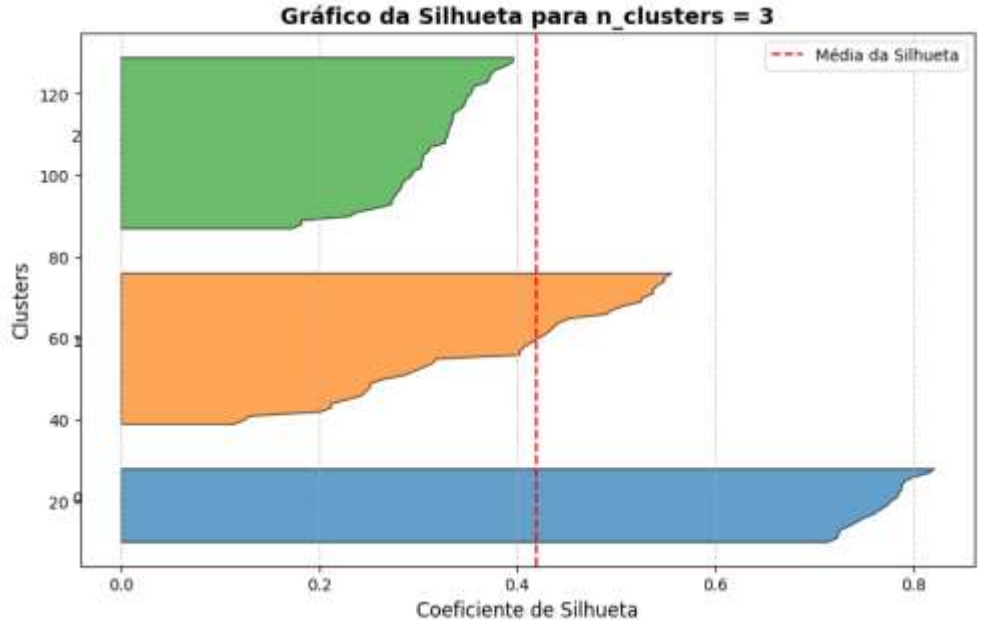
3.5 - Análise dos Clusters

Para avaliar a qualidade e a adequação dos clusters, foram gerados os seguintes gráficos:

- **Gráfico do Cotovelo**: utilizado para determinar o número ideal de clusters, variando **n** de 1 a 8, e observando a curva da soma dos erros quadráticos (inertia).



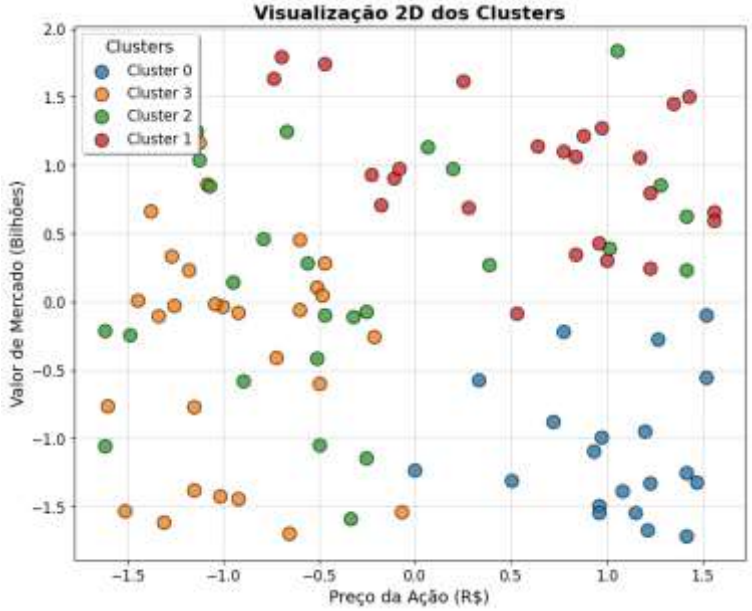
- **Gráfico da Silhueta**: para avaliar o quão bem os dados foram agrupados, medindo a coesão interna e a separação entre os clusters.



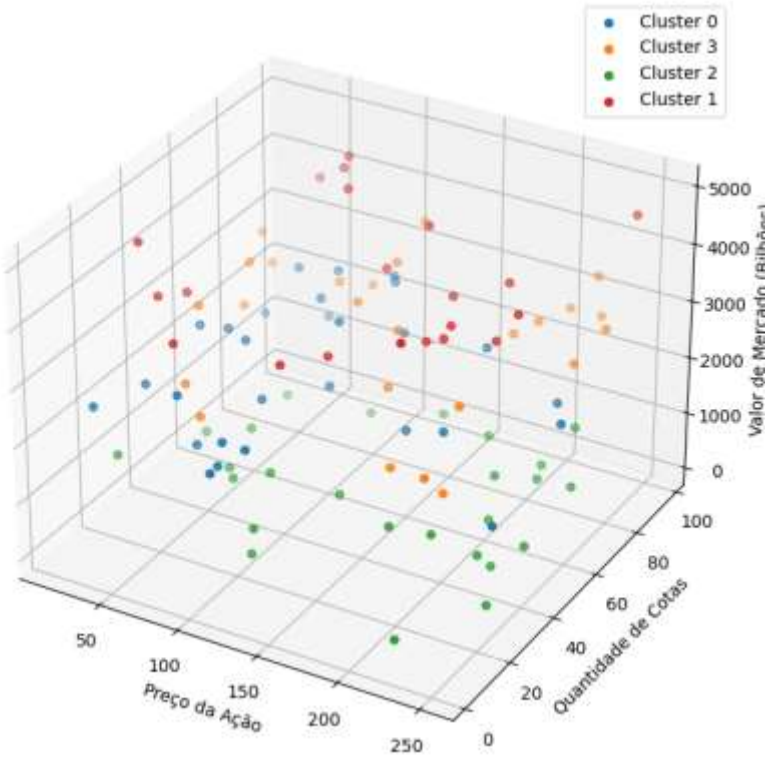
3.6 - Visualização dos Clusters Formados

Por fim, as visualizações dos clusters foram realizadas para facilitar a interpretação dos resultados:

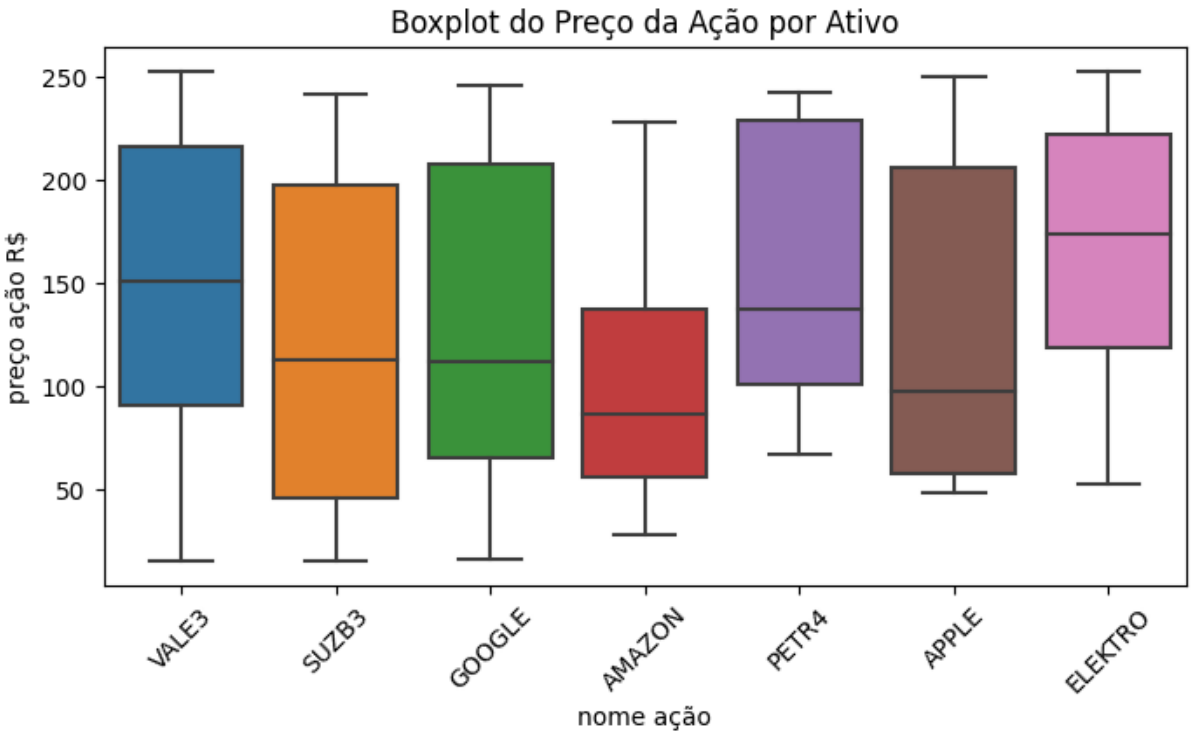
- **Visualização 2D**: foi gerado um gráfico bidimensional para representar a distribuição dos clusters com base nas duas variáveis mais relevantes.

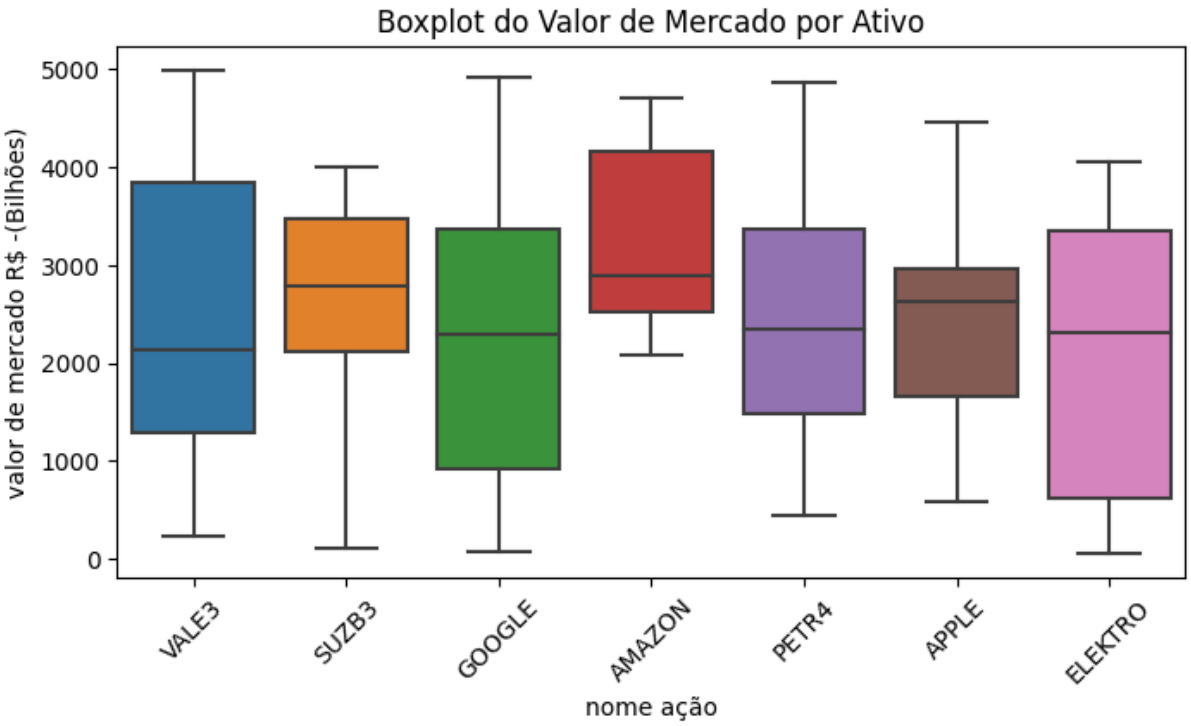


- **Visualização 3D**: um gráfico tridimensional foi criado para representar os clusters com base nas três variáveis (preço da ação, quantidade de cotas e valor de mercado), utilizando a biblioteca **Matplotlib**.



4 RESULTADOS E DISCUSSÕES





dados.info() para análise das informações do conjunto de dados

```
1  dados.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   nome ação                       100 non-null    object
 1   preço ação R$                   100 non-null    int64
 2   qtde cotas                      100 non-null    int64
 3   valor de mercado R$ -(Bilhões)  100 non-null    int64
dtypes: int64(3), object(1)
memory usage: 3.3+ KB
```

dados.describe() para descrição das informações do conjunto de dados

```
1  dados.describe()
```

✓ 0.0s

	preço ação R$	qtde cotas	valor de mercado R$ -(Bilhões)
count	100.000000	100.000000	100.00000
mean	136.140000	52.010000	2433.70000
std	75.237942	27.475791	1397.29373
min	15.000000	2.000000	52.00000
25%	67.000000	28.250000	1189.25000
50%	121.500000	55.500000	2433.00000
75%	209.000000	76.500000	3616.50000
max	253.000000	97.000000	4993.00000

Lidando com valores ausentes, codificando variáveis categóricas etc.

```
1  dados['preço ação R$'].fillna(dados['preço ação R$'].mean(), inplace=True)
2  dados['qtde cotas'].fillna(dados['qtde cotas'].mean(), inplace=True)
3  dados['valor de mercado R$ -(Bilhões)'].fillna(dados['valor de mercado R$ -(Bilhões)'].mean(), inplace=True)
4
5  dados.isnull().sum()
6  # Os valores ausentes nas colunas 'preço ação R$', 'qtde cotas' e 'valor de mercado R$ -(Bilhões)' são preenchidos com a média de cada coluna
```

✓ 0.0s

```
nome ação                       0
preço ação R$                   0
qtde cotas                      0
valor de mercado R$ -(Bilhões)  0
dtype: int64
```

5 CONCLUSÃO

O principal benefício do aprendizado não orientado em relação ao supervisionado é a sua capacidade de detectar padrões e estruturas nos dados sem a exigência de rótulos ou supervisão direta. Embora o aprendizado supervisionado necessite de dados etiquetados para treinar o modelo, o aprendizado não supervisionado pode utilizar dados não etiquetados, tornando-o mais flexível em circunstâncias onde a obtenção de rótulos é dispendiosa, complexa ou inviável.

Ademais, é eficaz na identificação de padrões escondidos ou agrupamentos nos dados, como em métodos de agrupamento (clustering) e diminuição de dimensionalidade, possibilitando uma análise mais aprofundada e a descoberta de informações surpreendentes. Esta estratégia é particularmente eficaz ao lidar com grandes quantidades de dados não categorizados, como textos, imagens ou dados numéricos.

REFERÊNCIAS

- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Lin, T., Jegelka, S., & Sra, S. (2018). "Cluster analysis in high-dimensional data spaces". *Journal of Machine Learning Research*, 19(1), 1-36.
- MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Xu, R., & Wunsch, D. (2005). "Survey of Clustering Algorithms". *IEEE Transactions on Neural Networks*, 16(3), 645-678.