Fundação Educacional Serra dos Órgãos

Disciplina: Data Science

Alunos:

Gabriel Maciel - Matricula: 06006665

Kleber Daniel - Matricula: 06007199

Hugo Norte - Matricula: 06006259

Professor: Mauro Sérgio

Relatório de Aprendizado Não Supervisionado: Análise de Clientes de Shopping

Teresópolis, novembro de 2025

1. Introdução

O presente trabalho tem como objetivo aplicar técnicas de aprendizado não supervisionado utilizando o dataset **“Mall Customers”**.
A análise concentra-se nas variáveis **“Age” (idade)** e **“Spending Score (1-100)” (pontuação de gasto)**, permitindo identificar padrões de comportamento e segmentar clientes em grupos com perfis semelhantes.

O foco é comparar o desempenho dos algoritmos **K-Means** e **Hierarchical Clustering** em termos de qualidade de agrupamento, interpretando seus resultados por meio de métricas de avaliação e validação cruzada.

2. Etapas do workflow

O pipeline de análise seguiu as seguintes etapas:

1. **Leitura e preparação dos dados:** extração das colunas *Age* e *Spending Score* do dataset original.
2. **Pré-processamento:** Na coluna “Gender” foi usado o One-Hot Encoding para manter todos os dados de forma numérica no dataset, e também os dados foram padronizados com StandardScaler para evitar que diferenças de escala afetassem os resultados.
3. **Modelagem:** aplicação dos algoritmos K-Means e Hierarchical Clustering.
4. **Validação cruzada:** uso de **K-Fold (k=5)** para estimar o erro de generalização e avaliar a estabilidade dos resultados.
5. **Avaliação:** cálculo das métricas **Silhouette Score**, **Davies-Bouldin Index** e **Calinski-Harabasz Index**.
6. **Visualização:** geração de gráficos de dispersão e dendrogramas para interpretar a formação dos clusters.

3. Resultados Obtidos

3.1. K-Means

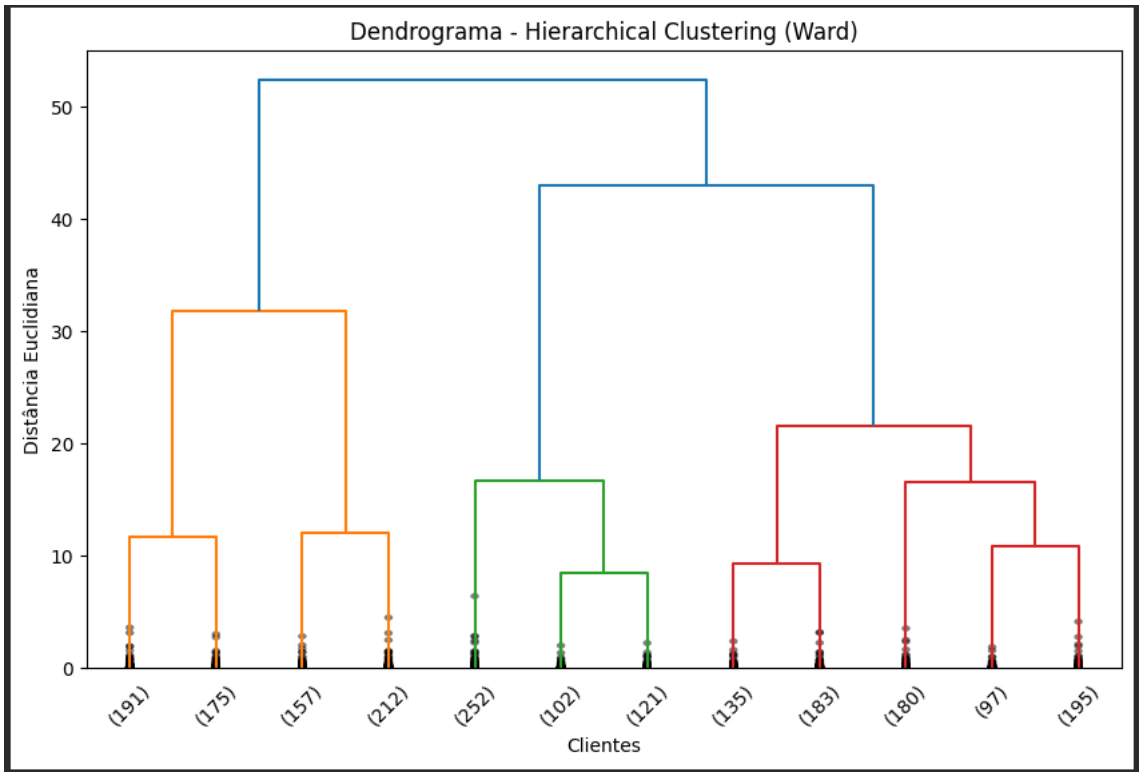
- Silhouette Score: **0.4049**
- Davies-Bouldin Index: **0.7758**
- Calinski-Harabasz Index: **1978.9935**

3.2. Hierarchical Clustering (HC)

- Silhouette Score: **0.3678**
- Davies-Bouldin Index: **0.8247**
- Calinski-Harabasz Index: **1679.3512**

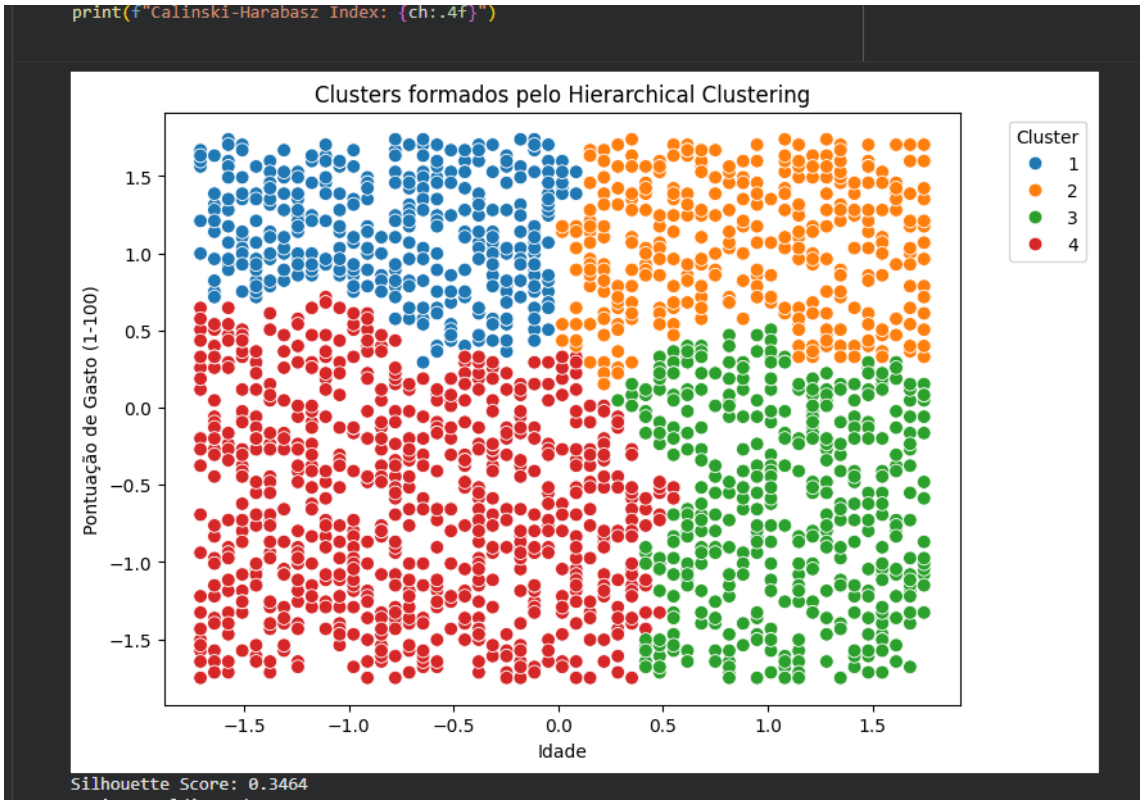
Ambos os modelos cumpriram o objetivo proposto de identificar perfis distintos de clientes.

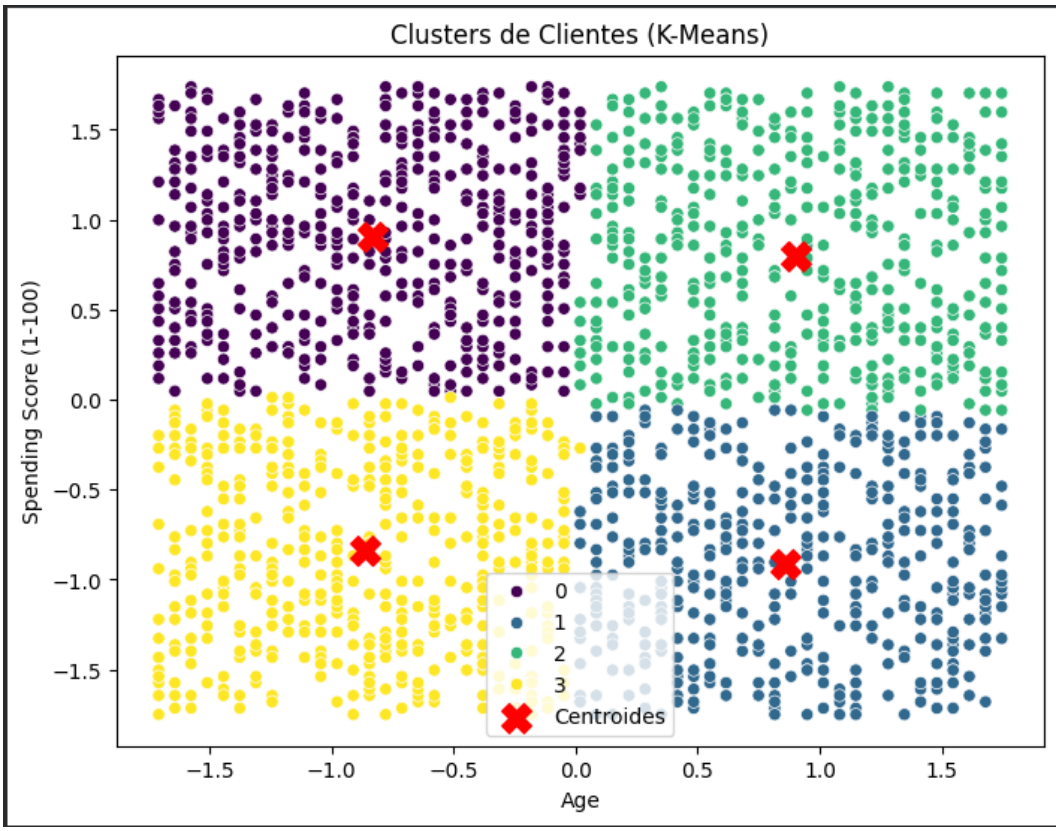
O K-Means mostrou melhor desempenho quantitativo, enquanto o Hierarchical Clustering se destacou pela interpretabilidade e estruturação visual.



O dendrograma mostra quatro agrupamentos bem separados, reforçando que o corte foi adequado.

As distâncias maiores entre os grupos indicam diferenças significativas no comportamento de consumo, especialmente entre grupos de idades e padrões de gasto diferentes.





4. Discussão e Comparação Crítica

- O **K-Means** apresentou melhor desempenho geral nas três métricas avaliadas, mostrando-se mais eficaz na identificação de grupos bem separados.
- O **Hierarchical Clustering** permitiu uma visualização interpretável por meio do **dendrograma**, porém a definição dos clusters foi menos nítida, refletida no menor Silhouette Score.
- O uso do **KFold** mostrou que ambos os modelos têm desempenho consistente, reforçando a robustez da análise.
- A segmentação com quatro clusters se mostrou adequada para distinguir perfis como:
    - Jovens com alto gasto (clientes potenciais);
    - Clientes mais velhos com gasto médio;
    - Clientes mais velhos com baixo gasto;
    - Consumidores jovens com pouco gasto;

5. Conclusão

Com base nos resultados obtidos:

- O **K-Means** foi o modelo que melhor segmentou os dados, com maior coesão interna e melhor separação entre clusters.
- O **Hierarchical Clustering**, apesar de apresentar resultados ligeiramente inferiores, é útil para compreender a estrutura hierárquica e as relações entre grupos.
- As métricas de validação cruzada confirmam que o modelo é estável e generalizável.
- A análise de **idade x pontuação de gasto** mostrou-se eficaz para compreender perfis de consumo e pode ser utilizada por empresas para estratégias de marketing segmentadas.