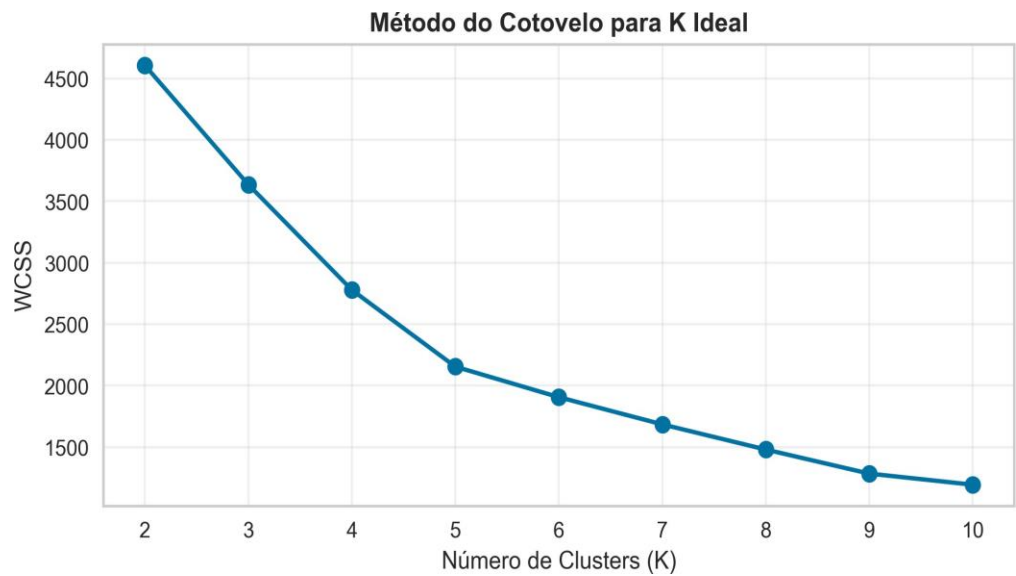


Relatório de Análise de Clusterização para K Ideal

Este relatório apresenta a análise das métricas de validação de clusters e a distribuição final dos clusters.

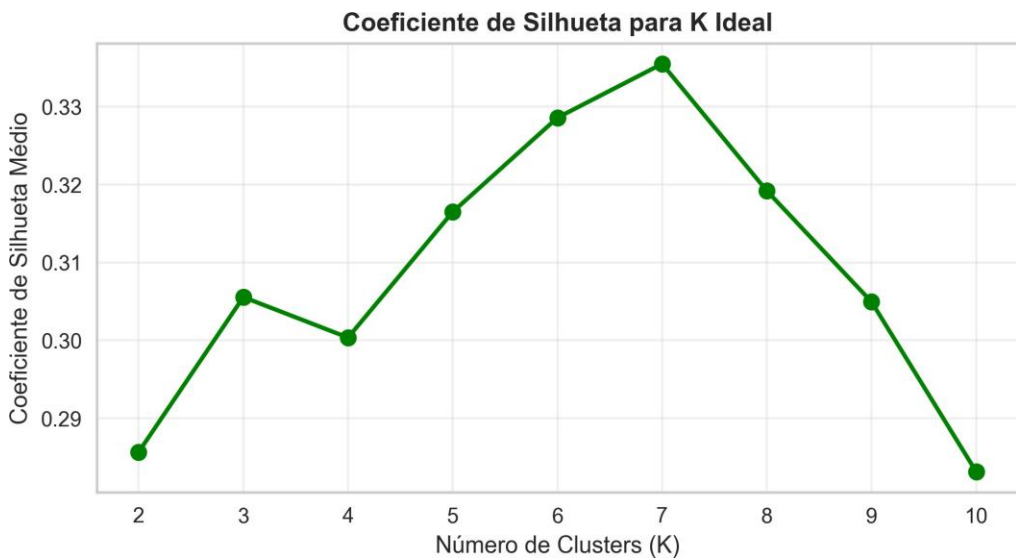
1. Análise de Métricas para Determinação do K Ideal

Método do Cotovelo (Elbow Method)

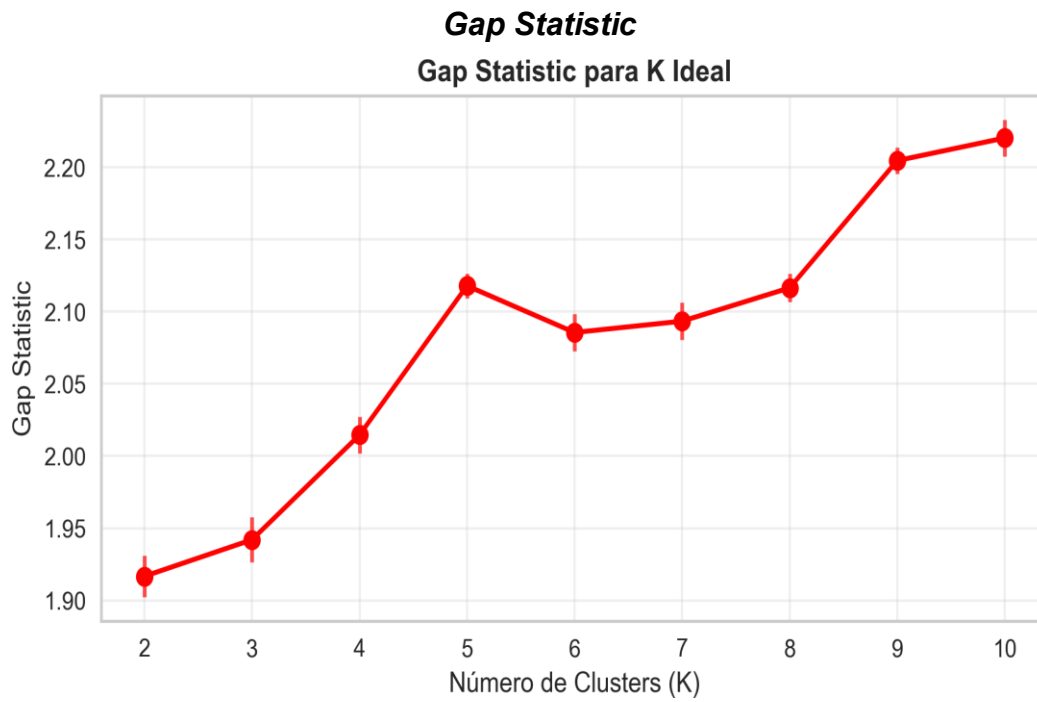


O Método do Cotovelo (WCSS) busca o ponto onde a soma dos quadrados intra-cluster (WCSS) começa a desacelerar significativamente. Visualmente, o 'cotovelo' aponta para um K onde os ganhos de agrupamento adicionais são marginais.

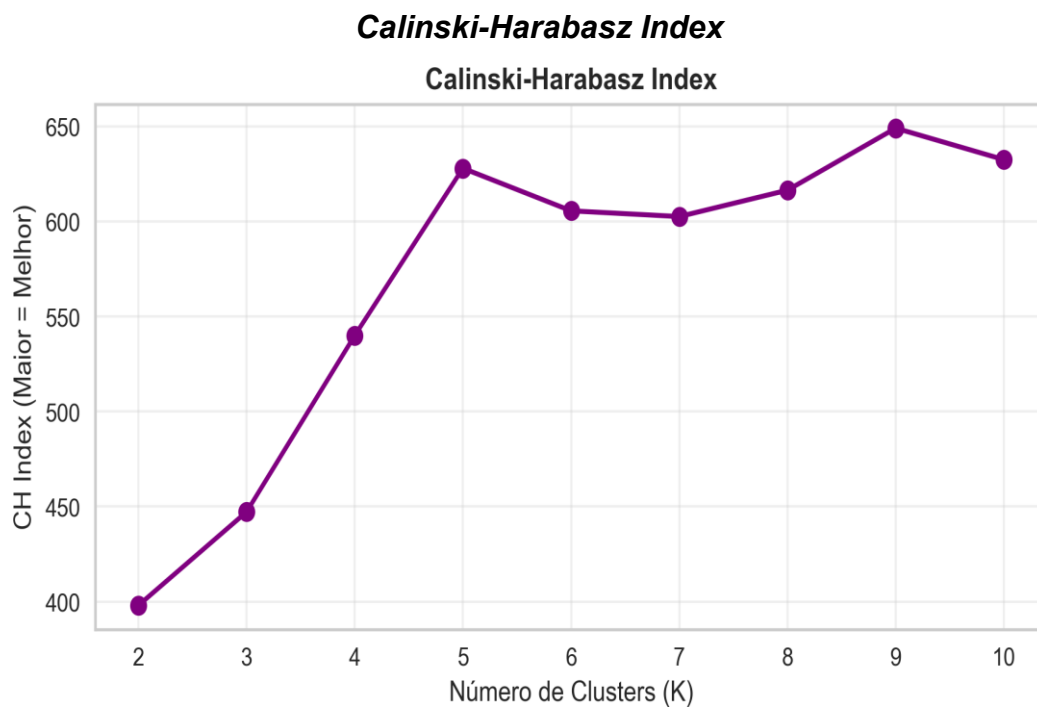
Coeficiente de Silhueta (Silhouette Score)



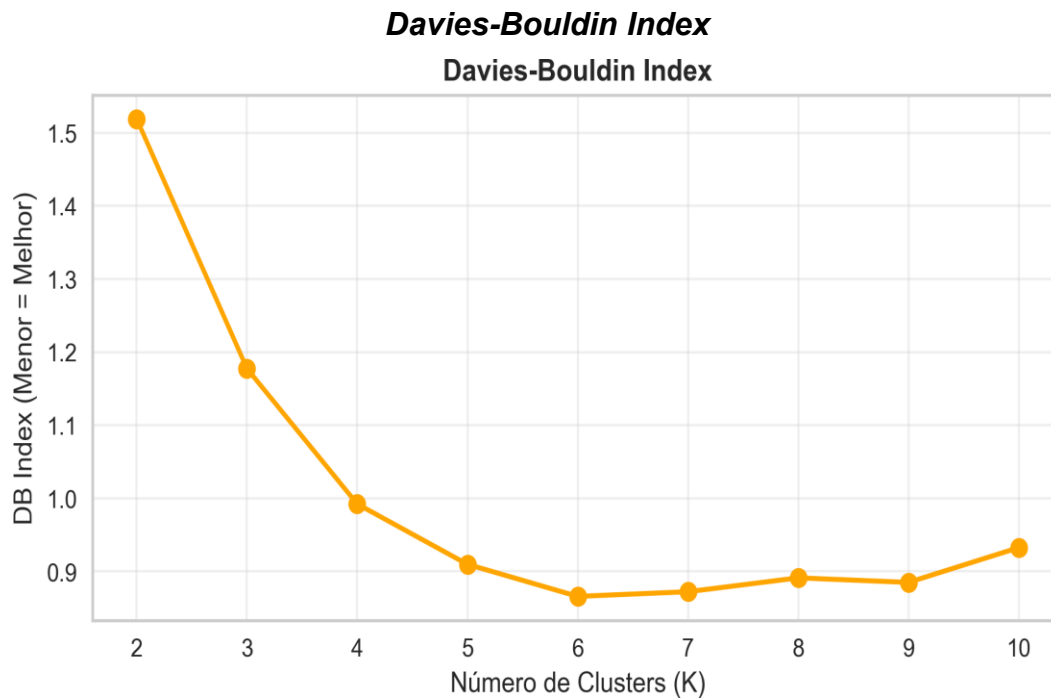
O Coeficiente de Silhueta mede o quão similar um objeto é ao seu próprio cluster em comparação com outros clusters. Valores próximos a 1 indicam uma boa separação e coesão, 0 indica sobreposição, e -1 indica uma possível atribuição incorreta.



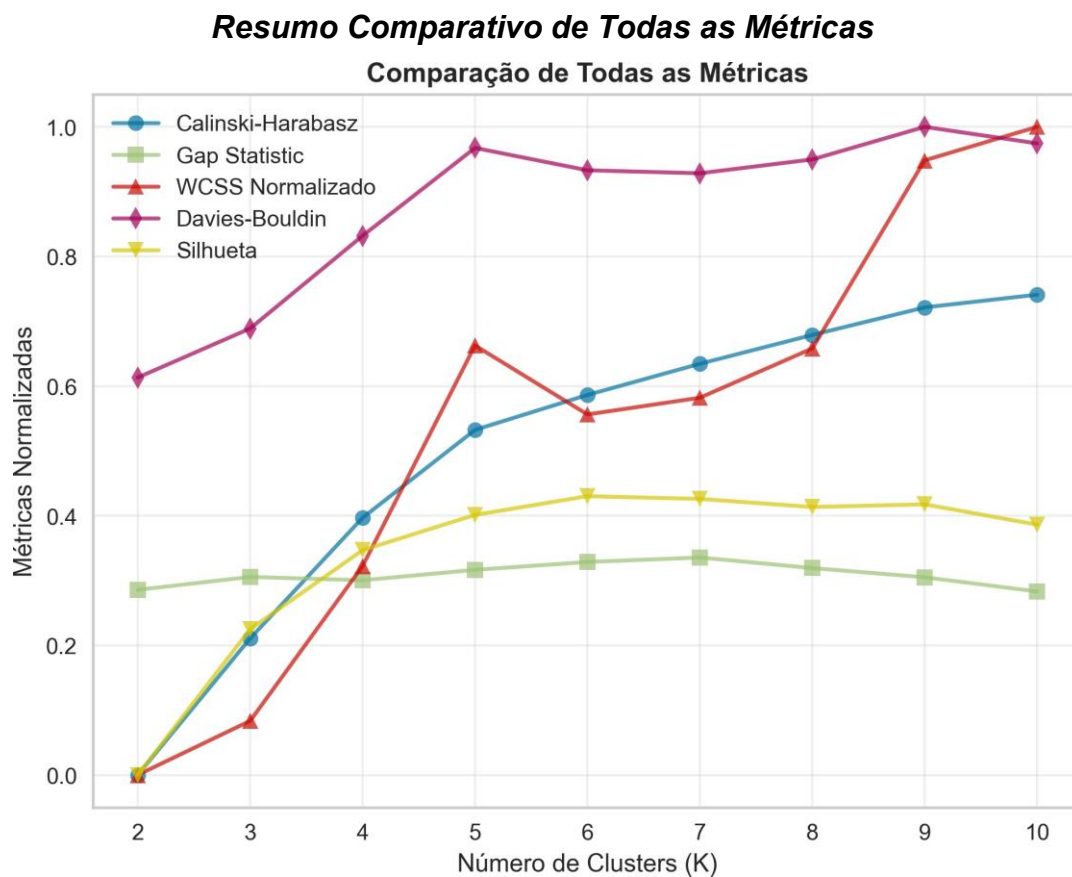
O Gap Statistic compara a dispersão intra-cluster dos dados reais com a dispersão esperada de uma distribuição de referência aleatória. Valores mais altos do Gap Statistic indicam uma melhor estrutura de agrupamento.



O Índice Calinski-Harabasz (também conhecido como Variance Ratio Criterion) mede a razão da dispersão inter-cluster para a dispersão intra-cluster. Um valor alto indica clusters densos e bem separados, buscando a maximização desta métrica.



O Índice Davies-Bouldin mede a similaridade média entre cada cluster e seu cluster mais similar. Um valor baixo indica que os clusters são bem separados e compactos, buscando a minimização desta métrica.



Este gráfico apresenta a comparação normalizada de diversas métricas de validação de clusters (Calinski-Harabasz, Gap Statistic, WCSS, Davies-Bouldin e Silhueta) em função do número de clusters (K). Ele oferece uma visão consolidada do desempenho de diferentes K's através de múltiplas perspectivas.

Resultados da Análise Automática por Métrica

Melhores K por métrica:

- Coeficiente de Silhueta: K = 7 (score: 0.3355)
- Gap Statistic: K = 10 (gap: 2.2202)
- Calinski-Harabasz Index: K = 9 (score: 649.04)
- Davies-Bouldin Index: K = 6 (score: 0.8657)

Recomendação para a Escolha do K Ideal

Análise e Recomendação do K Ideal

Com base na análise das diversas métricas de validação de clusters, observa-se que não há um valor único de K que optimize todas elas simultaneamente, o que é comum em problemas de clusterização. No entanto, podemos identificar um valor de compromisso:

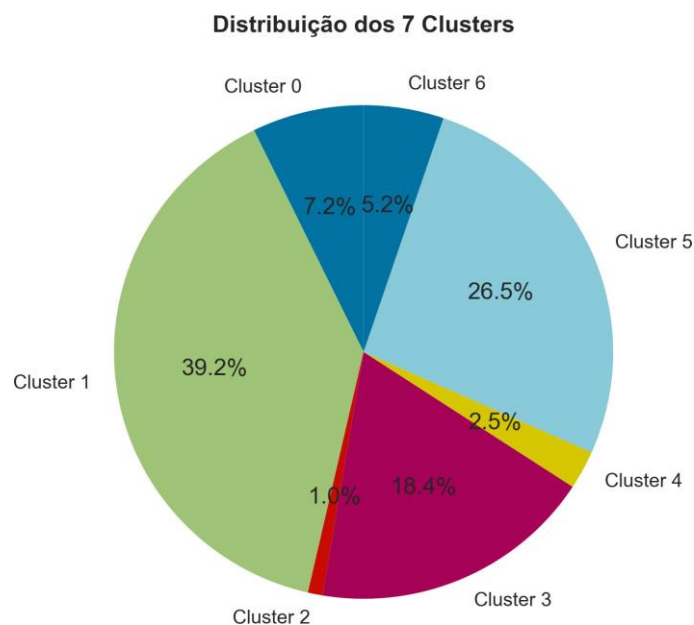
Métricas que buscam máxima separação e densidade (Calinski-Harabasz e Gap Statistic) tendem a indicar K's maiores (K=9 e K=10, respectivamente). Isso sugere uma estrutura de dados mais granular. Métricas que avaliam coesão e separação com menor complexidade (Silhueta e Davies-Bouldin) apontam para K's ligeiramente menores (K=7 e K=6, respectivamente).

Conclusão e Recomendação:

Considerando um balanço entre a qualidade do agrupamento e a interpretabilidade dos clusters, K=7 emerge como um forte candidato. Embora não seja o pico absoluto para todas as métricas, ele oferece um desempenho sólido para o Coeficiente de Silhueta e o Davies-Bouldin, e ainda apresenta valores muito bons para o Calinski-Harabasz e o Gap Statistic.

A escolha final de K=7 visa proporcionar clusters bem definidos que sejam, ao mesmo tempo, significativos para análise e aplicação prática, evitando a complexidade excessiva de um número muito elevado de grupos.

2. Distribuição dos 7 Clusters Finais



Este gráfico de pizza ilustra a distribuição percentual de cada um dos 7 clusters finais identificados, fornecendo uma visão da proporção de dados em cada grupo.