1 -

Silhouette Score k = 2: 0.629

Silhouette Score k = 3: 0.504

Silhouette Score k = 4: 0.443

Silhouette Score k = 5: 0.412

Silhouette Score k = 6: 0.341

Silhouette Score k = 7: 0.405

Silhouette Score k = 8: 0.330

Obtivemos Silhouette Scores de 0.629 para k = 2, 0.504 para k = 3 e valores decrescentes para valores maiores de k. Embora k = 2 tenha o maior Silhouette Score, k = 3 ainda apresenta uma separação razoável entre os clusters, permitindo capturar melhor a estrutura dos dados.

Agrupamentos:

- Cluster 1 (Iris-setosa): Com sépalas e pétalas menores, bem separado dos outros clusters.
- Cluster 2 (Iris-versicolour): Com dimensões intermediárias, posicionado entre *Iris-setosa* e *Iris-virginica*.
- Cluster 3 (Iris-virginica): Com sépalas e pétalas maiores, distante de *Iris-setosa* e próximo a *Iris-versicolour*.

2-

Método Elbow: mede o quanto os pontos estão compactos dentro de cada cluster.

$$WCSS(K) = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

(Within-Cluster Sum of Squares - WCSS)

- K é o número de clusters.
- Ci é o conjunto de pontos no cluster i.
- x é um ponto de dados em Ci.
- μi é o centróide (média) do cluster i.
- ||x−μi ||^2 representa a distância euclidiana ao quadrado entre o ponto x e o centróide μi.

consiste em calcular o WCSS para diferentes valores de K e escolher o valor onde a taxa de redução da WCSS diminui significativamente, formando o "cotovelo" da curva.

Silhouette Score: mede a separação e coesão dos clusters. Para cada ponto i, definimos duas variáveis:

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)}$$

μ_{in} (x_i) é a distância média de Xi para os pontos do seu próprio cluster.

$\mu^{max}{}_{out}$ é a distância média de Xi para os pontos dos clusters mais próximos.

Si está entre –1 e 1:

- Um valor próximo a 1 indica que xi está mais próximo dos pontos do seu próprio cluster e distante dos clusters vizinhos.
- Um valor próximo de 0 indica que xi está próximo da fronteira de dois clusters.
- Um valor próximo a −1 indica que xi está próximo dos pontos que pertencem a outro cluster

Sillhouette Index: mede o valor médio Si entre todos os pontos.

$$SilhouetteIndex = \frac{1}{n}\sum_{i=e}^{n} S_i$$

Os valores do silhouette index são avaliados como:

Valor	Significado
0,71 - 1	Estrutura forte
0,54 – 0,70	Estrutura razoável
0,26 – 0,50	Estrutura fraca e pode ser artificial
< 0,26	Nenhuma estrutura substancial

3-

Davies-Bouldin Score k = 2: 0.488

Davies-Bouldin Score k = 3: 0.787

Davies-Bouldin Score k = 4: 0.901

Davies-Bouldin Score k = 5: 0.950

Davies-Bouldin Score k = 6: 1.031

Davies-Bouldin Score k = 7: 1.036

Davies-Bouldin Score k = 8: 1.061

O Índice de Davies-Bouldin (DBI) é uma métrica que combina a separação e a compactação dos clusters para avaliar a qualidade de um agrupamento. Quanto menor o índice, melhor é a qualidade do agrupamento, indicando que os clusters são bem definidos, bem separados e com pouca variação interna.

O índice é baseado nas distâncias médias entre os centros dos clusters e nas dispersões dentro dos clusters. A ideia central é que, para um bom agrupamento:

1. Separação entre clusters deve ser grande: Quanto maior a distância entre os centros dos clusters, melhor.
2. Compactação dentro de cada cluster deve ser pequena: Os pontos dentro de cada cluster devem estar próximos uns dos outros.

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

- k é o número de clusters.
- σi é a dispersão do cluster i (média das distâncias dos pontos de i para o centro de i)
- d(ci ,cj) é a distância entre os centros dos clusters i e j.
- A fórmula calcula o valor máximo de cada par de clusters i e j e tira a média desse valor para todos os clusters

4-

K-means:

- O K-means encontrou o melhor desempenho com k = 2 clusters, apresentando um Silhouette Score de 0.629.

- À medida que o número de clusters aumentou, o Silhouette Score diminuiu, sugerindo que a divisão dos dados em mais de 2 clusters não resultou em agrupamentos mais coesos.
- Conclusão: O K-means indica que a melhor quantidade de clusters é 2, já que esse valor gerou a melhor separação entre os grupos.

SOM (Self-Organizing Maps):

- O SOM foi testado para diferentes configurações de mapa (com diferentes valores para x e y, que representam o número de unidades nas direções do mapa).
- Os melhores resultados ocorreram para x = 3, y = 2 e x = 5, y = 2, com um Silhouette Score de 0.487.
- O SOM não está limitado ao número de clusters como o K-means, mas sim à topologia do mapa. A ideia é que o SOM ajuda a organizar dados em uma estrutura que favorece a separação dos grupos.
- Conclusão: O SOM obteve bons resultados para mapas com dimensões de 3x2 ou 5x2, sugerindo que ele encontrou uma forma de organizar os dados em grupos coesos.

DBSCAN:

- O DBSCAN foi testado para diferentes valores de eps (tamanho da vizinhança) e min_samples (número mínimo de pontos necessários para formar um cluster).
- A maioria das combinações de eps e min_samples não gerou clusters válidos, indicando que o DBSCAN teve dificuldades com os parâmetros escolhidos.
- Quando os parâmetros estavam bem ajustados (eps = 0.3, min_samples = 5 ou 7), o Silhouette Score foi de 0.468, mas o desempenho geral foi inferior ao K-means.
- Conclusão: O DBSCAN teve dificuldades em encontrar clusters válidos para a maioria dos parâmetros. Ele gerou apenas uma boa solução com eps = 0.3, min_samples = 5 ou 7, mas ainda assim, o desempenho foi inferior ao K-means.
-

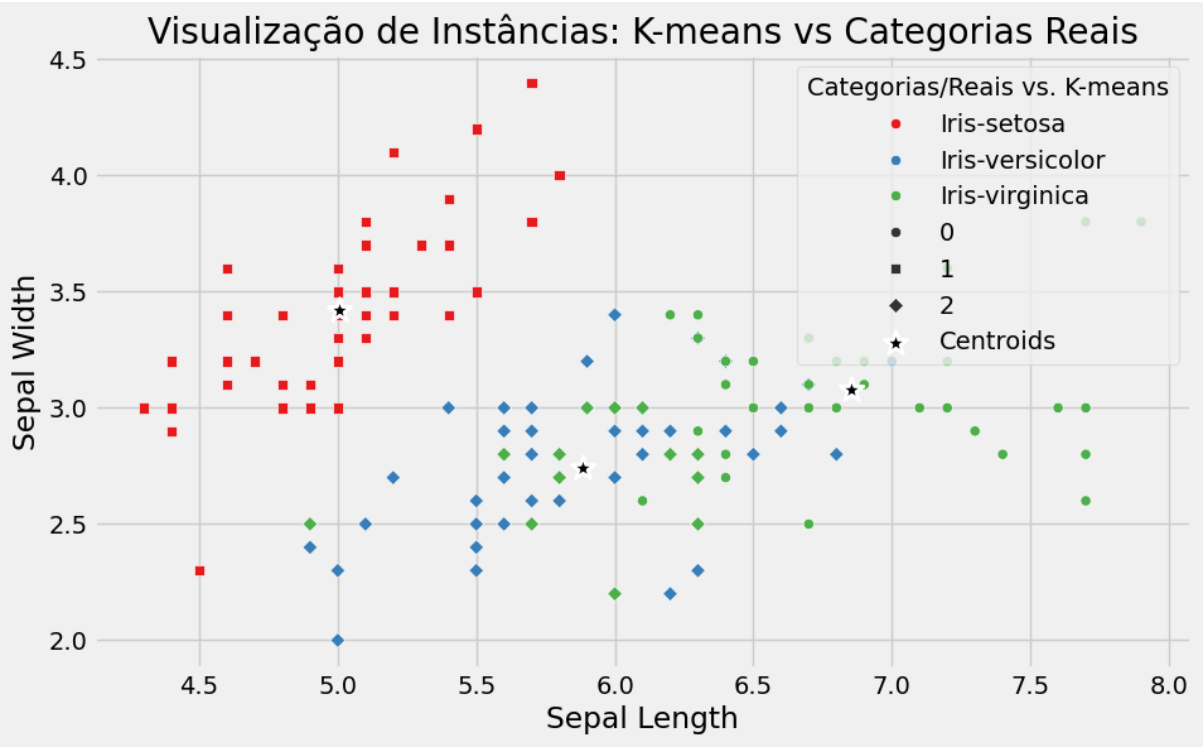
Comparação dos métodos:

K-means sugeriu 2 clusters como o número ideal, com o melhor Silhouette Score.

SOM não depende do número de clusters, mas encontrou uma estrutura de mapa de 3x2 ou 5x2 como a mais adequada para os dados.

DBSCAN apresentou dificuldades em encontrar clusters válidos na maioria das configurações, mas quando foi bem-sucedido, gerou uma solução com eps = 0.3, min_samples = 5 ou 7.

5-



6 - Por ser uma base de dados construída para fins didáticos e amplamente aceita para experimentos básicos, a base Iris já foi preparada e validada para que os pesquisadores e desenvolvedores possam focar diretamente na modelagem dos algoritmos de aprendizado, ao invés de gastar tempo com etapas de pré-processamento. Dessa forma, todos os valores são consistentes e as variáveis estão escaladas de maneira adequada para uma análise inicial.

LINKS:
https://colab.research.google.com/drive/1XittKE_XGkLmb9Ld2dkJrrWqM1zk9YXT?usp=sharing