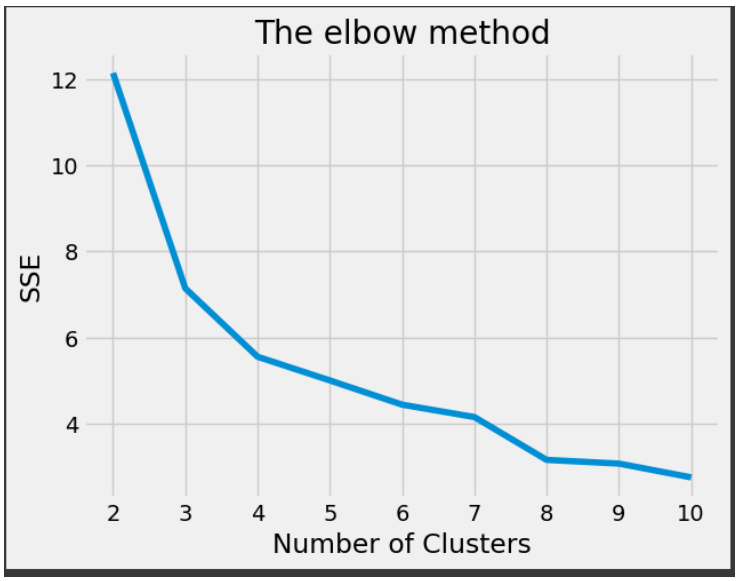
Lista 6 – IA

1_



No gráfico do método Elbow, o cotovelo mais evidente está em k=2. A redução do SSE é menos acentuada, assim, k=2 é uma escolha razoável para o número de clusters com base nesse método.

Silhouette Score k = 2: 0.629

Silhouette Score k = 3: 0.504

Silhouette Score k = 4: 0.444

Silhouette Score k = 5: 0.353

Silhouette Score k = 6: 0.319

Silhouette Score k = 7: 0.415

Silhouette Score k = 8: 0.324

K=2 ter a maior pontuação *Silhouette*.

Conclusão:

k = 2 proporciona clusters mais coesos e bem separados, mas com menos detalhes dos dados.

2_

Método Elbow:

Usa a Soma dos Erros Quadrados (SSE, do inglês Sum of Squared Errors) para avaliar a qualidade dos agrupamentos. A SSE mede a coesão dentro de cada cluster, ou seja, o quão próximos os pontos estão do centroide do cluster.

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

- k é o número de clusters.
- Ci é o conjunto de pontos pertencentes ao cluster i.
- x é um ponto de dados pertencente ao cluster Ci .
- μi é o centroide do cluster i, ou seja, a média dos pontos em Ci.
- ‖x−μi‖^2 é a distância euclidiana quadrada entre o ponto x e o centroide μi.

Conforme o número de clusters aumenta, a SSE diminui, mas em certo ponto, a redução passa a ser menos significativa. Esse ponto é onde o número de clusters k é ideal.

Silhouette Score:

A pontuação Silhouette mede a qualidade de um ponto em relação ao seu cluster e aos clusters vizinhos, e varia de -1 a 1. Valores próximos de 1 indicam que o ponto está bem ajustado ao seu próprio cluster e mal ajustado aos clusters vizinhos. Valores próximos de -1 indicam que o ponto pode estar mal alocado no cluster.

Para cada ponto xi, a pontuação Silhouette s é calculada como:

$$s = \frac{b - a}{\max(a, b)}$$

Onde:

- a é a distância média entre xi e todos os outros pontos no mesmo cluster. Essa métrica mede a coesão dentro do cluster.
- b é a distância média entre xi e todos os pontos no cluster mais próximo, ou seja, o cluster vizinho mais próximo ao ponto xi. Essa métrica mede a separação do cluster.

A pontuação *Silhouette* para o agrupamento inteiro é a média das pontuações *Silhouette* de todos os pontos:

$$\text{Silhouette Score} = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

3

O **Índice de Dunn** é uma métrica útil para avaliar a qualidade dos agrupamentos, pois mede o quão bem separados e compactos os clusters estão. Em geral, quanto maior o índice de Dunn, melhor a qualidade do agrupamento.

Essa métrica é calculada considerando a menor distância entre diferentes clusters e a maior distância entre pontos dentro do mesmo cluster. Em outras palavras, o Índice de Dunn visa maximizar a separação entre os clusters e minimizar a dispersão interna de cada um, o que resulta em agrupamentos bem definidos quando o índice é alto.

Com o Índice de Dunn calculado em aproximadamente 0.067 utilizando 4 clusters na base de dados *Iris*, o valor sugere uma qualidade de agrupamento relativamente baixa. O índice, que mede a coesão e a separação entre clusters, indica que, mesmo com 4 clusters, há uma separação limitada entre eles ou uma alta dispersão interna dentro de pelo menos um dos grupos.

4_

Resultados dos agrupamentos:

KMeans - Número de Clusters: 4, Silhouette Score: 0.38724679456128813

indica uma qualidade moderada dos agrupamentos, com algum nível de sobreposição entre os clusters. Essa métrica sugere que há certa coesão e separação, mas os clusters podem estar se sobrepondo levemente, o que não é ideal.

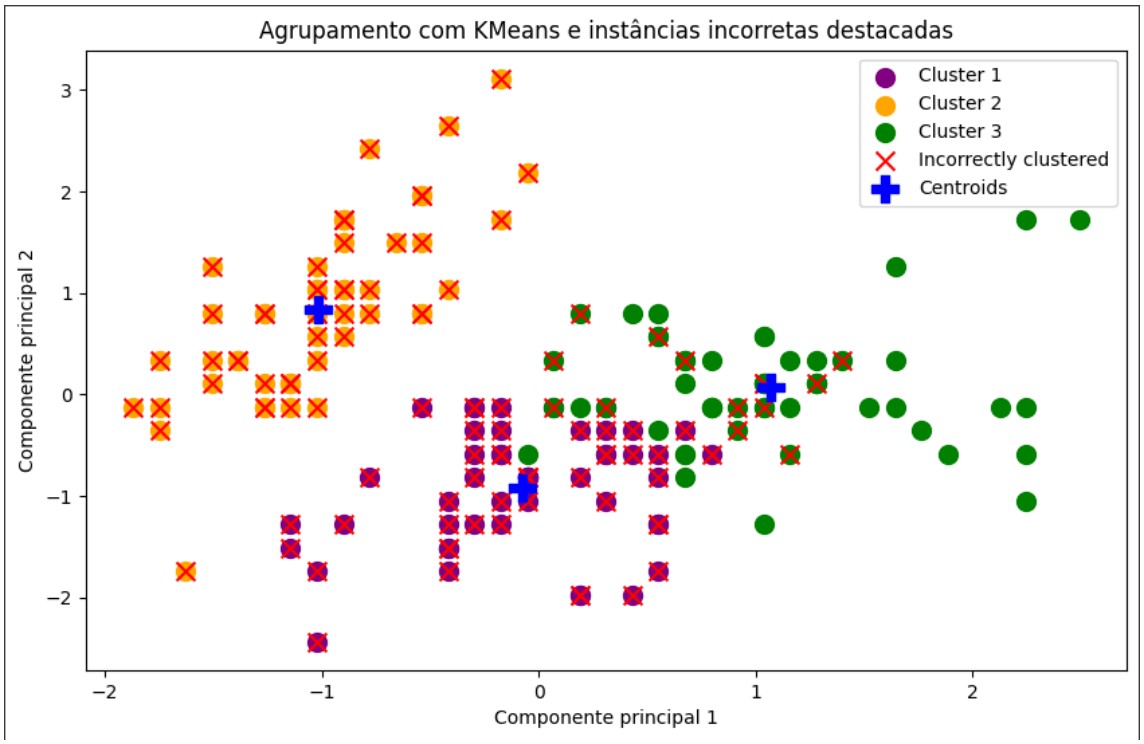
DBSCAN - Número de Clusters: 2, Silhouette Score: 0.6531989922140501

O DBSCAN identificou 2 clusters de alta densidade. O Silhouette Score de 0.653 indica uma boa qualidade de agrupamento, com uma alta coesão interna e boa separação entre os clusters. Isso sugere que o DBSCAN foi mais eficaz em encontrar grupos bem separados na base Iris. No entanto, o fato de identificar apenas 2 clusters implica que o algoritmo não encontrou tantas subdivisões nos dados, o que é uma limitação já que o objetivo é identificar 3 espécies.

SOM - Número de Clusters: 4, Silhouette Score: 0.24399888381147813

O SOM foi configurado com uma grade 2x2, resultando em 4 clusters. No entanto, o Silhouette Score de 0.244 indica uma baixa qualidade de agrupamento. Esse valor sugere que os clusters estão se sobrepondo significativamente ou que a coesão interna dos grupos é fraca.

5_



O KMeans conseguiu identificar corretamente um cluster bem distinto (Iris-setosa), mas falhou em separar os clusters sobrepostos (Iris-versicolor e Iris-virginica).

6_

1. Pré-processamento dos Dados

Para preparar a base de dados para análise, foram realizadas as seguintes etapas de pré-processamento:

1. Carregamento dos Dados:
    - A base de dados *Iris* foi carregada com colunas representando quatro características (comprimento e largura de sépala e pétala) e uma coluna de rótulos de classe (Iris-setosa, Iris-versicolor e Iris-virginica).
2. Remoção de Rótulos:
    - Para o processo de agrupamento, foram removidos os rótulos de classe, deixando apenas as características numéricas. Isso é necessário para que o algoritmo de agrupamento funcione de forma não supervisionada.
3. Padronização dos Dados:
    - Para evitar o viés em relação à escala das características, os dados foram padronizados utilizando o StandardScaler. Este processo garantiu que todas as características tivessem uma média de 0 e desvio padrão de 1, permitindo que o algoritmo atribuísse igual importância a cada dimensão.

2. Aplicação dos Algoritmos de Agrupamento

Após o pré-processamento, aplicamos três algoritmos de agrupamento: KMeans, DBSCAN e SOM. Abaixo estão os detalhes e resultados de cada aplicação.

1. KMeans:
    - Configuramos o KMeans para formar 4 clusters, conforme a análise de métodos como o Elbow e o Índice de Dunn, que sugeriram um agrupamento razoável em torno desse número.
    - O Silhouette Score foi de aproximadamente 0.387, indicando uma qualidade moderada dos clusters. Observamos que um dos clusters correspondia predominantemente à classe *Iris-setosa*, enquanto os demais clusters mostraram considerável sobreposição, especialmente entre *Iris-versicolor* e *Iris-virginica*.
2. DBSCAN:
    - O algoritmo DBSCAN encontrou apenas 2 clusters, identificando agrupamentos baseados na densidade. Com um Silhouette Score de 0.653, este método demonstrou maior coesão e separação, sugerindo clusters bem definidos.
    - O DBSCAN foi eficaz em identificar grupos densos, mas, por utilizar a densidade como base, ele agrupou algumas instâncias de *Iris-versicolor* e *Iris-virginica* juntas, o que gerou um agrupamento com menos clusters.
3. SOM (Self-Organizing Map):
    - Utilizando uma grade 2x2, o SOM produziu 4 clusters. O Silhouette Score foi de 0.244, indicando uma baixa qualidade de agrupamento. Isso ocorreu porque o SOM, ao projetar os dados em uma grade, não separou claramente as classes sobrepostas.
    - Os clusters formados pelo SOM mostraram bastante sobreposição, e o algoritmo teve dificuldade em separar as instâncias de *Iris-versicolor* e *Iris-virginica*.

3. Análise Comparativa e Métrica de Avaliação

- Índice de Dunn: O índice de Dunn para os clusters formados com o KMeans foi calculado e apresentou um valor de 0.0669. Esse valor baixo indica que, mesmo com 4 clusters, o KMeans teve dificuldades em manter uma boa separação entre os grupos, especialmente nas classes com características semelhantes.
- Silhouette Score:
    - O DBSCAN teve o maior Silhouette Score (0.653), indicando que os clusters eram bem definidos e coesos. No entanto, ele identificou menos clusters do que o esperado (2 clusters), devido à dependência da densidade.
    - O KMeans e o SOM apresentaram Silhouette Scores mais baixos (0.387 e 0.244, respectivamente), destacando limitações em separar dados sobrepostos.

4. Conclusões

- Desempenho dos Algoritmos:
    - O KMeans foi o mais apropriado quando a tarefa exigia uma divisão em 4 grupos, mas teve dificuldades com classes sobrepostas.
    - O DBSCAN apresentou a melhor coesão e separação, mas encontrou apenas 2 clusters. Ele é mais apropriado para dados com densidade variável e menos sobreposição.
    - O SOM teve o menor desempenho em termos de separação dos clusters. O SOM projetou os dados em uma grade, mas não foi eficaz em clusters com fronteiras pouco definidas.
- Implicações:
    - Para a base de dados *Iris*, em que duas das três classes possuem características sobrepostas, o KMeans e o DBSCAN apresentaram vantagens e desvantagens distintas. O KMeans foi mais flexível em relação ao número de clusters, enquanto o DBSCAN mostrou maior qualidade para grupos de alta densidade.

Link que com todos os códigos utilizados:
https://colab.research.google.com/drive/1rKjZmEk5ZTFN4Vp5HbHgLGNwcPqQn-Eh?usp=sharing