Lista #7

Disciplina: Inteligência Artificial

Iago Fereguetti Ribeiro

Código no colab:
https://colab.research.google.com/drive/1ILcxC6UTn81YUNPaYPCYpXxVW_VUKJed#scrollTo=qdvdgfKbXCyt

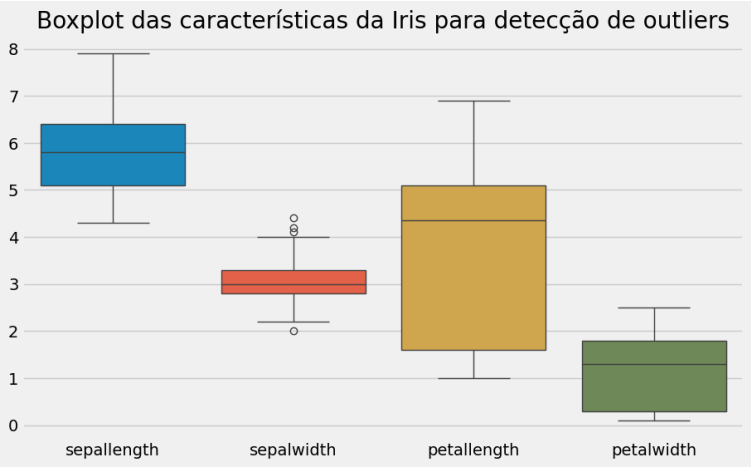
Código no github:

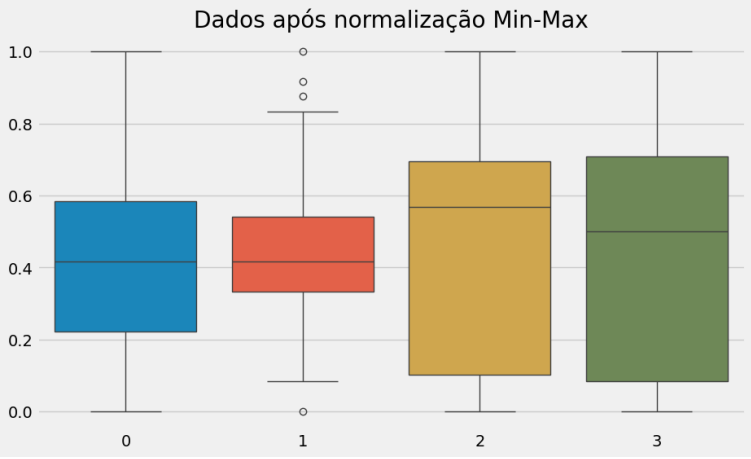
https://github.com/Iago-Fereguetti18/IAListA-07.git

Questão 01:

Número de outliers detectados: 4

Índices dos outliers: [32, 33, 60, 15]

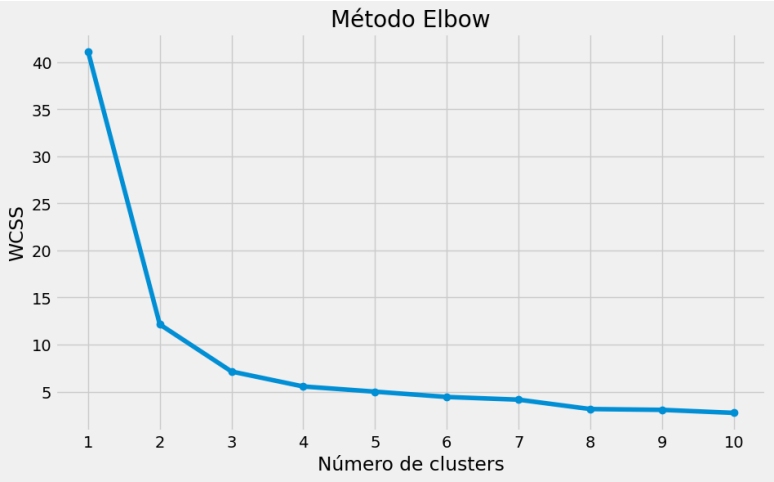


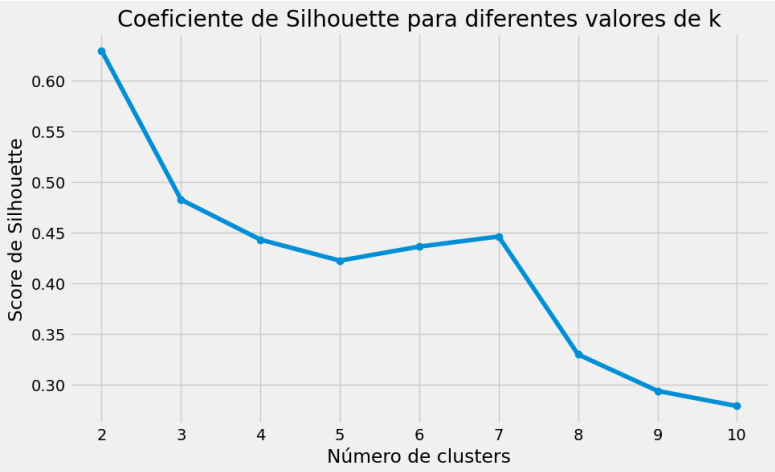


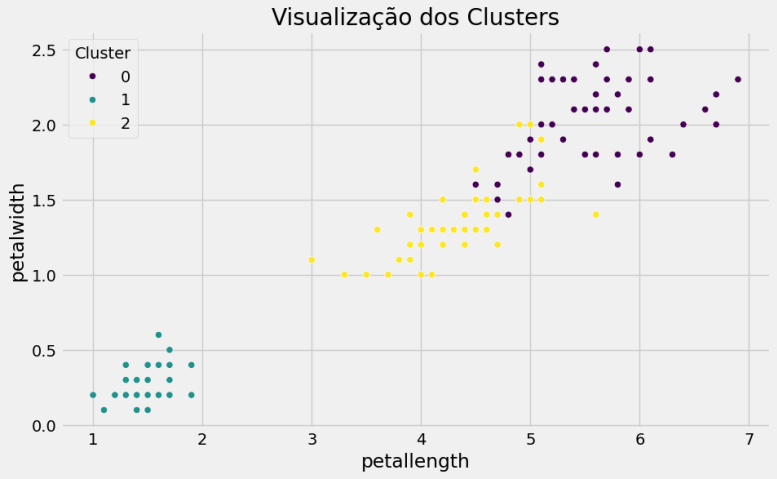
Questão 02:

Cluster

0.	6.703846	3.063462	5.467308	1.976923
1.	5.006000	3.418000	1.464000	0.244000
2.	5.783333	2.664583	4.297917	1.350000







Questão 03:

Escolha do centroide inicial

O K-means++ é uma melhoria no algoritmo K-means tradicional que seleciona os centroides iniciais de forma mais inteligente, reduzindo a probabilidade de convergência para ótimos locais.

Métricas de distância

Euclidiana: Distância "em linha reta" entre dois pontos no espaço euclidiano

Manhattan: Soma das diferenças absolutas entre as coordenadas

Cosine: Mede o cosseno do ângulo entre dois vetores

Podemos testar diferentes métricas usando o KMeans personalizado ou outros algoritmos como o K-medoids.

Questão 04:

WCSS

WCSS (Within-Cluster Sum of Squares) calcula a soma das distâncias quadradas entre cada ponto e o centroide do seu cluster:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i}^{\square} \| x - \mu_i \|^2$$

Onde:

k = número de clusters

Ci = pontos no cluster i

μi = centroide do cluster i

Coeficiente de Silhouette

Para cada ponto i:

Calcule a(i)a(i) = distância média para todos os pontos no mesmo cluster

Calcule b(i)b(i) = menor distância média para pontos em qualquer outro cluster

O score para o ponto i é:

Coeficiente de Silhouette

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

O score global é a média de s(i) para todos os pontos.

Questão 05:

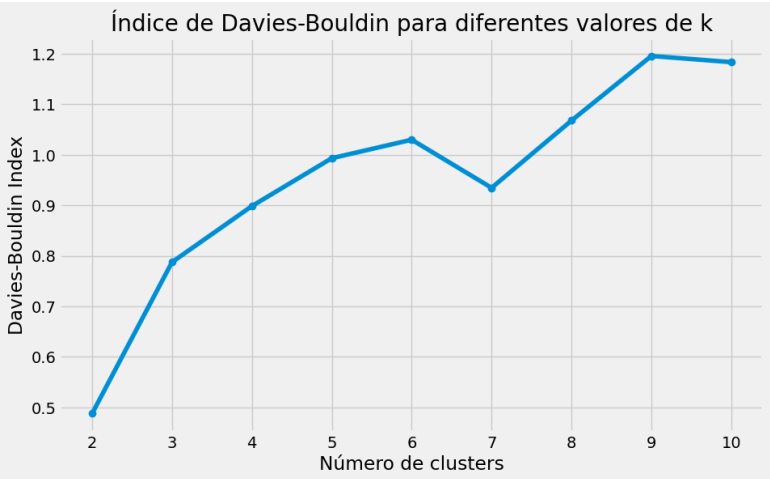
O índice Davies-Bouldin (DBI) é uma métrica utilizada para avaliar a qualidade de uma clusterização, o DBI mede o quão bem separados e compactos estão os clusters. Quanto menor o valor, melhor a clusterização (clusters mais distintos entre si e mais coesos internamente), o índice compara a distância entre os centróides dos clusters e a dispersão interna de cada cluster. Para cada par de clusters i e j, calcula-se:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

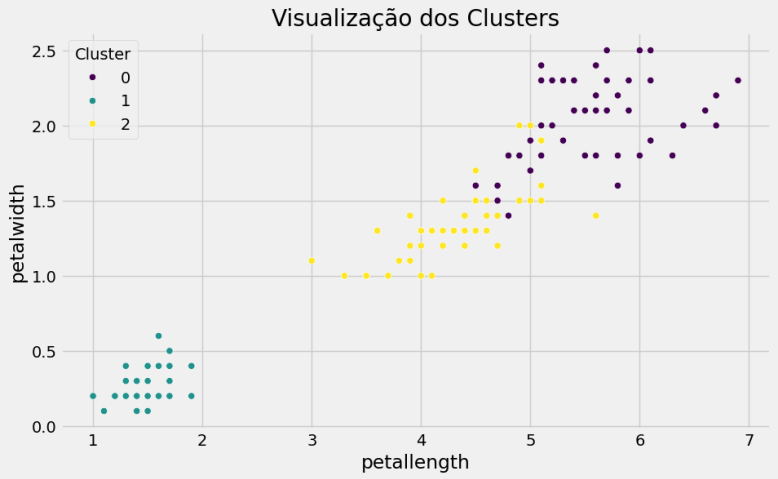
$$d_{ij} = d(v_i, v_j), \; s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i)$$

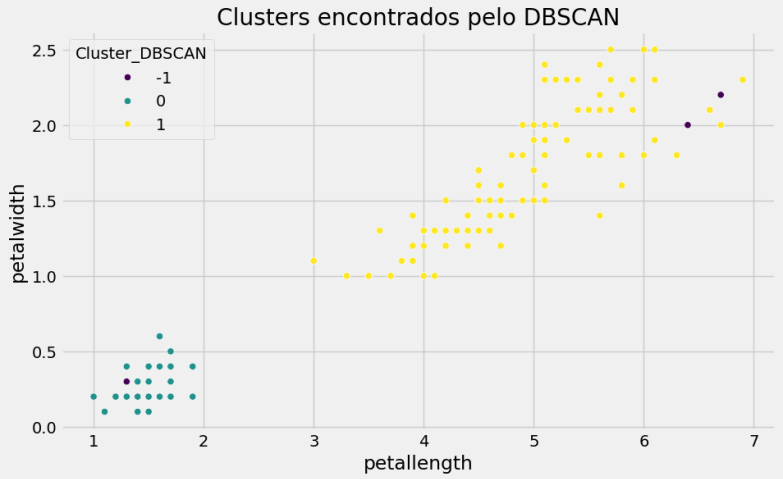
Implementação do DBI:

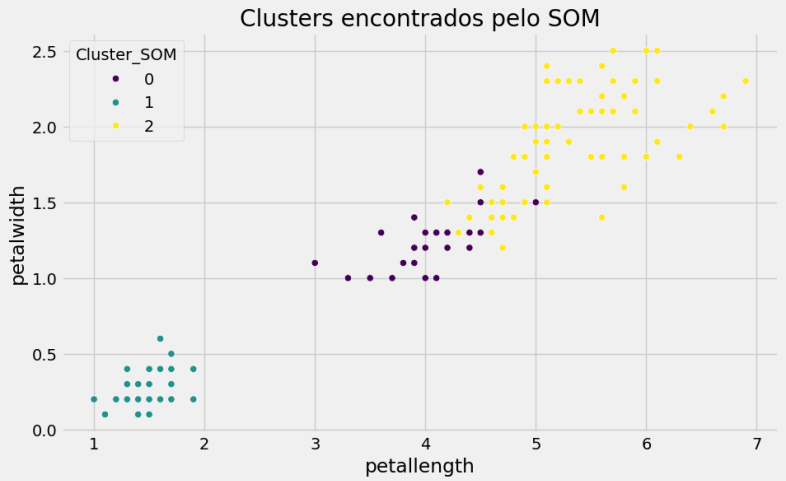
Index para k=2: 0.488 Index para k=3: 0.787 Index para k=4: 0.899 Index para k=5: 0.993 Index para k=6: 1.030 Index para k=7: 0.935 Index para k=8: 1.068 Index para k=9: 1.196 Index para k=10: 1.184



Questão 06:

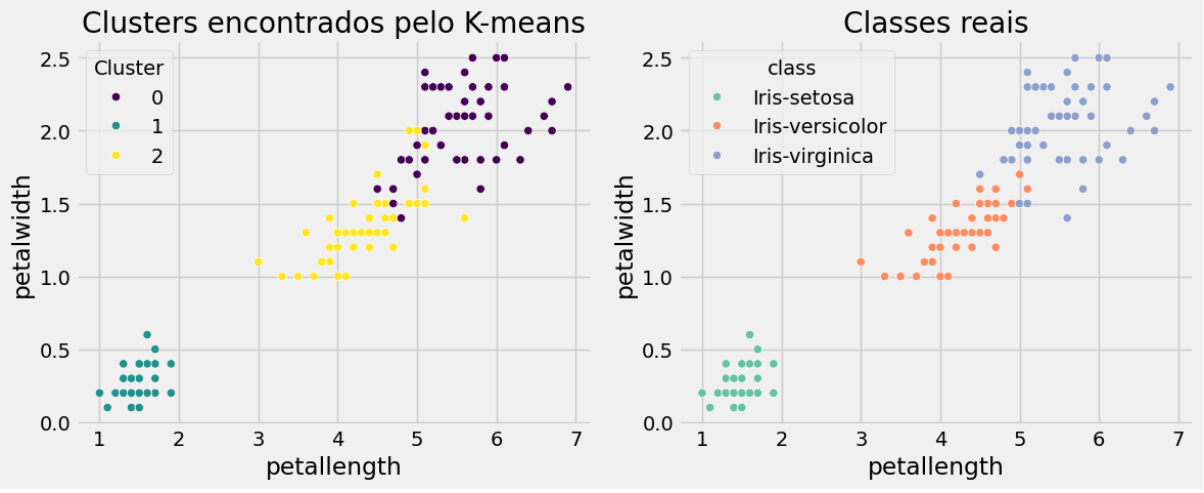


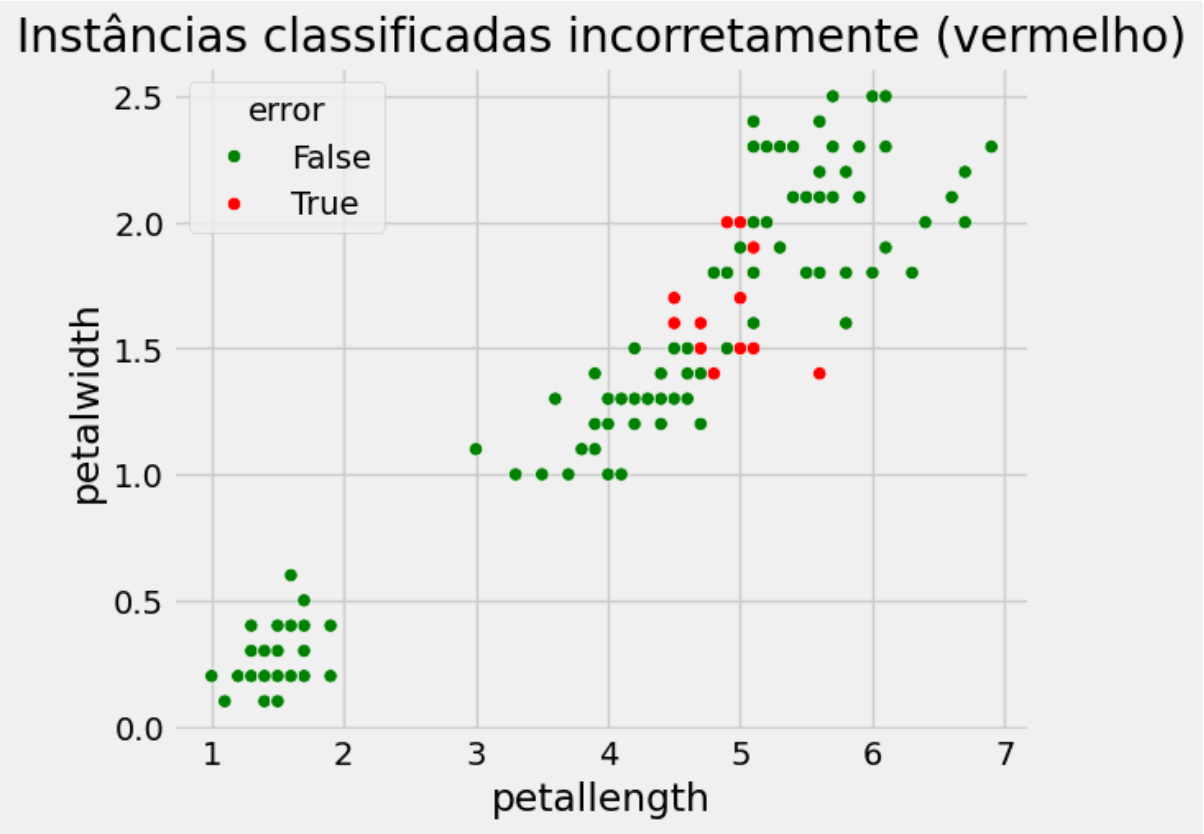




Não encontraram o mesmo número, pois o DBSCAN encontrou 2 e o restante encontrou 3.

Questão 07:





Ele teve uma acurácia: 88.00%

Questão 08:

Questão 01 – Detecção de Outliers

- Quantidade de outliers detectados: 4
- Índices identificados: 32, 33, 60, 15
- Objetivo: Identificar e tratar valores extremos que possam comprometer a análise.

Questão 02 – Agrupamento com K-means

- Número de clusters: 3
- Centroides obtidos:
    - Cluster 0: 6.703846, 3.063462, 5.467308, 1.976923
    - Cluster 1: 5.006000, 3.418000, 1.464000, 0.244000
    - Cluster 2: 5.783333, 2.664583, 4.297917, 1.350000

Questão 03 – Inicialização e Métricas de Distância

- Inicialização: K-means++
- Métricas utilizadas: Euclidiana, Manhattan, Cosseno
- Observação: Essas métricas afetam diretamente a formação dos agrupamentos.

Questão 04 – Avaliação de Agrupamentos

- WCSS: Soma das distâncias quadradas entre pontos e centroides.
- Coeficiente de Silhouette: Mede o quão bem cada ponto está agrupado.
- Valores próximos de 1 indicam boa separação.

Questão 05 – Índice Davies-Bouldin (DBI)

- Finalidade: Avaliar separação e compactação dos clusters.
- Resultados por valor de k:
    - k=2: 0.488
    - k=3: 0.787
    - k=4: 0.899
    - k=5: 0.993
    - k=6: 1.030
    - k=7: 0.935
    - k=8: 1.068

    - k=9: 1.196
    - k=10: 1.184
- Melhor valor: k=2

Questão 06 – Comparação de Métodos de Agrupamento

- DBSCAN encontrou 2 clusters.
- K-means e outros métodos encontraram 3 clusters.
- A diferença se deve ao critério de densidade usado no DBSCAN.

Questão 07 – Classificação

- Acurácia do modelo: 88,00%
- Interpretação: resultado satisfatório.