Inteligência Artificial - Lista #7

Aluno: Samuel Horta de Faria - 801528

Questão 1

A base de dados Iris foi carregada, e os atributos numéricos foram normalizados utilizando o método **Min-Max Scaling**, garantindo que os valores sejam entre 0 e 1.
Além disso, foi utilizado o **Z-score** para identificar e remover possíveis outliers (valores que estão além de 3 desvios-padrão da média).

Questão 2

Foi aplicado o algoritmo **KMeans** e foi avaliado a qualidade dos agrupamentos utilizando duas métricas:

- **Elbow Method:** indicou que o número ideal de clusters é **3**, o que coincide com as classes reais da base.
- **Silhouette Score:** também apresentou valor máximo para k = 3, confirmando que os dados foram bem agrupados.

Os grupos formados são coerentes com as classes reais: Setosa, Versicolor e Virginica.

Questão 3

Foi explorado os principais hiperparâmetros do KMeans:

- init='k-means++': melhora a inicialização dos centróides;
- n_init=10: executa o algoritmo várias vezes para evitar mínimos locais;
- max_iter=300: garante convergência;
- A métrica usada pelo scikit-learn é a **distância euclidiana**, que mede a proximidade entre pontos e centróides.

Questão 4

As duas fórmulas essenciais foram Distância Euclidiana e Silhouette Score.
Onde:

- a(i): distância média entre o ponto e os demais do mesmo cluster.
- b(i): menor distância média até outro cluster.

Questão 5

Foi implementado o **Davies-Bouldin Index**, que mede a qualidade dos clusters com base na dispersão intra-cluster e separação inter-cluster.
Valores menores indicam agrupamentos mais bem definidos. Essa métrica complementa a análise do Silhouette.

Questão 6

Foi aplicado:

- **DBSCAN:** detecta clusters baseando-se em densidade. Identificou menos grupos, ignorando alguns ruídos (outliers).
- **SOM (Self-Organizing Maps):** rede neural não supervisionada. Os grupos formados foram similares ao KMeans, mas organizados em uma grade.

Ao comparar com o KMeans, observamos que DBSCAN pode detectar clusters de formato arbitrário, enquanto o KMeans assume que são esféricos e de tamanho similar.

Questão 7

Como a base Iris possui rótulos reais, foi possível comparar os clusters do KMeans com as classes reais.
Utilizei **PCA** para reduzir os dados para 2D e gerar uma visualização.
O índice **ARI (Adjusted Rand Index)** foi calculado para medir a correspondência entre os rótulos reais e os clusters.
A maioria das instâncias foi agrupada corretamente, com exceção de algumas sobreposições entre Versicolor e Virginica.

Questão 8

Foi realizado um processo completo de agrupamento, com as seguintes etapas:

- **Pré-processamento:** normalização e remoção de outliers.
- **Execução:** aplicação do KMeans, DBSCAN e SOM.
- **Avaliação:** Elbow, Silhouette, Davies-Bouldin, ARI.
- **Visualização:** análise gráfica com PCA.

Conclui-se que o KMeans teve um desempenho muito bom nesta base, mas que o DBSCAN pode ser mais eficiente em dados com clusters de formatos complexos ou com outliers significativos.

Link para código ipynb.