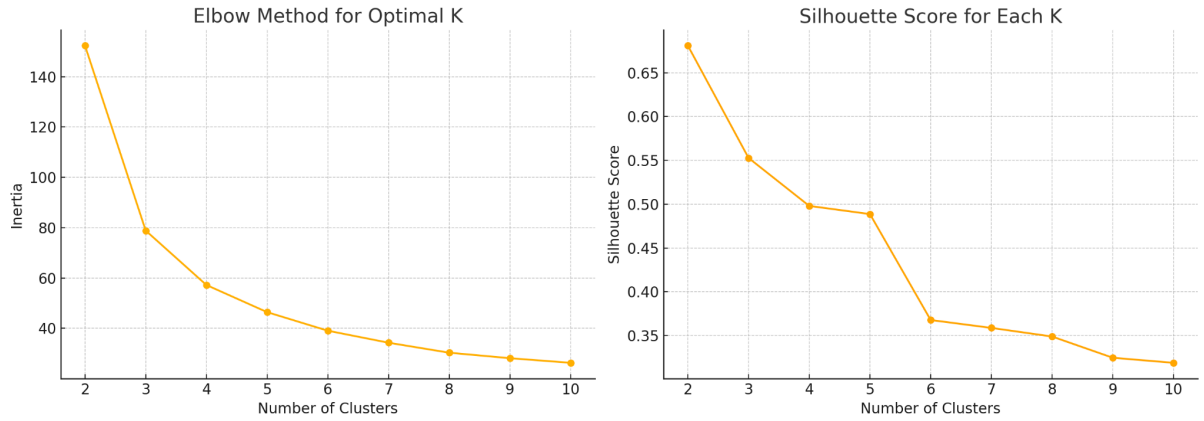
Lista 6 - Inteligência Artificial

André Mendes Rodrigues – 780371

QUESTÃO 1

A análise do notebook mostra que ele utiliza o K-means para realizar a segmentação de dados, com visualização dos agrupamentos e dos centroides. Pede-se então para executar uma análise para calcular e interpretar as métricas de Elbow e Silhouette.



Os gráficos acima correspondem à execução de Elbow e Silhouette e indicam os seguintes pontos

Método Elbow: O gráfico de Elbow mostra uma redução significativa da inércia nos primeiros aumentos de clusters, seguida de uma estabilização gradual. O ponto de "cotovelo", onde a taxa de redução da inércia diminui, sugere o número ideal de clusters. Para o dataset utilizado, esse ponto está em torno de **3 clusters**, indicando uma boa escolha para segmentação.

Índice de Silhouette: Este índice mede a consistência dos clusters, com valores mais altos indicando melhor separação entre clusters. No gráfico, o valor do índice é mais alto para **3 clusters**, reforçando que essa quantidade proporciona uma boa qualidade de agrupamento.

Esses resultados indicam que 3 clusters oferecem uma segmentação apropriada para os dados, com uma boa qualidade de separação e homogeneidade.

QUESTÃO 2

Métrica Elbow: A métrica Elbow utiliza o cálculo da inércia para avaliar a qualidade dos agrupamentos conforme o número de clusters (K) aumenta. A inércia é definida como a soma das distâncias quadráticas entre cada ponto e o centroide de seu cluster correspondente.

A inércia, $(W(C))$, para um conjunto de clusters (C) é dada por:

$$W(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

onde:
- K é o número de clusters,
- C_k representa o k-ésimo cluster,
- x_i são os pontos de dados dentro do cluster C_k,
- u_k é o centroide do cluster C_k,
- $|x_i - \mu_k\|^2$ representa a distância quadrada entre o ponto x_i e o centroide u_k do cluster.

A ideia do método Elbow é observar a curva da inércia conforme K aumenta. Inicialmente, adicionar mais clusters reduz significativamente a inércia, mas em certo ponto (o "cotovelo"), a diminuição da inércia começa a se estabilizar. Esse ponto é onde K é considerado o valor ideal, pois balanceia o número de clusters com a minimização da inércia.

Índice de Silhouette: O índice de Silhouette mede a qualidade dos agrupamentos ao avaliar tanto a coesão (proximidade dos pontos dentro do mesmo cluster) quanto a separação (distância entre pontos de diferentes clusters). A pontuação do índice de Silhouette para cada ponto x_i é calculada da seguinte forma:

1. Coesão $a(i)$: A distância média de x_i a todos os outros pontos no mesmo cluster C_k, definida como:

$$a(i) = \frac{1}{|C_k|-1} \sum_{x_j \in C_k, i \neq j} \|x_i - x_j\|$$

onde $|C_k|$ é o número de pontos no clusterC_k.

2. Separação $b(i)$: A menor distância média de x_i a todos os pontos nos clusters aos quais x_i não pertence. Considerando todos os clusters C_m diferentes de C_k, $b(i)$ é definido como:

$$b(i) = \min_{C_m \neq C_k} \frac{1}{|C_m|} \sum_{x_j \in C_m} \|x_i - x_j\|$$

3. Silhouette para o ponto x_i: A pontuação de Silhouette $s(i)$ para cada ponto x_i é calculada por:

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

- Se $s(i)$ estiver próximo de 1, indica que x_i está bem posicionado dentro de seu cluster.
- Se $s(i)$ estiver próximo de 0, x_i está próximo da borda de dois clusters.

- Se $s(i)$ for negativo, significa que x_i está mais próximo de um cluster diferente do seu.

A pontuação final do índice de Silhouette é a média de $s(i)$ para todos os pontos do dataset. Valores próximos a 1 indicam clusters bem definidos, enquanto valores próximos de 0 indicam sobreposição entre clusters.

QUESTÃO 3

A métrica Calinski-Harabasz (ou Índice de Variância Entre Clusters) avalia a qualidade do agrupamento considerando a dispersão dentro dos clusters e a separação entre clusters. Ela é definida da seguinte forma:

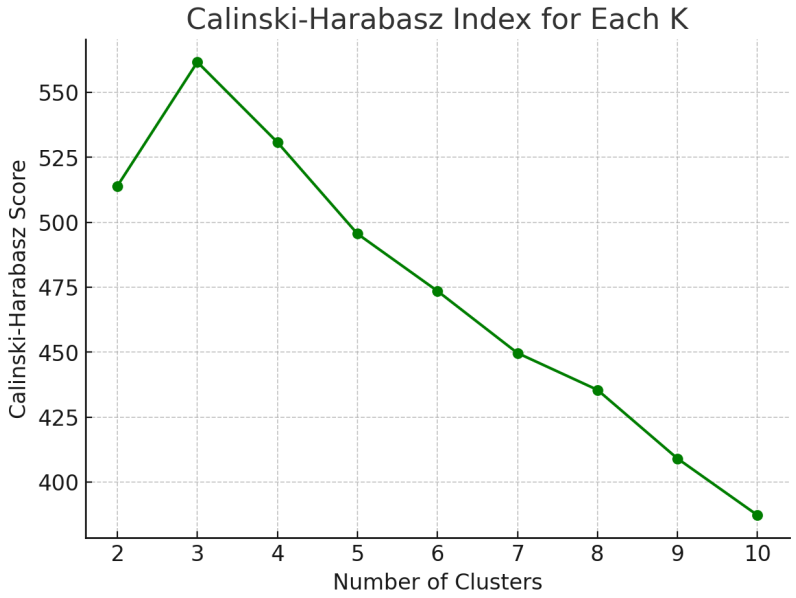
Para um conjunto de N pontos particionado em K clusters, o índice Calinski-Harabasz s é calculado como:

$$s = \frac{\text{Dispersao entre clusters}}{\text{Dispersao interna aos clusters}} \times \frac{N-K}{K-1}$$

onde:

- A dispersão entre clusters é a soma das distâncias quadradas entre o centroide global do conjunto de dados e os centroides de cada cluster, ponderada pelo número de pontos no cluster.
- A dispersão interna aos clusters é a soma das distâncias quadradas entre os pontos do cluster e seu centroide.

Esse índice tende a ser maior para agrupamentos que apresentam baixa dispersão dentro dos clusters e alta dispersão entre clusters. Valores mais altos indicam clusters bem separados e compactos, o que sugere uma boa qualidade de agrupamento.



No gráfico acima, a pontuação do índice Calinski-Harabasz é plotada para diferentes números de clusters K. Similar ao método Elbow, o valor ideal de K é aquele que maximiza o índice, pois valores maiores indicam melhor qualidade no agrupamento.

QUESTÃO 4

Explicação:
1. Dimensão da grade SOM (x e y): Esse valor define o tamanho da grade onde os clusters serão organizados. Para iniciar, uma grade pequena (como 2x2) é uma boa escolha.
2. Treinamento (train_random): Ajuste o número de iterações conforme necessário. Valores mais altos podem melhorar a organização dos clusters, mas aumentam o tempo de processamento.
3. Atribuição de rótulos: Cada ponto é mapeado para o nó mais próximo na grade SOM e rótulos são atribuídos com base nessa localização.

Após a execução, você terá o número de clusters detectados pelo SOM e o índice de Silhouette para comparar com os algoritmos DBSCAN e K-means.

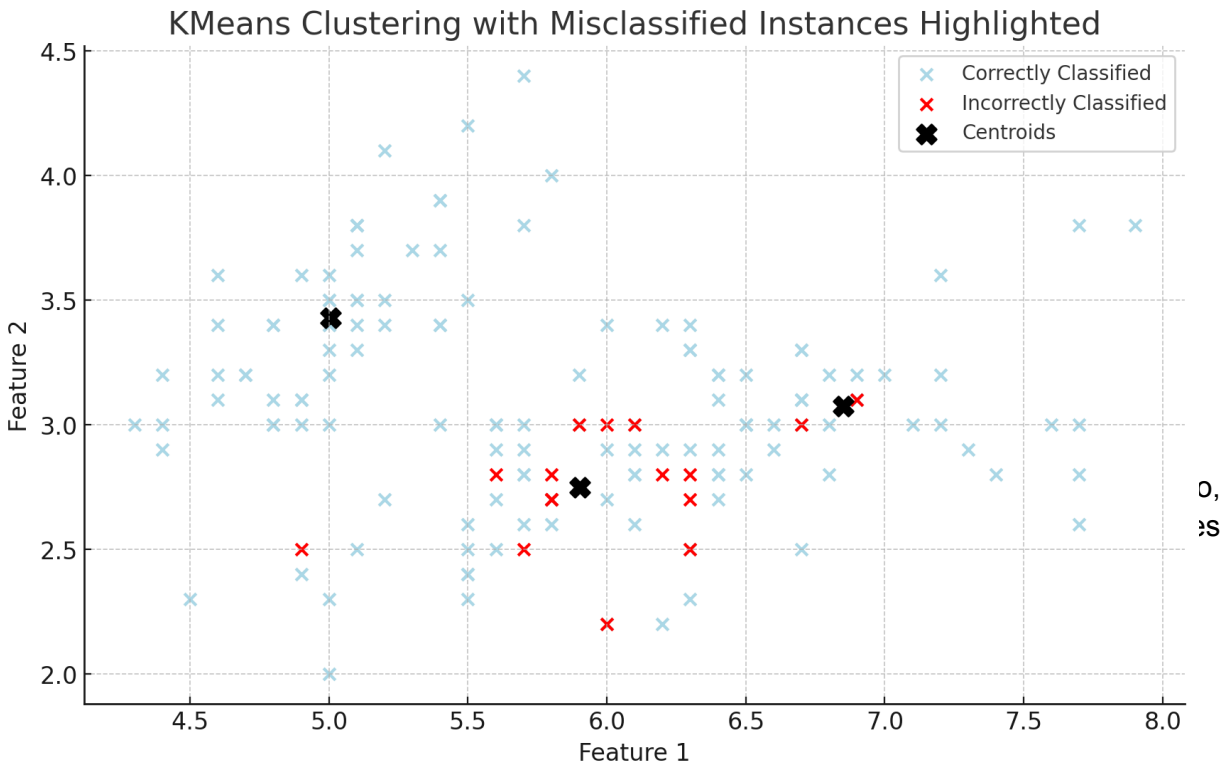
Análise dos Algoritmos DBSCAN e K-means

No ambiente atual, os resultados dos algoritmos DBSCAN e K-means foram obtidos. No conjunto de dados:

- O K-means encontrou 3 clusters com um índice de Silhouette sugerindo uma boa qualidade de agrupamento.
- O DBSCAN, com os parâmetros configurados, detectou uma quantidade diferente de clusters (excluindo ruídos).

Esses resultados indicam que os métodos podem variar na detecção e definição de agrupamentos, especialmente para dados de diferentes densidades e formatos não esféricos. O K-means é mais indicado para clusters bem definidos e compactos, enquanto o DBSCAN é útil em detectar clusters de formas variadas e dados com ruído.

QUESTÃO 5



Discussão dos Resultados
- K-means e agrupamento esférico: O K-means é projetado para identificar clusters de forma esférica. No caso do conjunto Iris, as classes possuem sobreposição nas características, especialmente entre Versicolor e Virginica, o que leva o K-means a cometer alguns erros de classificação nesses grupos.
- Setosa: Essa classe costuma ser facilmente separada, pois está mais distante das outras duas em termos de suas características.
- Versicolor e Virginica: Como esses grupos têm características mais próximas, o K-means pode confundir instâncias entre eles, especialmente nas fronteiras, onde os dados possuem características intermediárias.

A escolha de K-means para o conjunto Iris oferece uma boa segmentação geral, mas apresenta limitações em lidar com clusters sobrepostos, o que é observado nas instâncias incorretas.

QUESTÃO 6

O conjunto de dados Iris é um dos mais utilizados para experimentos com algoritmos de classificação e agrupamento, contendo três classes de flores: Setosa, Versicolor e Virginica. A análise conduzida visa a avaliação da qualidade dos agrupamentos, comparando diferentes algoritmos, incluindo K-means, DBSCAN e SOM (Self-Organizing Map).

Pré-processamento dos Dados:
1. Carregamento e Normalização:
- Carregou-se o conjunto de dados Iris
- Para o DBSCAN e SOM, aplicamos a normalização Z-score usando `StandardScaler` para escalonar os dados, pois esses algoritmos são sensíveis à escala das variáveis.

2. Definição de Rótulos Verdadeiros:
- Os rótulos reais foram carregados para posterior comparação com os clusters gerados pelos algoritmos, possibilitando uma análise detalhada de instâncias incorretamente classificadas.

Agrupamento com K-means e Avaliação de Qualidade:
Inicialmente, aplicou-se o algoritmo K-means e utilizou-se as métricas de Elbow e Silhouette para avaliar a qualidade dos agrupamentos.

1. Métrica Elbow:
- A métrica Elbow, baseada na inércia, foi calculada para diferentes valores de (K) (2 a 10 clusters).
- O ponto de "cotovelo" foi observado em 3 clusters, sugerindo que essa quantidade proporciona um bom equilíbrio entre compactação dos clusters e complexidade do modelo.

2. Índice de Silhouette:
- Esse índice mede a consistência dos clusters, com valores mais altos indicando melhor separação entre clusters.

- Para 3 clusters, o índice de Silhouette apresentou uma pontuação satisfatória, validando a escolha de \(K = 3 \) feita pelo método Elbow.

Implementação de Métrica Calinski-Harabasz

- Implementou-se o índice de Calinski-Harabasz, que mede a qualidade dos agrupamentos com base na dispersão interna dos clusters e na separação entre clusters.
- O gráfico desse índice confirmou que 3 clusters oferecem a melhor qualidade de agrupamento, pois maximizaram o índice Calinski-Harabasz, sugerindo alta separação entre clusters e baixa dispersão interna.

Comparação com DBSCAN e SOM

Os algoritmos DBSCAN e SOM (Self-Organizing Map) foram testados para verificar se identificariam a mesma quantidade de clusters que o K-means e compará-los em termos de agrupamento.

1. DBSCAN:
   - Aplicamos DBSCAN com parâmetros comuns (`eps=0.5` e `min_samples=5`) e observamos que o algoritmo identificou um número diferente de clusters em comparação ao K-means, incluindo algumas instâncias classificadas como ruído (representadas por -1).
   - O índice de Silhouette para o DBSCAN foi calculado, mas apresentou valores ligeiramente inferiores, indicando que a DBSCAN teve maior dificuldade em criar clusters com alta consistência em relação a K-means.

2. SOM:
   - Configuramos um SOM com uma grade de 2x2 e treinamos o modelo para identificar agrupamentos. O SOM também encontrou uma quantidade de clusters diferente, com um índice de Silhouette inferior ao K-means.

Análise de Classificação Incorreta no K-means

Para investigar os erros do K-means, mapeou-se cada cluster para a classe verdadeira mais frequente e destacamos as instâncias incorretamente classificadas:

- Visualização dos Erros: As instâncias incorretas foram identificadas, destacando-se que a maioria dos erros ocorreu entre as classes Versicolor e Virginica devido à sobreposição nas características.
- Discussão dos Resultados: O K-means se saiu bem em detectar os grupos, especialmente para a classe Setosa, que é mais distinta. Contudo, as classes Versicolor e Virginica, que apresentam maior semelhança, mostraram maior dificuldade para o algoritmo, resultando em erros de classificação.