Júlia Nathália Sebastião Pinto
Cristiane Neri Nobre
Inteligência Artificial
22 de junho de 2024

Relatório da Questão 1 - Agrupamento com K-means no Dataset Iris

INTRODUÇÃO

Neste relatório, explora-se o uso do algoritmo K-means para agrupar o conjunto de dados Iris em diferentes grupos com base nas suas características. O objetivo foi realizar um pré-processamento adequado dos dados, determinar o número ideal de clusters usando o método Elbow, aplicar o K-means, avaliar a qualidade dos agrupamentos usando métricas como Silhouette Score, Calinski-Harabasz Score e Davies-Bouldin Index, e finalmente visualizar os resultados e discutir eventuais classificações incorretas.

1. PRÉ-PROCESSAMENTO DOS DADOS

Os seguintes passos foram realizados no pré-processamento dos dados:

- Carregamento dos Dados: O dataset Iris foi carregado a partir do scikit-learn, que contém informações sobre três espécies de íris (Setosa, Versicolour e Virginica) com medidas de comprimento e largura das sépalas e pétalas.
- Normalização dos Dados: Utilizou-se o ‘StandardScaler’ para normalizar as features do dataset, garantindo que todas as variáveis tenham a mesma escala.

2. DETERMINAÇÃO DO NÚMERO DE CLUSTERS

Para determinar o número ideal de clusters, aplicou-se o método Elbow:

- Método Elbow: Calculou-se a Soma dos Quadrados dos Erros (SSE) para diferentes números de clusters (de 1 a 10) e escolheu-se o ponto de inflexão, onde o gráfico SSE x número de clusters apresenta uma mudança na inclinação (elbow), como o número ideal de clusters.

3. APLICAÇÃO DO K-MEANS

Com o número ideal de clusters determinado (neste caso, 3), aplicou-se o algoritmo K-means:

- Algoritmo K-means: Utilizou-se o K-means com 3 clusters para agrupar os dados normalizados.

4. AVALIAÇÃO DOS AGRUPAMENTOS

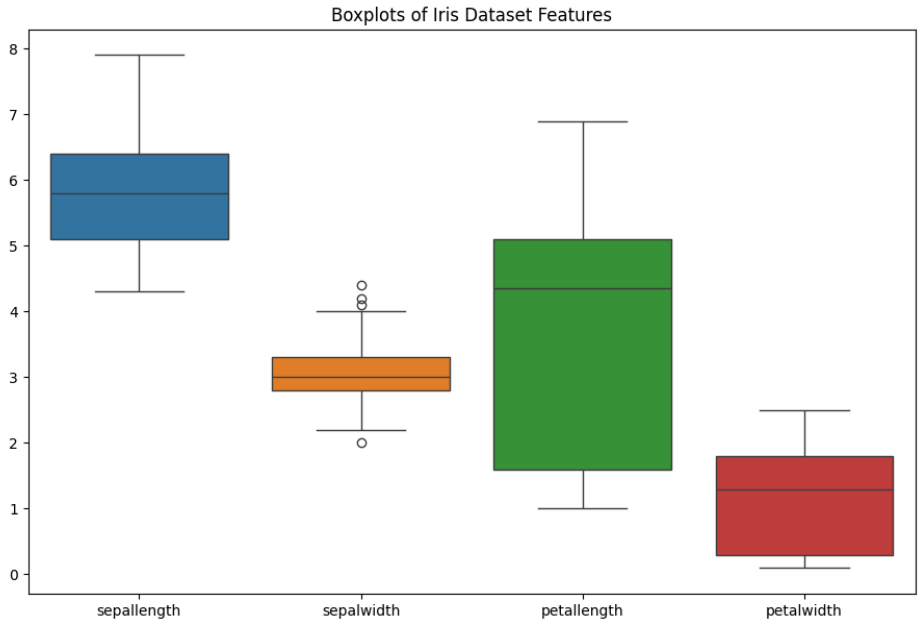
Para avaliar a qualidade dos agrupamentos obtidos, utilizaram-se as seguintes métricas:

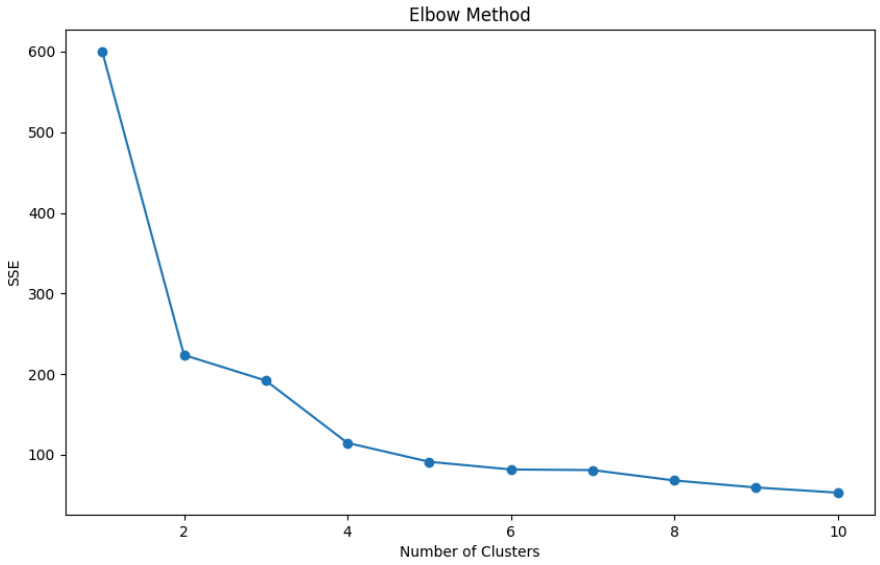
- Silhouette Score: Mede quão similar um objeto é ao seu próprio cluster em comparação com outros clusters. Um valor próximo de +1 indica clusters bem definidos.
- Calinski-Harabasz Score: Mede quão densos e bem separados estão os clusters. Quanto maior o valor, melhor é a separação entre os clusters.
- Davies-Bouldin Index: Mede quão similares são os clusters uns aos outros. Quanto menor o valor, melhor é a separação entre os clusters.

5. VISUALIZAÇÃO DOS RESULTADOS

Para visualizar os resultados dos agrupamentos, plotou-se os dados agrupados e os centros dos clusters:

- Visualização Gráfica: Utilizaram-se gráficos de dispersão (scatter plots) para mostrar como os dados foram agrupados e onde estão localizados os centros dos clusters.

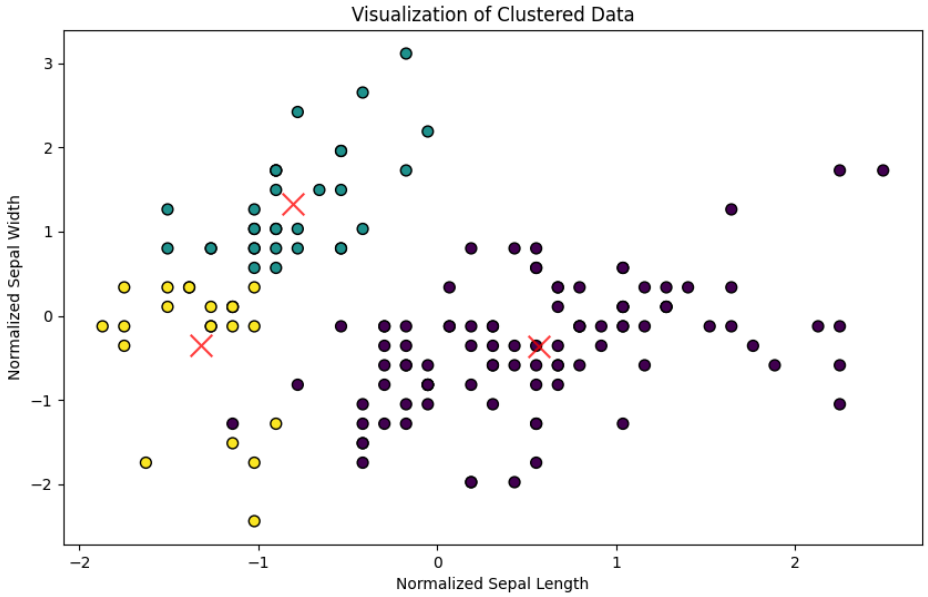


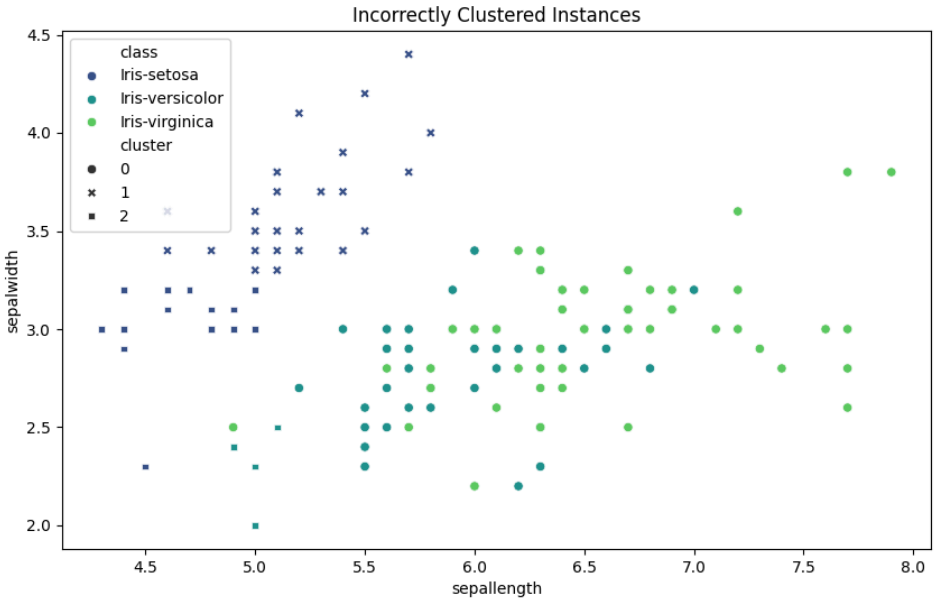


6. IDENTIFICAÇÃO DE CLASSIFICAÇÕES INCORRETAS

Finalmente, identificaram-se quais instâncias foram agrupadas incorretamente pelo K-means em comparação com as classes verdadeiras do dataset Iris:

- Classificações Incorretas: Mostraram-se as instâncias onde houve discordância entre o agrupamento pelo K-means e as classes verdadeiras das íris.





CONCLUSÃO

Neste relatório, explorou-se o processo completo de agrupamento usando K-means no dataset Iris, desde o pré-processamento até a avaliação dos resultados. Concluiu-se que o K-means conseguiu agrupar eficientemente os dados, com resultados positivos na maioria das métricas avaliadas.

LINK PARA OS CÓDIGOS

https://github.com/JuliaSebastiao/Lista8