CLUSTERIZAÇÃO K-MEANS

TRABALHO PRÁTICO 2
INTELIGÊNCIA ARTIFICIAL

Maria Eduarda Ferreira da Silva
Vitória Christie Amaral Santos

OBJETIVOS DO PROJETO

01

IMPLEMENTAR O ALGORITMO K-MEANS DO ZERO EM PYTHON.

02

APLICAR O ALGORITMO NA BASE DE DADOS IRIS, DESCONSIDERANDO AS CLASSES ORIGINAIS.

03

AVALIAR A QUALIDADE DOS CLUSTERS PARA K=3 E K=5 UTILIZANDO O SILHOUETTE SCORE.

04

COMPARAR OS RESULTADOS COM A IMPLEMENTAÇÃO DA BIBLIOTECA SCIKIT-LEARN.

05

UTILIZAR A TÉCNICA DE PCA PARA REDUZIR A DIMENSIONALIDADE E VISUALIZAR OS CLUSTERS.

O QUE É O K - MEANS?

Algoritmo de aprendizado não supervisionado para clusterização de dados.

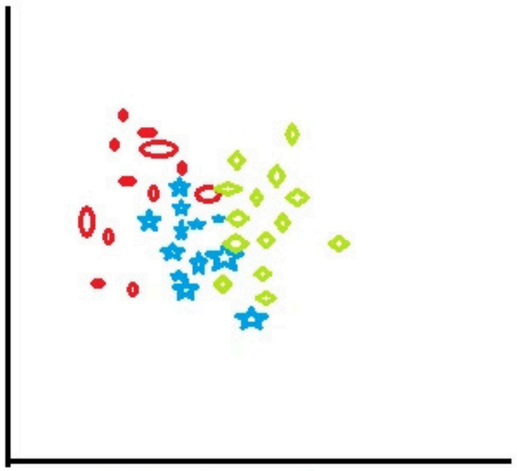


fig 1: before applying k-means clustering

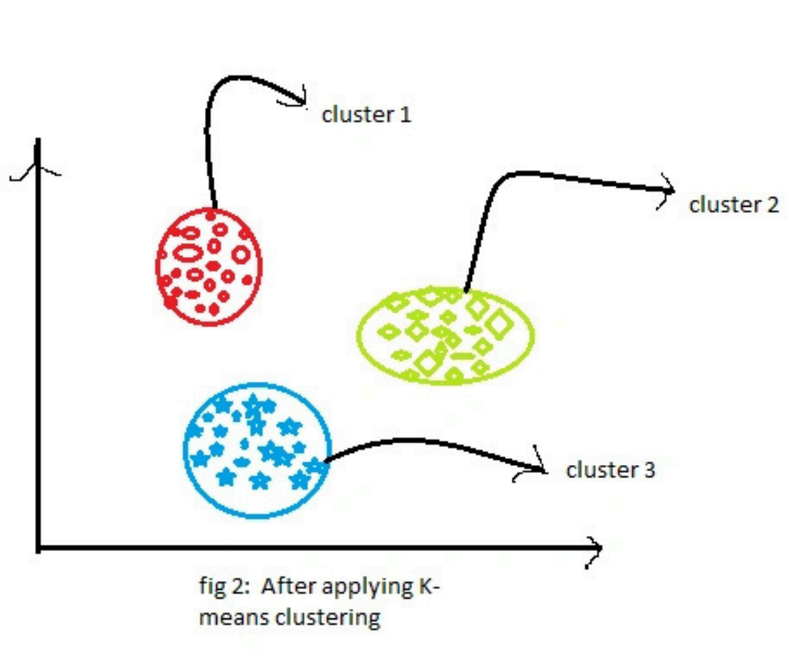


fig 2: After applying K-means clustering

01

OBJETIVO

Agrupar dados em 'k' clusters com base na similaridade.

02

PROCESSO ITERATIVO

- Associa cada ponto ao centróide (centro do cluster) mais próximo.
- Recalcula o centróide como a média dos pontos do cluster.

IMPLEMENTAÇÃO HARDCORE

Desenvolvido em Python utilizando NumPy para cálculos matemáticos.

Critério de Parada: Convergência dos centróides (quando não há mais mudanças).

FUNÇÕES PRINCIPAIS

01

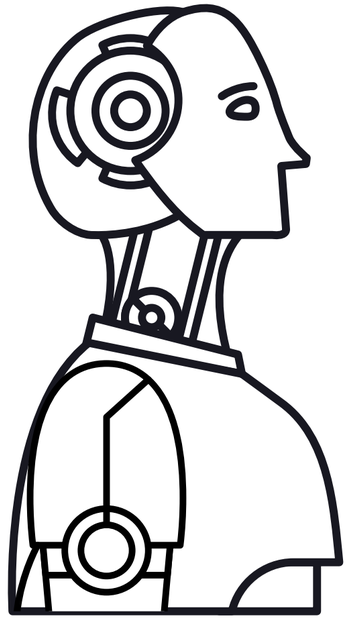
`inicializar_centroides`
Escolha aleatória de 'k' pontos iniciais.

02

`atribuir_clusters`
Cálculo da distância Euclidiana para associação.

01

`atualizar_centroides`
Cálculo da média para novos centrórides.



RESULTADOS

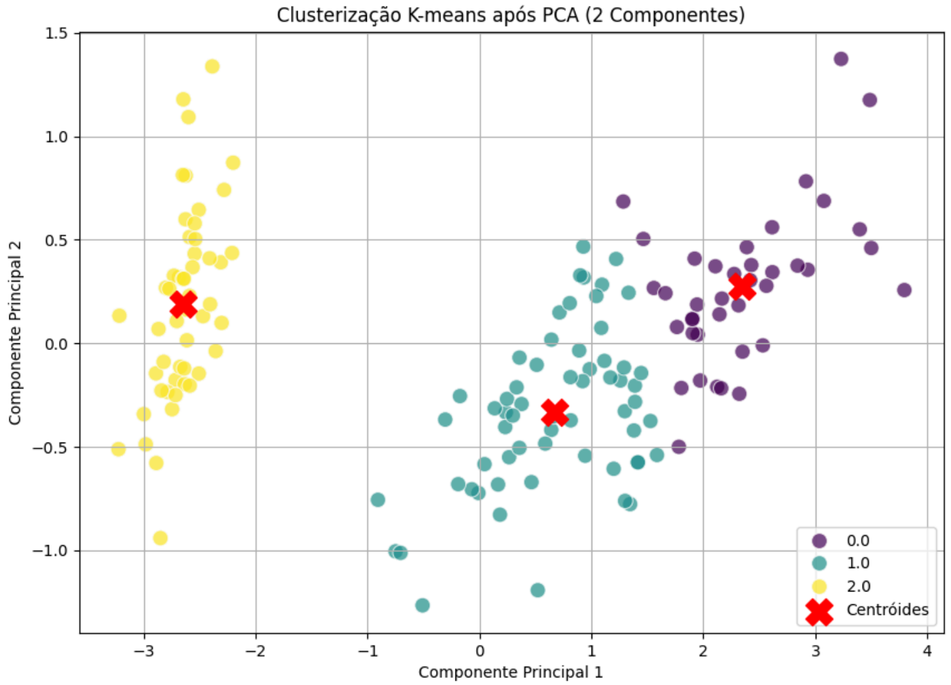
Métrica de Avaliação: Silhouette Score
(quanto mais perto de 1, melhor).

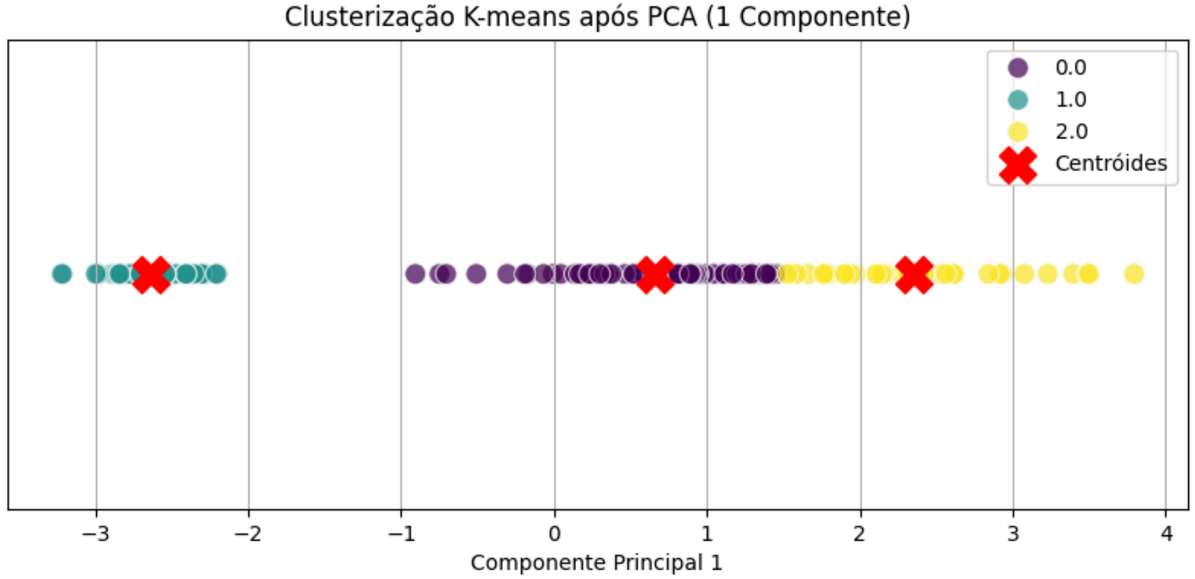
```
Resultado da implementação Hardcore para K=3:
Silhouette Score: 0.5528
Resultado da implementação Hardcore para K=5:
Silhouette Score: 0.4922
```

Conclusão: k=3 apresentou a melhor formação de clusters para a base Iris.

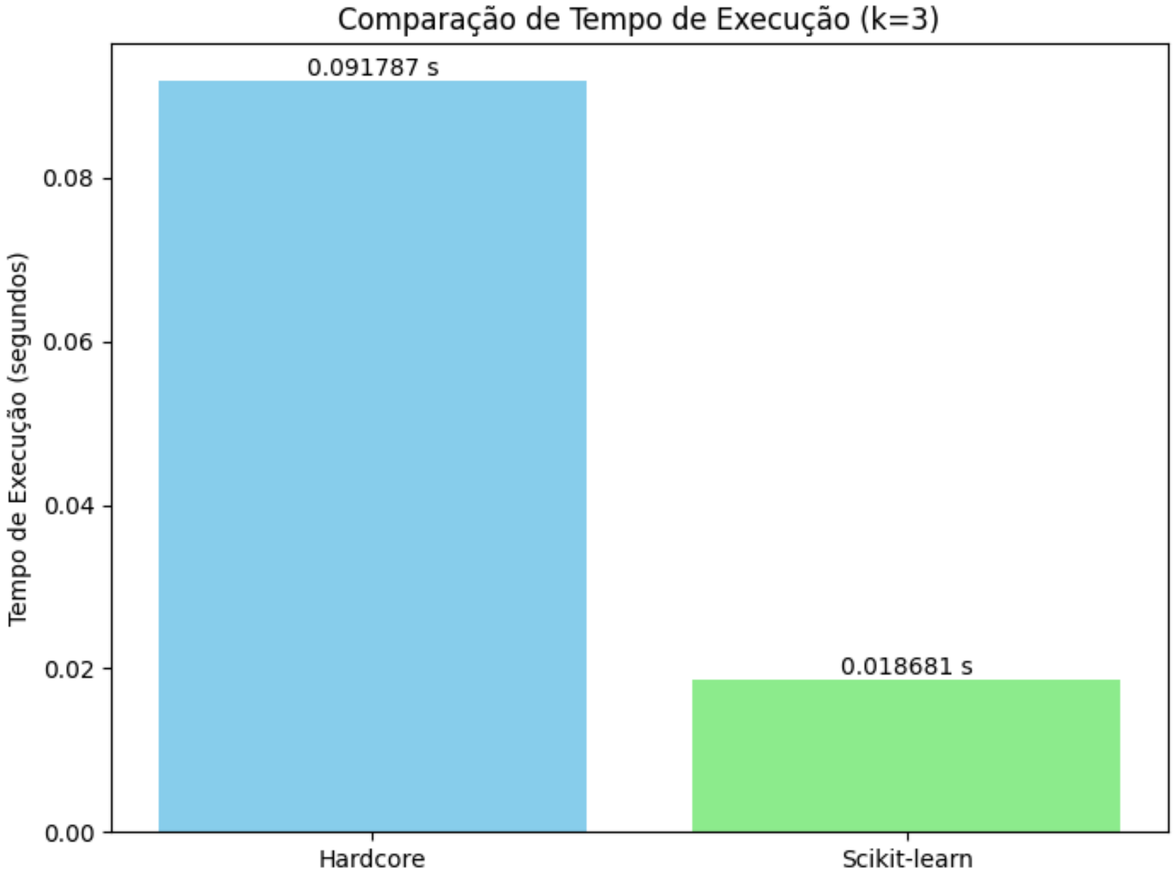
VISUALIZAÇÃO COM PCA

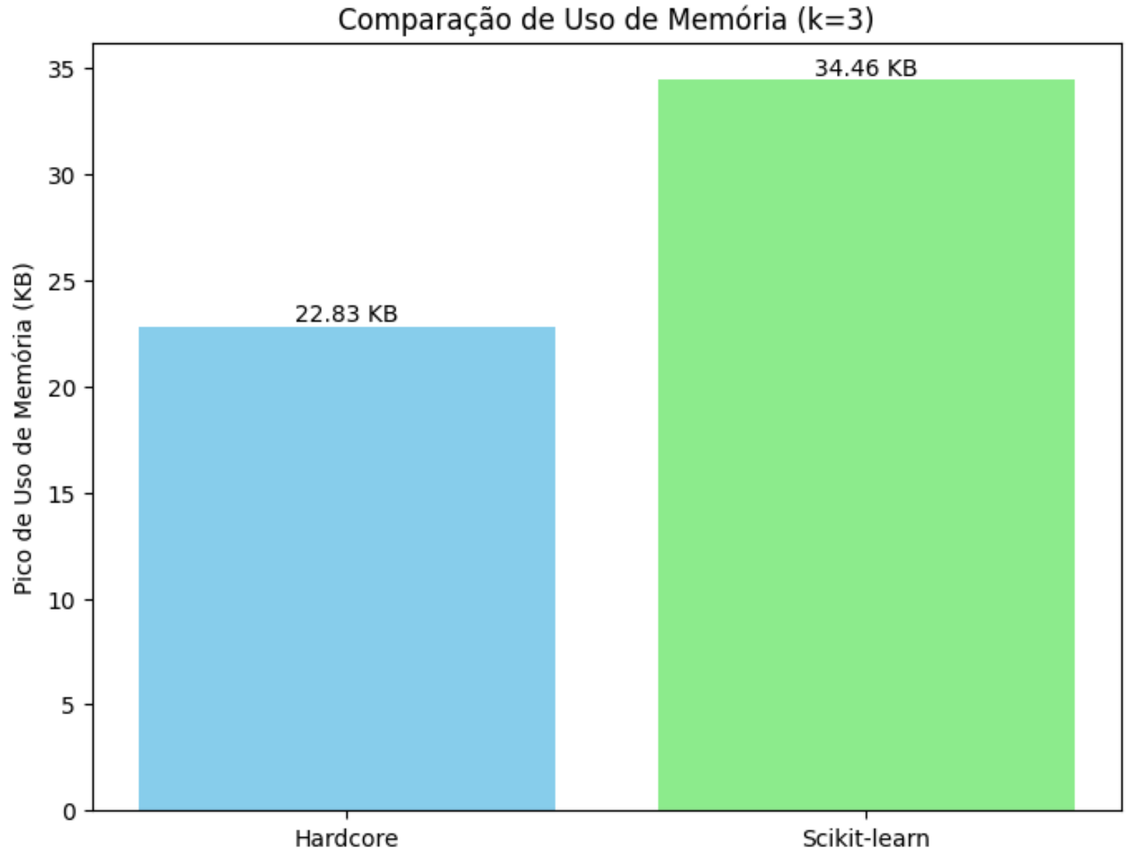
- A base Iris possui 4 dimensões (features), impossibilitando a visualização direta.
- Usamos PCA para reduzir os dados para 2 dimensões, preservando o máximo de informação.



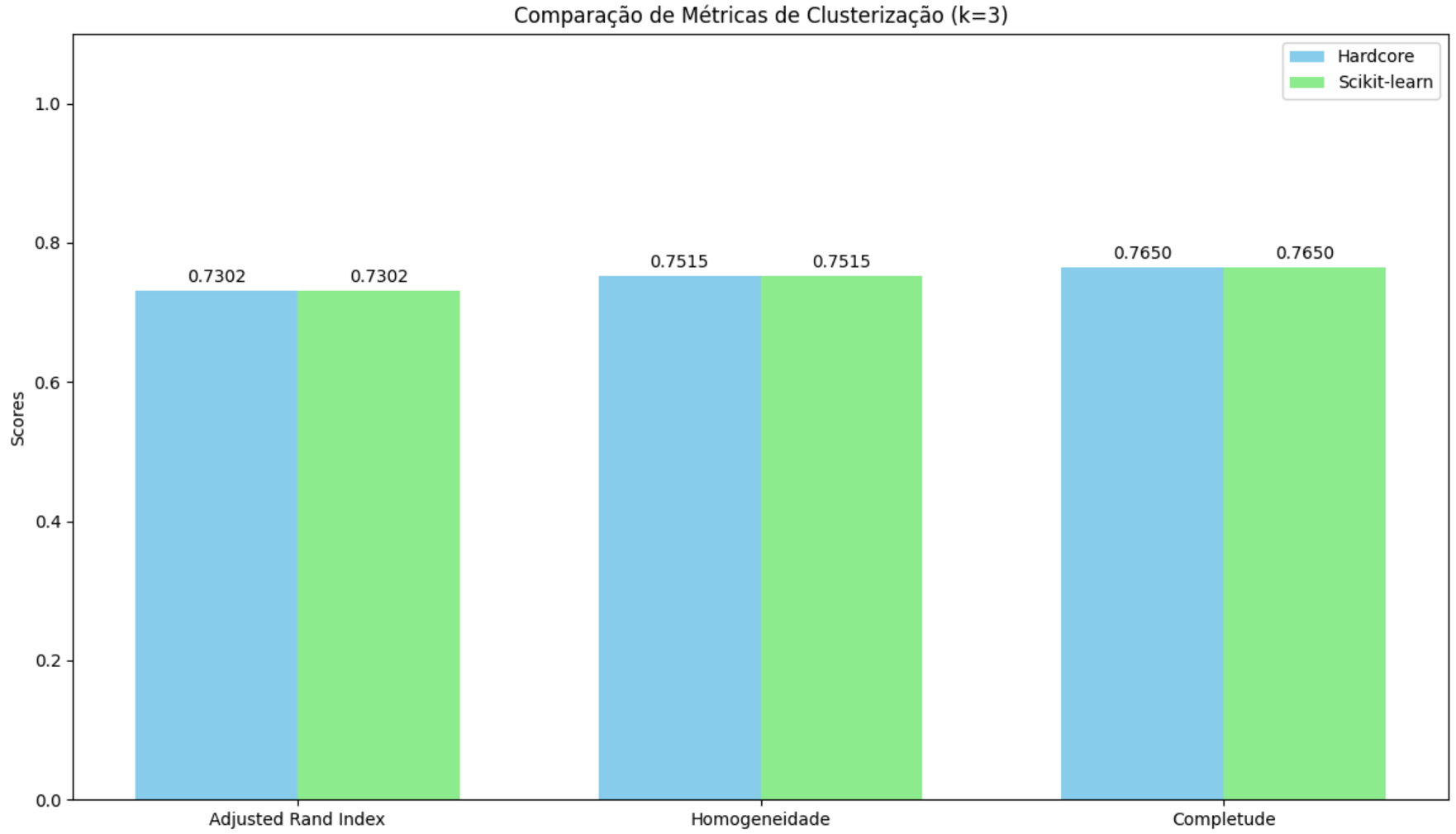


COMPARAÇÃO COM SCIKIT-LEARN





COMPARAÇÃO COM SCIKIT-LEARN



CONCLUSÕES

- A implementação do K-means do zero foi bem-sucedida e produziu resultados consistentes.
- O Silhouette Score foi eficaz para determinar o número ideal de clusters.
- O PCA provou ser uma ferramenta indispensável para a visualização e interpretação de dados multidimensionais.
- A comparação de desempenho ressaltou a importância das otimizações em bibliotecas padrão.

LINK DA APRESENTAÇÃO

https://youtu.be/YTNUruuPM48