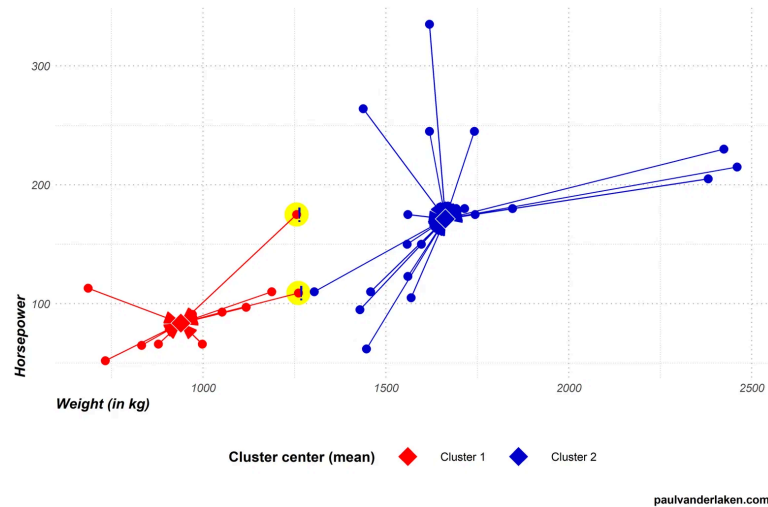


Trabalho Prático 02 - Clusterização K-means

Comparação de Implementações: Hardcode vs Scikit-learn



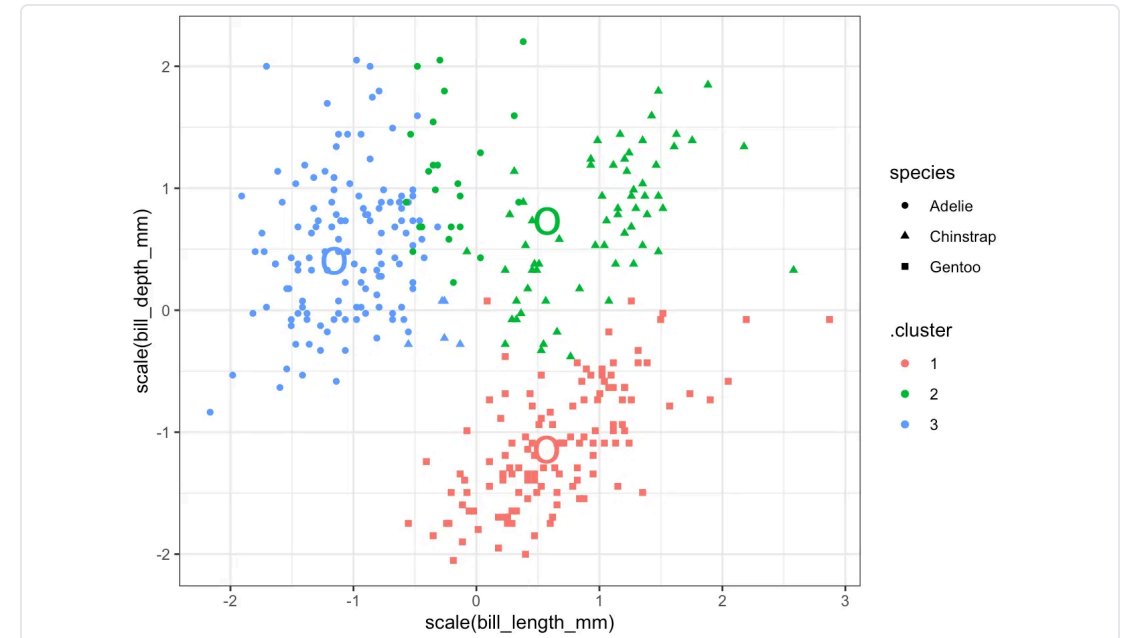
GCC128 - Inteligência Artificial
Sistemas de Informação – 14A

Ahmed Ali Abdalla Esmin
Anna Paula Figueiredo






Lavras - MG

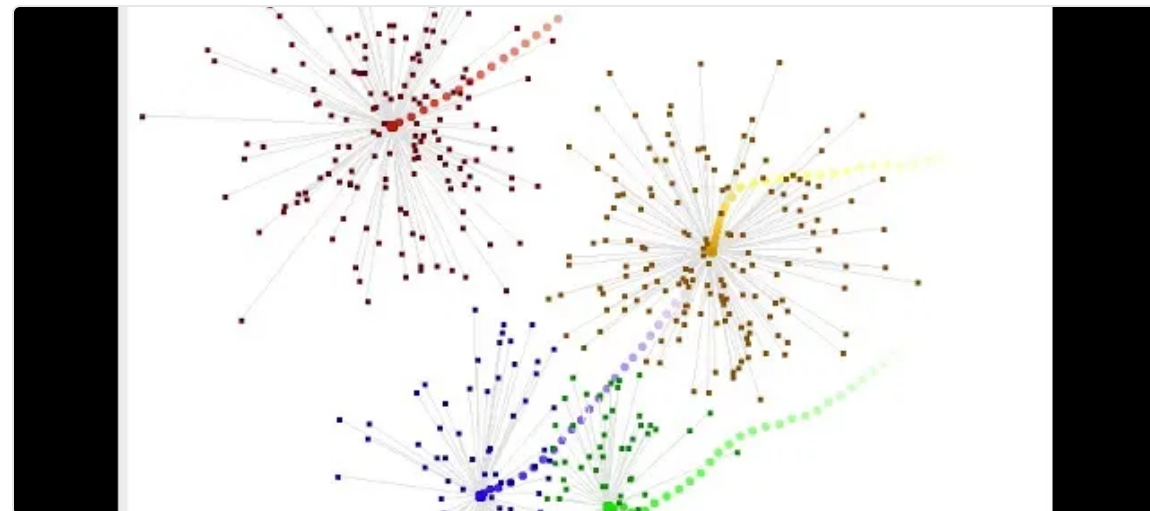
Introdução à Clusterização K-means

- ▶ **K-means** é um algoritmo de clusterização não supervisionada, que agrupa dados em K clusters distintos.
- ▶ O objetivo é particionar N observações em K clusters, onde cada observação pertence ao cluster com a média mais próxima.
- ▶ Aplicável em problemas de segmentação de clientes, compressão de imagens, análise de dados biológicos, entre outros.
- ▶ **Objetivo do trabalho:** Fixar conhecimentos desenvolvendo o algoritmo K-means do zero (hardcore) e comparando com uma implementação de biblioteca.



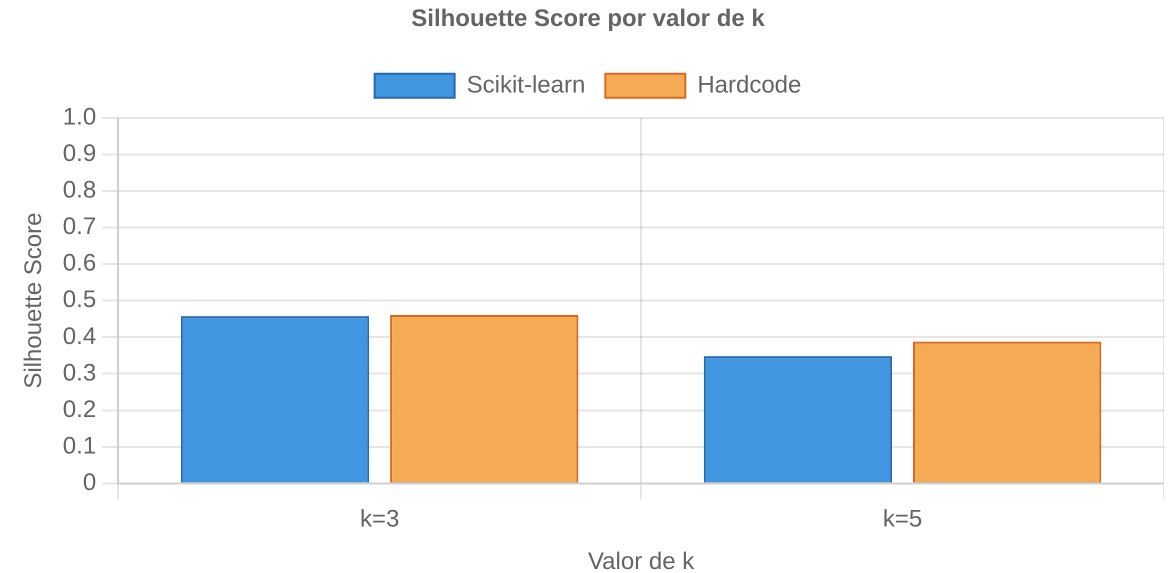
Metodologia de Comparação

-  **Dataset:** Iris (150 amostras), desconsiderando a classe alvo para clusterização.
-  **Implementação Hardcode:** Desenvolvida do zero em Python, usando bibliotecas básicas.
-  **Implementação Scikit-learn:** Utiliza a versão otimizada da biblioteca, com múltiplas inicializações (`n_init=20`).
-  **Valores de k testados:** Experimentos para $k=3$ e $k=5$.
-  **Métricas de Avaliação:** Inércia, Silhouette Score, Número de Iterações, Tempo de Execução e Tamanhos de Clusters.



Avaliação e Análise de Desempenho

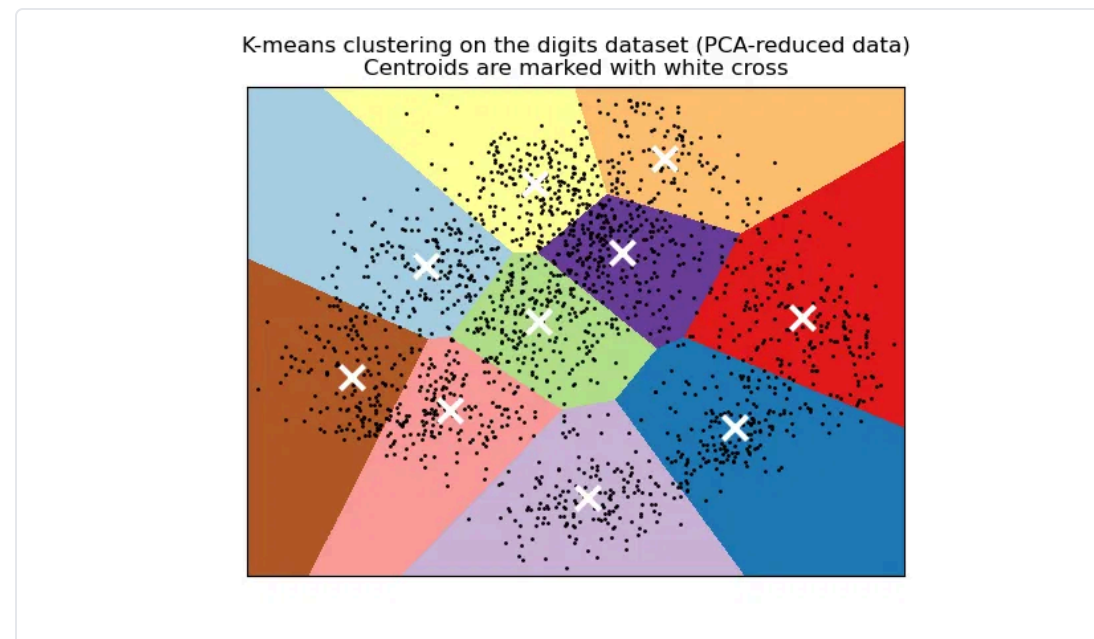
- Resultados para k=3:** Hardcore: inércia ≈ 140.0 , silhouette ≈ 0.463 . Scikit-learn: inércia ≈ 139.8 , silhouette ≈ 0.460 .
- Resultados para k=5:** Silhouette scores menores (Hardcore ≈ 0.39 , Scikit-learn ≈ 0.35), indicando clusters menos definidos.
- Tempo de Execução:** Ordem de centésimos de segundo para ambos, com ligeira vantagem para Hardcore em k=3.
- Convergência:** Scikit-learn convergiu em menos iterações, demonstrando maior robustez e estabilidade.
- Melhor k:** O valor mais adequado de k é 3, coerente com as três espécies da base Iris.



Comparação de Silhouette Score entre implementações para diferentes valores de k

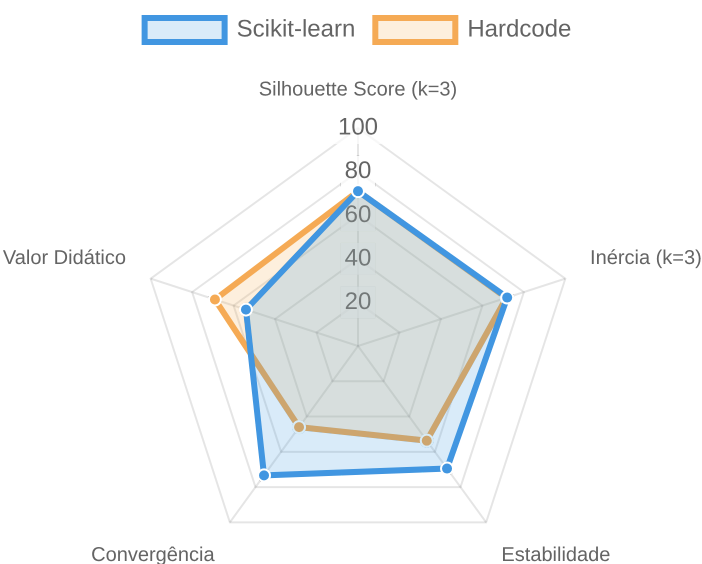
Redução de Dimensionalidade com PCA

- ▶ **Aplicação:** PCA foi aplicado com o melhor $k=3$ para projetar os dados.
- ▶ **1 Componente Principal:** Explicou cerca de 73% da variância, mas com sobreposição entre os clusters.
- ▶ **2 Componentes Principais:** Explicou aproximadamente 96% da variância total.
- ▶ **Visualização:** Evidenciou uma separação mais clara entre os grupos e facilitou a visualização dos centróides.



Comparativo Geral das Implementações

- ▶ **Valor de k:** O valor mais adequado de k é 3, coerente com as três espécies de flores da base Iris.
- ▶ **Métricas Próximas:** Ambas as implementações apresentaram métricas próximas, validando a correteude da versão manual.
- ▶ **Scikit-learn:** Destacou-se por convergir em menos iterações e oferecer maior estabilidade devido às suas otimizações internas e múltiplas inicializações.
- ▶ **Hardcode:** Útil para fins didáticos e de compreensão do algoritmo, apesar de mais simples.



Característica	Hardcode	Scikit-learn
Valor didático	✓ Alto	✓ Médio
Silhouette Score (k=3)	0.463	0.460
Inércia (k=3)	140.0	139.8
Estabilidade	⚠ Menor	✓ Maior
Convergência	⚠ Mais iterações	✓ Menos iterações

Conclusão e Referências

Conclusões

- ✓ O trabalho demonstrou com sucesso a aplicação do K-means em duas abordagens distintas.
- ✓ Conclui-se que o valor ótimo de clusters é 3, e que a versão Scikit-learn apresenta desempenho mais eficiente e consistente.
- ✓ A versão Hardcode reforça o entendimento dos conceitos fundamentais do algoritmo.
- ✓ A análise com PCA confirmou visualmente a separação dos clusters e a adequação da escolha de k.

Referências

- SCIKIT-LEARN. K-means clustering. Disponível em:
<https://scikit-learn.org/stable/modules/clustering.html#k-means>

Link Para o vídeo de apresentação: K-means clustering.

Disponível

em: [https://drive.google.com/file/d/1zKjf7wBMRLCamjXw0fQ5WE6unDyOWdGv/view?](https://drive.google.com/file/d/1zKjf7wBMRLCamjXw0fQ5WE6unDyOWdGv/view?usp=sharing)

[usp=sharing](https://drive.google.com/file/d/1zKjf7wBMRLCamjXw0fQ5WE6unDyOWdGv/view?usp=sharing)