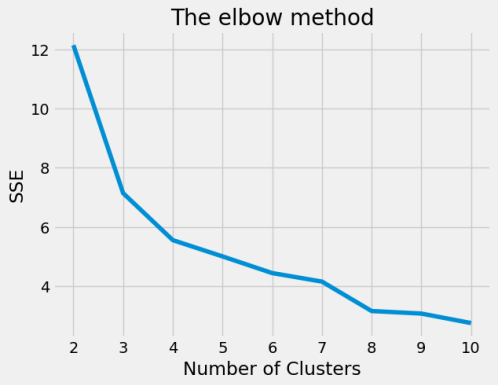
# Lista 6 - IA


## Sophia Carrazza Ventorim de Sousa

### PUC Minas - 2024


Código utilizado:
https://colab.research.google.com/drive/19obePzT0yWBB10avfZ4fF1DFZcOihKLK?usp=sharing

## 1- Elbow Method e Silhouette Score:




```
Silhouette Score k = 2: 0.629
Silhouette Score k = 3: 0.504
Silhouette Score k = 4: 0.444
Silhouette Score k = 5: 0.360
Silhouette Score k = 6: 0.317
Silhouette Score k = 7: 0.323
Silhouette Score k = 8: 0.325
```

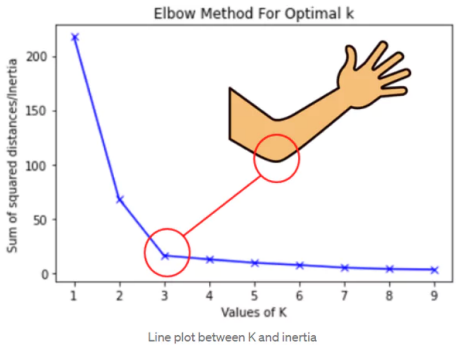
## Conclusões:

A métrica elbow indica que o “Optimal K” destes agrupamentos está em torno de 4 ou 5, quando a curva começa a se estabilizar depois de uma queda (onde forma um cotovelo). Já o silhouette mostra que há uma maior coesão e separação com 2 a 3 clusters.

Assim, para equilibrar uma boa separação, coesão e menor overfitting, 3 ou 4 clusteres seriam o ideal (entre 2 e 5).

## 2- SSE vs Silhouette Score:

O Método do Cotovelo, ou SSE (Soma dos Erros Quadráticos), se dá pela soma das distâncias quadradas entre cada ponto e o centróide de seu cluster. Ele ajuda a encontrar o ponto ótimo onde o SSE se estabiliza com o aumento de clusters.



Line plot between K and inertia

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- k é o número de clusters,
- Ci representa o conjunto de pontos no cluster i
- x é um ponto de dado dentro do cluster
- μi é o centróide do cluster Ci
- ∥x−μi∥^2 é a distância entre o ponto x e o centróide μi.

O Silhouette Score é calculado com a média da fórmula a seguir para todos os pontos, e ele mede o quão bem cada instância de dados se encaixa em seu próprio cluster em comparação com os seus vizinhos, indicando a qualidade do agrupamento em uma escala de -1 a 1.

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

- b(x) e a(x) são as distâncias médias entre x e todos os outros pontos no mesmo cluster (**a** é a coesão e **b** é a separação).

## 3- Índice de Davies-Bouldin:

A nova métrica escolhida foi o Índice de Davies-Bouldin, muito usado para definir qualidade de agrupamentos. Ele mede a qualidade a partir da média das similaridades entre cada cluster e o cluster mais parecido com ele. Quão menor o valor do índice, melhor agrupado estarão os clusteres

**Resultados:**

Índice de Davies-Bouldin para 5 clusters: 0.722

Índice de Davies-Bouldin para 4 clusters: 0.438 (índice baixo, ou seja, boa qualidade de agrupamento)

Índice de Davies-Bouldin para 3 clusters: 0.626

Índice de Davies-Bouldin para 2 clusters: 0.676

## 4- DBSCAM e SOM:

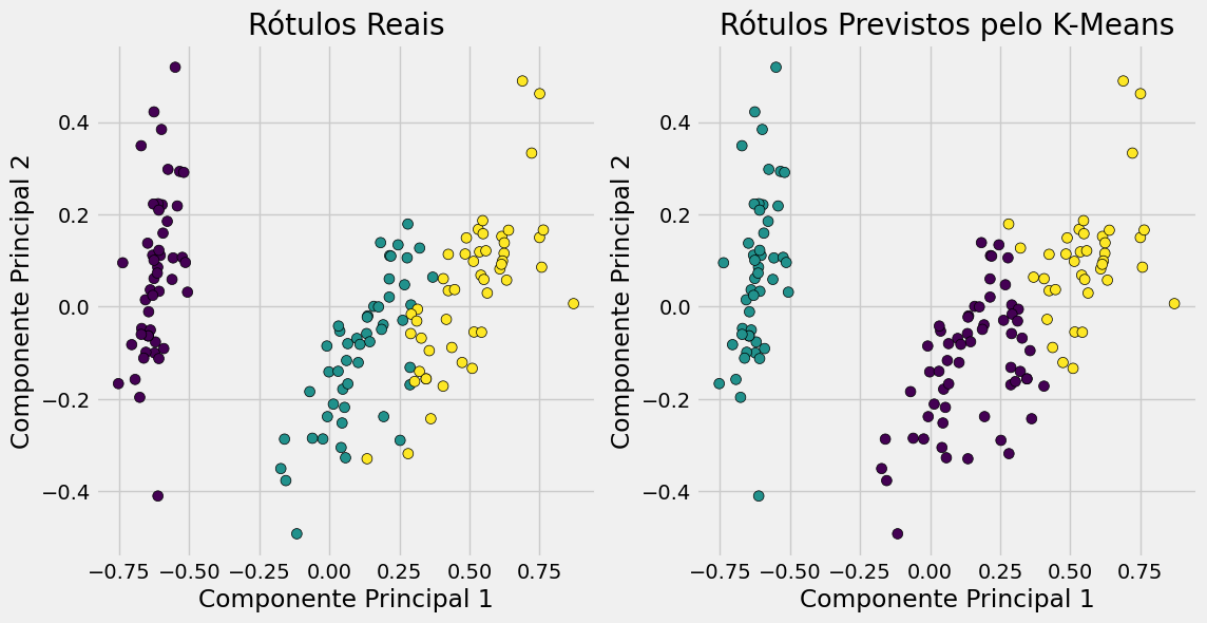
- **K-Means**: Encontrou 4 clusters, como esperado pelo raciocínio já feito;
- **DBSCAN**: Foi encontrado somente 1 cluster (mas pode ter encontrado uma outra quantidade, devido ao ruído - se os dados fossem mais bem separados, talvez desse um número mais próximo ao do K-means);
- **SOM**: Também encontrou 4 clusters (com uma grade de 2×2).

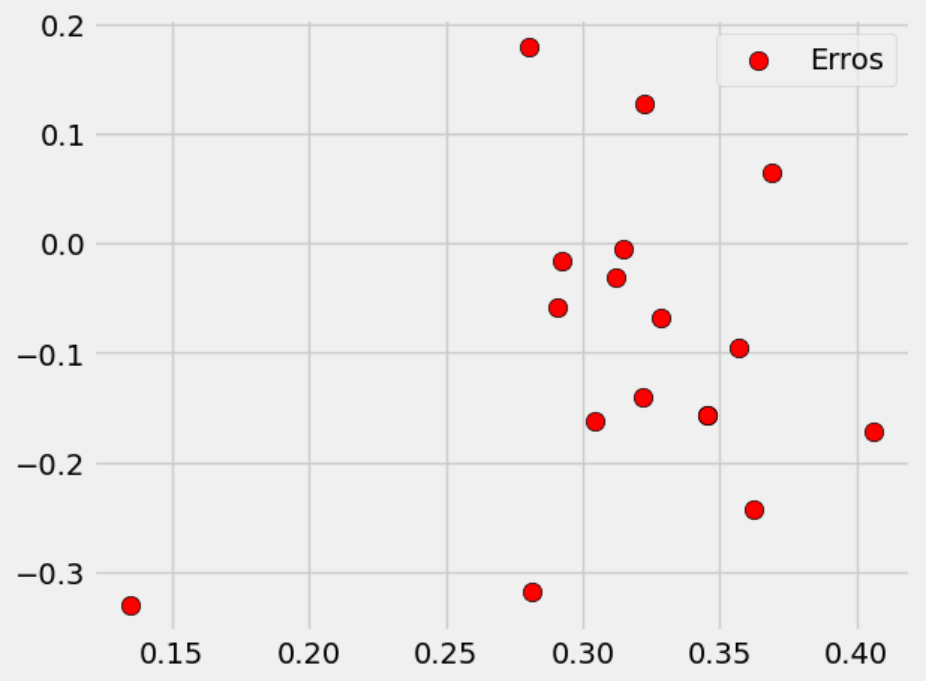
## 5- Verificação de instâncias agrupadas incorretamente:

Para verificar as instâncias agrupadas incorretamente, comparei os rótulos previstos pelo K-Means com os rótulos reais das classes.

**Resultados:**

Acurácia do K-Means em relação aos rótulos reais: 88.67%






## 6- Relatório:

**Pré-processamento:**

- Os dados foram carregados de Iris.csv, com 150 amostras e 4 características;
- Para garantir a mesma escala de todas as características, o MinMaxScaler foi usado. Ele transforma os valores para um intervalo entre 0 e 1;
- Para encontrar o número ideal de clusters, foram utilizadas 3 métricas: Elbow Method, o qual indicou 4 a 5 clusteres, Silhouette Score, que indicou 2 a 3 clusteres, e Índice de Davies-Bouldin, que indicou 4 clusteres.

**Algoritmo K-means:**

- Com  3 clusteres e random_state=0, aplicamos o algoritmo K-Means;
- A acurácia foi calculada comparando os rótulos previstos pelo K-Means aos rótulos reais.

**Resultados:**

- O K-Means recebeu uma acurácia alta (88%) ao agrupar as instâncias de acordo com suas classes reais, especialmente para a classe setosa, embora tenha errado para algumas instâncias da virginica e da versicolor;