Lista 6

código:https://github.com/beatriz-fulgencio/AI/tree/main/lista6

1.

a. **Silhouette**: varia de -1 a 1, sendo que um coeficiente maior indica que os clusters estão mais separados.

   No código:

```
Silhouette Score k = 2: 0.629
Silhouette Score k = 3: 0.504
Silhouette Score k = 4: 0.445
Silhouette Score k = 5: 0.355
Silhouette Score k = 6: 0.348
Silhouette Score k = 7: 0.344
Silhouette Score k = 8: 0.327
```

   K =2 : .629 → maior coeficiente, ou seja, o mais indicado

b. E**lbow**: utiliza o SSE para encontrar o numero ideal de clusters.

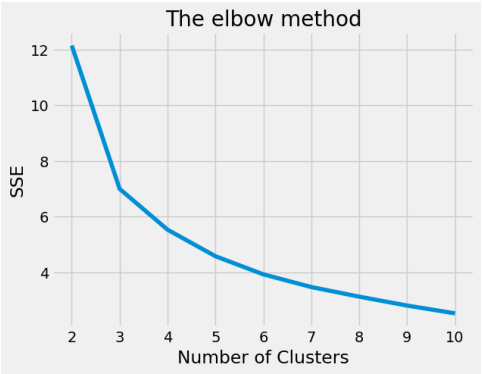


   gráfico de SSE se estabiliza em 3 → o mais indicado

2.

a. **Silhouette:**

O coeficiente de silhouette é uma métrica de avaliação de clusters que mede o quão semelhantes os objetos de um cluster são entre si em comparação com o quão diferentes eles são dos objetos de outros clusters. O coeficiente de silhouette varia de -1 a 1.

Fórmula do Silhouette index:

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}\}}$$

1. Calcule a média do ponto com todos os outros pontos no mesmo cluster. `(a)`
2. Calcule a média entre o ponto e todos os pontos no cluster mais próximo `(b)`
3. O coeficiente de silhouette para o ponto é então calculado como `(b - a) / max(a, b)`. Quanto mais próximo esse valor estiver de 1, melhor o agrupamento.
4. Faça isso para cada amostra

b. **Elbow:**

O método do cotovelo é uma técnica usada para determinar o número ideal de clusters em um conjunto de dados.

Fórmula:

$$SSE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

SSE → a soma dos quadrados das distâncias entre os pontos e seus centróides dentro dos clusters.

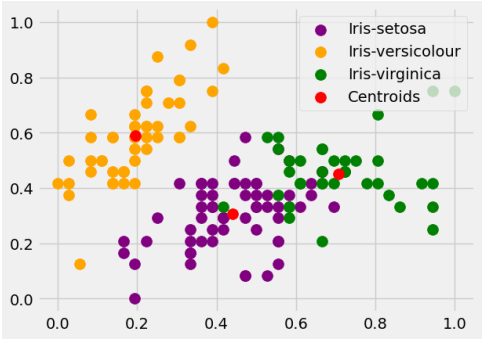
1. Execute o algoritmo K-Means para diferentes valores de k (o número de clusters)
2. Para cada valor de k, calcule o SSE
3. Plote o valor do SSE em relação ao número de clusters k.
4. Observe o gráfico resultante. O ponto onde a curva do gráfico começa a se dobrar (semelhante a um cotovelo) é geralmente considerado o número ideal de clusters.

3. Rand score

O Rand Score (Índice de Rand) é uma métrica de avaliação usada para medir a similaridade entre os agrupamentos (clusters) gerados por um algoritmo de clustering, como o K-Means, e os agrupamentos de referência ou "verdadeiros" conhecidos. O Rand Score varia de 0 a 1.

Para o código o índice foi de 0.7302382722834697 ou cerca de 73% de acerto

4.



Visualmente, podemos notar que a divisão de iris-versicolor ficou melhor que de virginica e setosa, devido a maior distancia entre seus elementos. Assim, tiveram mais erros de classificação da virginica e setosa, em que algumas setosas foram classificadas como virginicas e vice-versa.

5. Etapas de pre processamento do código:
   a. Normalização dos valores: escalando os valores para entre 0 e 1 para cada coluna

```
scaler = MinMaxScaler()
Entrada = scaler.fit_transform(Entrada)
```