PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Instituto de Ciências Exatas e de Informática

Marcos Antonio Lommez Candido Ribeiro[1]

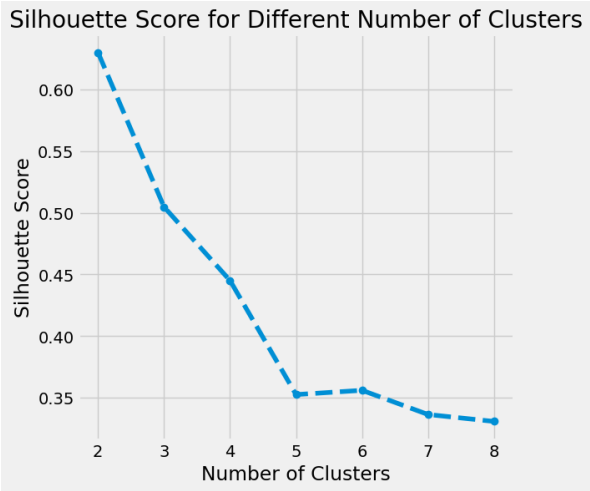
# Lista #6

**Inteligência Artificial**

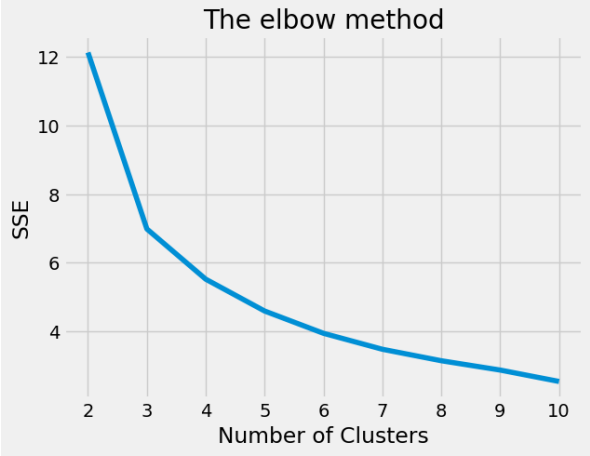
[1]Aluno de Graduação em Ciência da Computação – tonilommez@hotmail.com

## 1. Encontre os agrupamentos, discuta a qualidade destes agrupamentos (usando Silhouette e Elbow) e caracterize os agrupamentos obtidos

Ao utilizarmos k-menas na base e analisar o resultado obtido com silhouette chegamos aos valores de clusters presentes na figura 1. A partir dele podemos ver que o ponto ideal de separação acaba sendo de 2 clusters

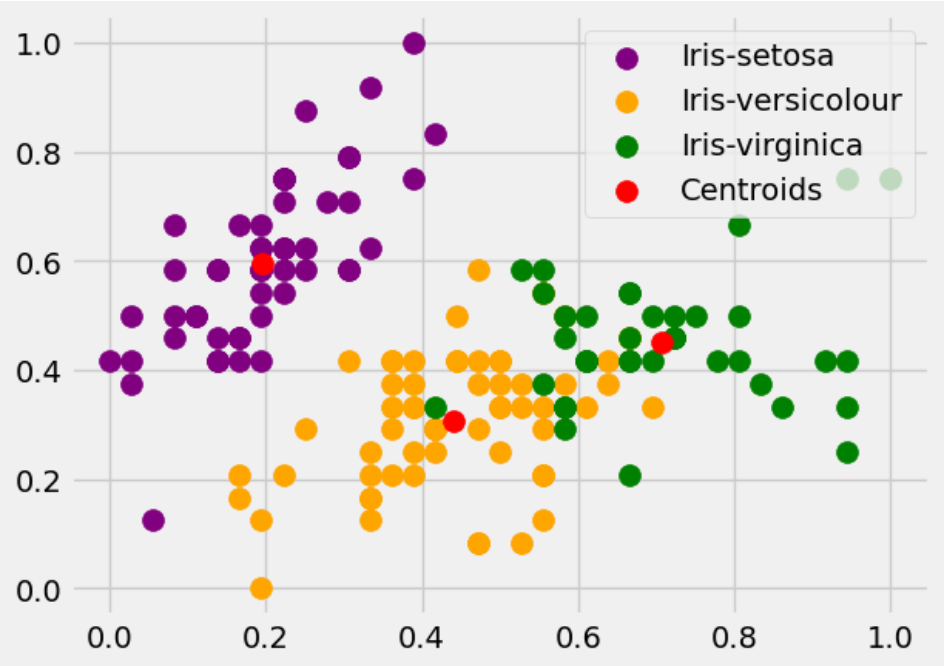



Por outro lado através do método de Elbow chegamos a conclusão de um numero ideal de 3 a 4 clusters baseados nos pontos da curva onde possui maior inflexão.




Junto a isso podemos caracterizar os clusters gerados através dos centroides respectivos. Utilizando-se 3 clusters chegamos as seguintes caracteristicas, que separam de maneira satisfatoria os 3 grupos presentes na base.

| cluster | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) |
|---|---|---|---|---|
| 0 | 5.006 | 3.428 | 1.462 | 0.246 |
| 1 | 6.846 | 3.082 | 5.703 | 2.079 |
| 2 | 5.889 | 2.738 | 4.397 | 1.418 |



## 2. Explique como se obtém estas duas métricas, ou seja, explique as equações matemáticas.

### Silhouette

O coeficiente de Silhouette mede o quão próximo cada ponto em um cluster está dos pontos nos clusters vizinhos. Para um único ponto de dado, o coeficiente de Silhouette é calculado da seguinte forma:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

Onde:

- $s(i)$ é o coeficiente de Silhouette para o ponto de dado $i$.
- $a(i)$ é a distância média do ponto $i$ para os outros pontos no mesmo cluster (coesão).
- $b(i)$ é a distância média do ponto $i$ para os pontos no cluster mais próximo do qual $i$ não faz parte (separação).

O coeficiente de Silhouette varia entre -1 ate +1, sendo que se próximo de +1 o objeto esta proximo de um cluster e longe do cluster vizinho. Igualmente um valor proximo de -1 indica que o objeto esta dentro do cluster vizinho.

### WCSS (Within-Cluster Sum of Squares)

WCSS é a soma das distâncias quadradas entre cada ponto de dado e o centroide de seu cluster. É uma métrica de coesão que tentamos minimizar ao usar o algoritmo K-means. Para um único cluster, WCSS é calculado como:

$$\text{WCSS}_k = \sum_{i=1}^{n} (x_i - c_k)^2 \tag{2}$$

Onde:

- $n$ é o número de pontos no cluster $k$.
- $x_i$ é um ponto de dado no cluster.
- $c_k$ é o centroide do cluster $k$.

A metrica WCSS é feita para diferentes valores de K. Para usa-la deve-se analisar o ponto de cotovelo da curva que é onde a curva começa a se tornar horizontal.

## 3. Investigue, explique e implemente, pelo menos, mais 1 métrica de avaliação dos agrupamentos, diferentes das 2 anteriores

### Índice Davies-Bouldin (DBI)

O Índice Davies-Bouldin é uma métrica que avalia a qualidade dos clusters com base na similaridade entre cada cluster e o seu cluster mais similar, onde a similaridade é a relação entre as distâncias dentro do cluster e as distâncias entre os clusters. Um valor menor de DBI indica uma melhor separação entre os clusters.

O DBI é definido como:

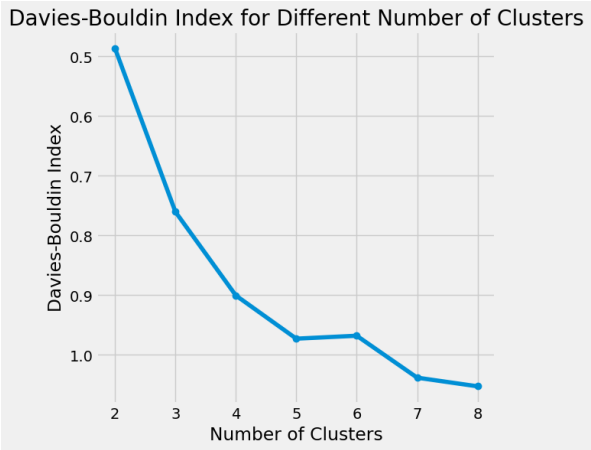
$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left( \frac{S_i + S_j}{D_{ij}} \right) \tag{3}$$

Onde:

- $k$ é o número total de clusters.
- $S_i$ é a média da distância de todos os pontos no cluster $i$ ao centróide de $i$.
- $D_{ij}$ é a distância entre os centróides dos clusters $i$ e $j$.

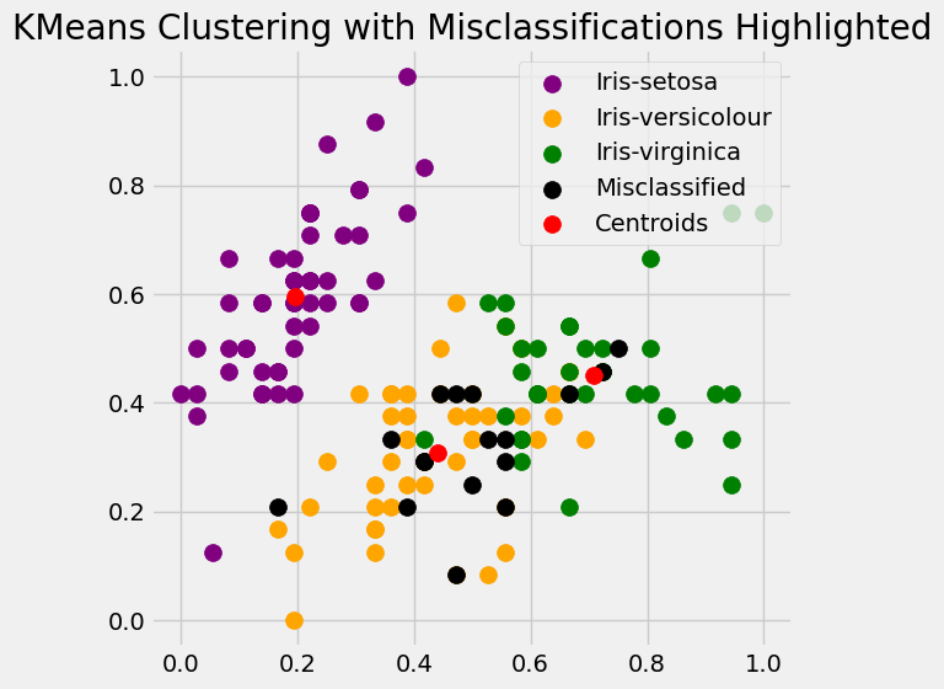
Para utilizar a metrica deve-se entender que valores menores do índice indicam uma melhor separação entre os clusters, porque um valor menor implica que os clusters estão mais distantes entre si (maior Dij) e/ou mais compactos (menor Si). Um DBI igual a 0 indica uma partição perfeita dos clusters.

Aplicando a metrica na base de dados da Iris, chegamos aos valores apresentados na figura a seguir:




A partir dele vemos um valor muito próximo do gerado com a o Silhouette Score, onde o numero ideal de clusters começa fortemente em 2 e decai rapidamente ate 4.

**4. Uma vez que a base é classificada (setosa, virgínica e versicolor), mostre visualmente que instâncias foram agrupadas incorretamente pelo kmeans.**
**Discuta os resultados.**




Ao primeiro momento podemos dizer que o K-means fez um bom trabalho classificando os grupos por ter acertado majoritariamente os valores das classes. Confusões existiram apenas entre alguns valores entre as classes Virginica e Versicolour, porque ambas possuem instancias cujo estao muito proximas aos centroides gerados das outras classes, sendo assim valores ambiguos para a classificação de grupos do k-means

## 5. Faça um pequeno relatório explicando todas as etapas de pré-processamento realizadas e explicando todos os resultados obtidos.

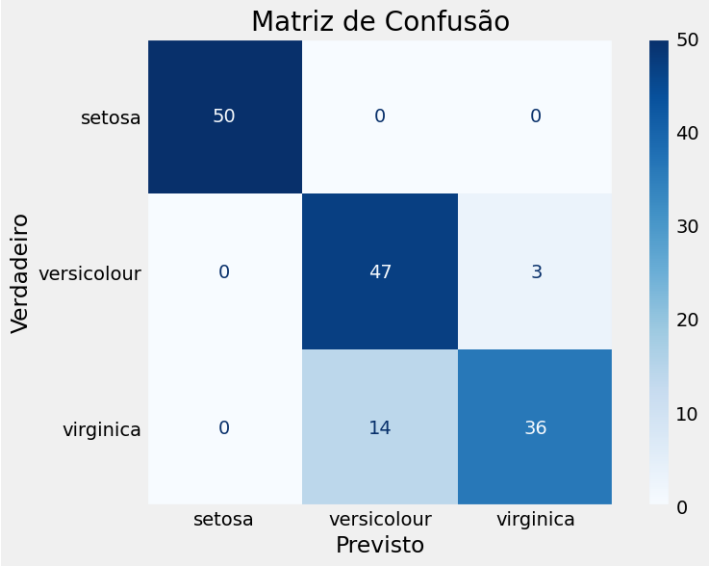
Os dados foram carregados da base da biblioteca scikit-learn. Sendo uma base artificial, já estão balanceados e isentos de problemas típicos de bases reais. A única normalização aplicada foi o MinMaxScaler, ajustando os valores entre 0 e 1.

Além disso devido a natureza do k-means ser de agrupamento e nao de classificação, foi necessario converter a saida do k-means para os devidos valores da base de dados, onde a ordem de classificação original era 0, 1, 2 e a gerada pelo k-means foi 1, 0, 2.

Com as devidas correções e adaptações montadas foram feitas analises da qualidade dos clusters gerados através de 3 métricas, sendo estas o Silhouette Score, WCSS (Within-Cluster Sum of Squares) e o índice Davies-Bouldin.

Com as adaptações, analisamos a qualidade dos clusters usando o Silhouette Score, WCSS e o índice Davies-Bouldin. Estas métricas indicaram 2 clusters como ideais, mas, baseando-se na natureza da base, mantivemos 3 grupos, conforme a classificação original.

Com isso pode-se fazer uma analise sobre a capacidade de um modelo de clusterização como o k-means de possivelmente classificar uma base de dados. Os resultados obtidos foram bons, chegando a uma acuracia de 88%. E asim como podemos ver a partir da matriz de confusão, a classe setosa acertou 100% dos valores enquanto existiu uma leve confusão apenas ao classificar alguns valores de virginica como versicolour.



**Todo o codigo gerado pode ser encontrado no link a seguir**

Link do github