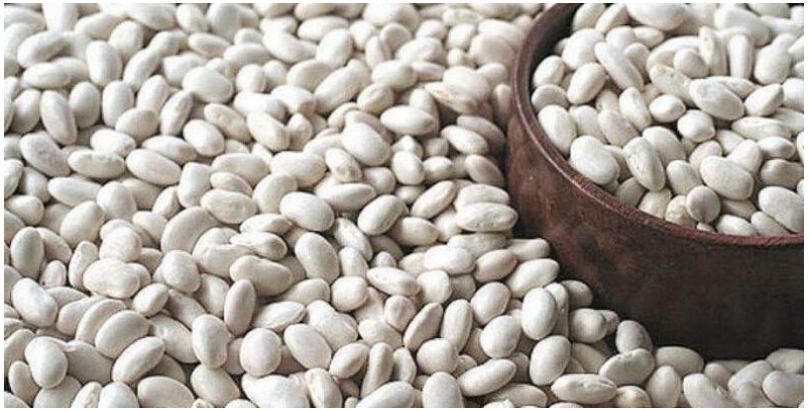
Dry Bean Dataset

Augusto Luchesi Matos, RA: 740871
Carlos Eduardo Nascimento dos Santos, RA: 791029
Gabriel Meirelles Carvalho Orlando, RA: 790728

Sobre o Dataset

- O Dataset escolhido é de feijões secos que podem ser separados em 7 tipos distintos
- É composto de 16 atributos + classe
- São 14 atributos numéricos contínuos e 2 atributos numéricos discretos
- O dataset não possui nenhum valor nulo em nenhuma classe

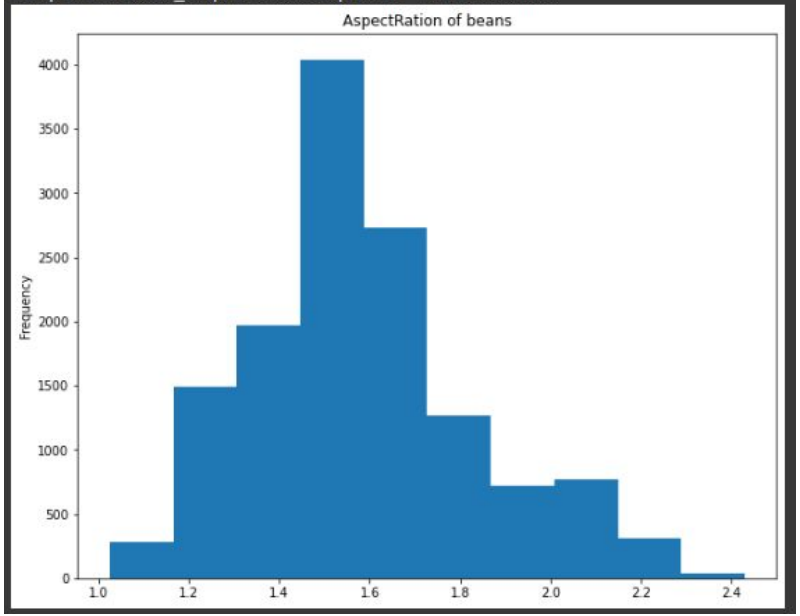


https://www.kaggle.com/datasets/muratkokludataset/dry-bean-dataset

Atributos Numéricos Contínuos

- Perímetro
- Comprimento do eixo principal (L)
- Comprimento do eixo menor (l)
- Proporção (entre L e l)
- Excentricidade
- Diâmetro Equivalente
- Extensão
- Solidez
- Redondeza
- Compacticidade
- Fator de formato 1, 2, 3 e 4

```
A media do atributo é: 1.58
O desvio padrão do atributo é 0.247
O valor máximo do atributo é 2.43
O valor mínimo do atributo é 1.02
O atributo não tem nenhum valor nulo
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ba796aa90>
```

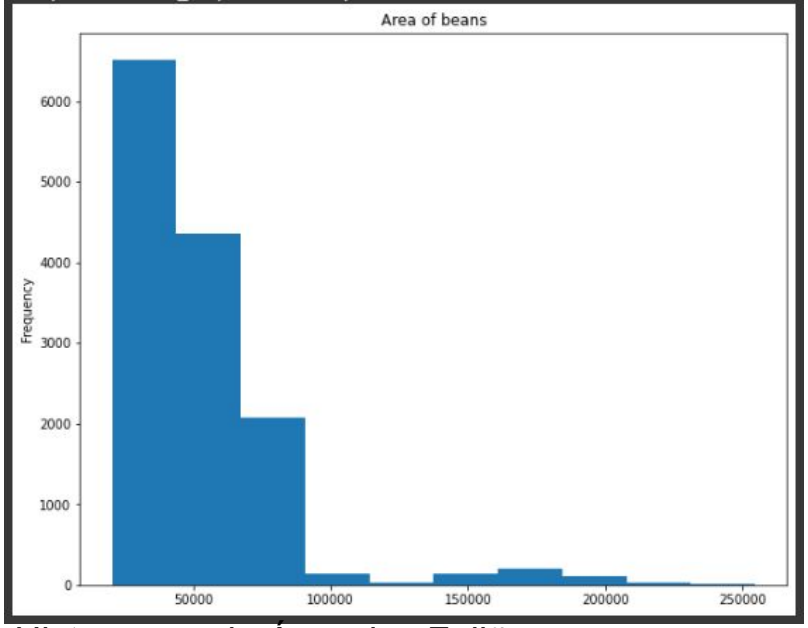


Histograma da Proporção dos Feijões (entre L e l)

Atributos Numéricos Discretos

- Área
- Área Convexa

```
A media do atributo é: 53048.28
O desvio padrão do atributo é 29324.096
O valor máximo do atributo é 254616.00
O valor mínimo do atributo é 20420.00
O atributo não tem nenhum valor nulo
<matplotlib.axes._subplots.AxesSubplot at 0x7f4bad478150>
```



Histograma da Área dos Feijões

Normalização

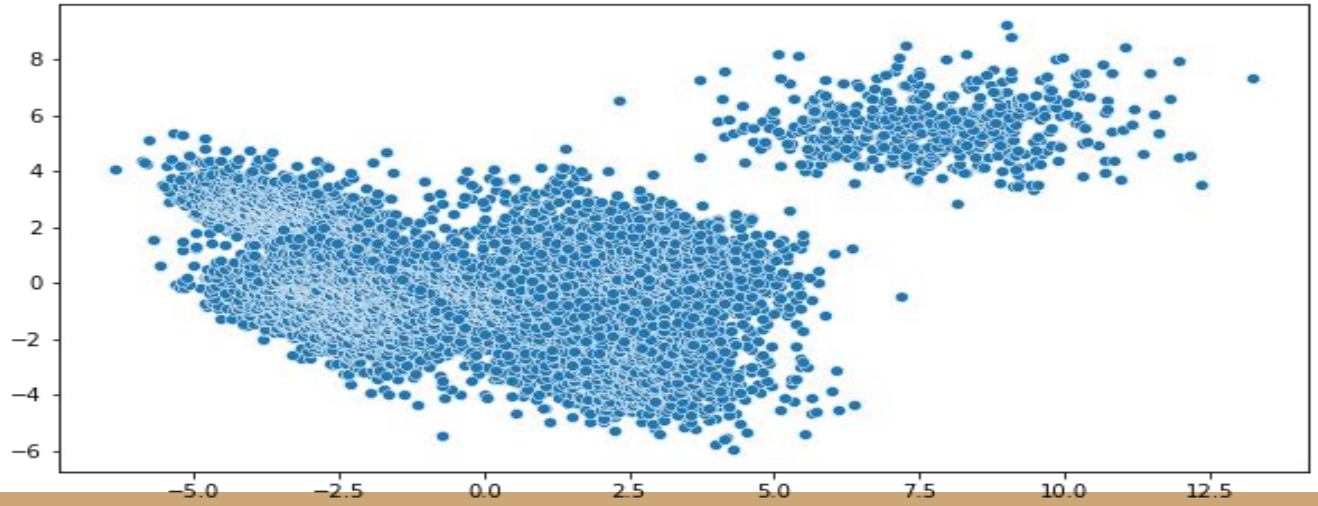
- Para evitar que a aplicação de medidas de distância utilizados nos algoritmos de agrupamento escolhidos seja afetado pela grande diferença de escala nos atributos, foi feita uma normalização do conjunto de dados.

```
df_normalized = StandardScaler().fit_transform(df)
```

- Os dados são normalizados de forma que fiquem com **MÉDIA = 0** e **DESVIO PADRÃO = 1.**

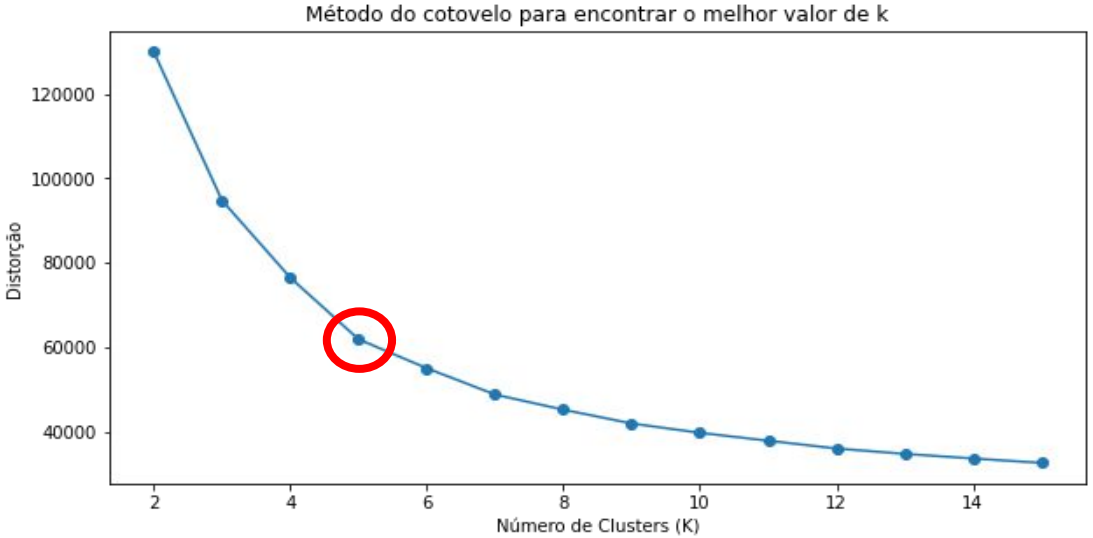
Redução de Dimensionalidade

- Para enxergar em 2 dimensões o conjunto escolhido, realizamos um método de redução de dimensionalidade para plotar o gráfico, o PCA.
- PCA
  - O PCA é baseado na variância dos dados, ou seja, ele tenta criar uma nova representação dos dados, com uma dimensão menor, mantendo a variância entre eles.



Algoritmos Utilizados

- K-Means
  - Método de Inicialização = **k-means++**
  - Foram testados diversos valores para K e foi escolhido o melhor utilizando o método do cotovelo, formado pela soma do quadrado dos erros.

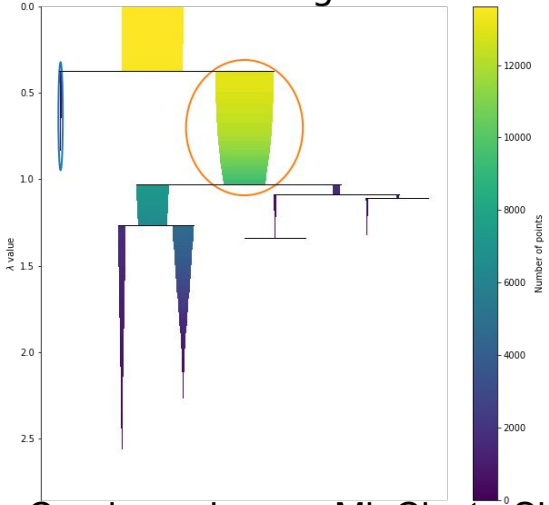


- Valor escolhido, K = 5

Algoritmos Utilizados

- HDBSCAN
  - Métrica = **euclidiana**
  - Foram testados com 5 valores para o número de pontos mínimos do cluster e, a cada iteração, foi testado o índice de silhueta. A partir deste, foi escolhido o melhor MinClusterSize.

Min cluster size igual à 30

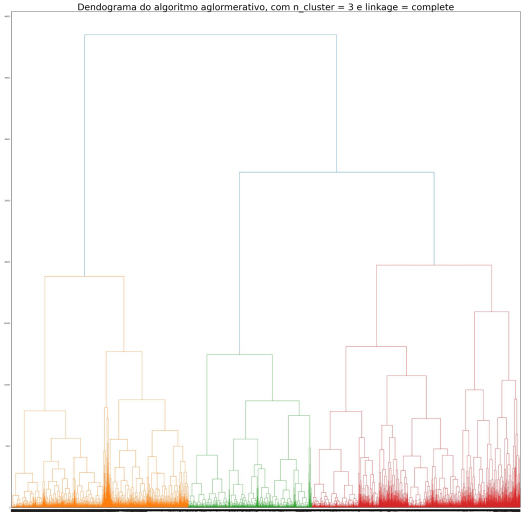


Árvore Condensada para MinClusterSize = 30

- Valor escolhido, MinClusterSize = 30

Algoritmos Utilizados

- Algoritmo Aglomerativo
  - O algoritmo foi testado utilizando 3 tipos de ligações: **simples, completa e média**
  - Cada tipo de ligação foi testada com valores de 2 a 7 para os clusters e, a cada iteração, foi testado o índice de silhueta. A partir deste, foi escolhido o melhor número de clusters.



Dendograma de ligação completa e n_cluster = 3

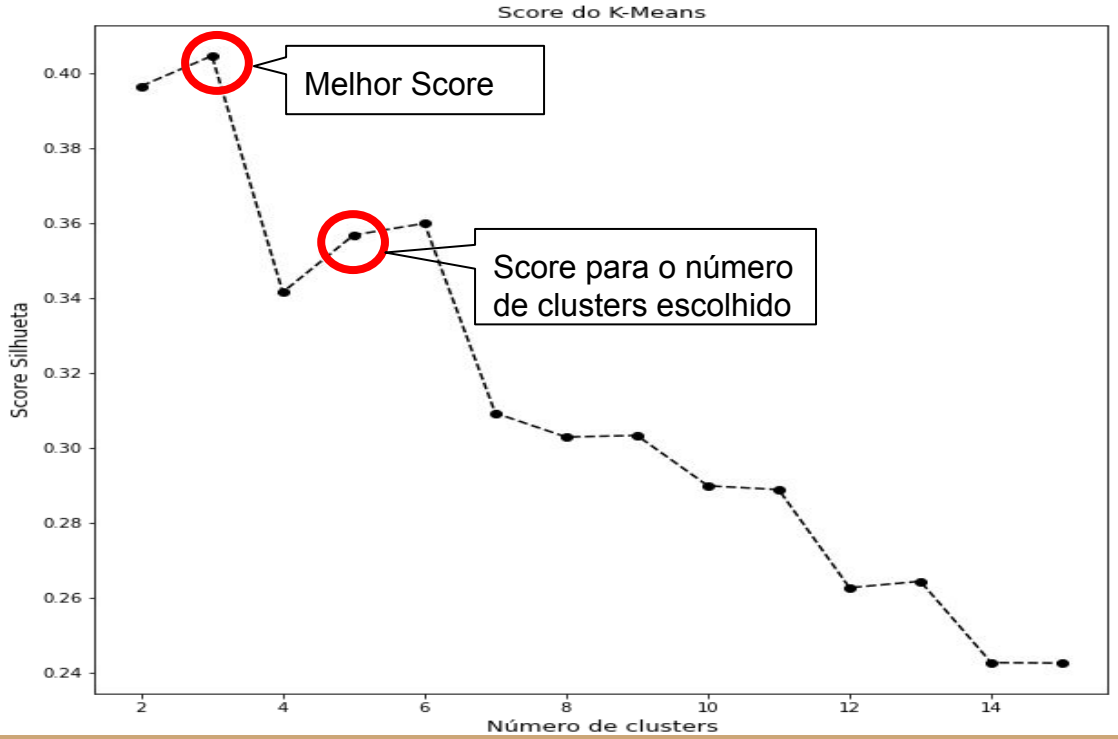
- Melhor valor para ligação **simples** = 2
- Melhor valor para ligação **completa** = 3
- Melhor valor para ligação **média** = 2

Métrica de Validação

- A métrica escolhida para a validação dos grupos foi o **Índice da Silhueta.**
- **Índice da Silhueta**
  - Quanto maior a distância do objeto para os outros grupos e menor a distância para seu grupo, melhor será a avaliação
- Utilizamos a função disponibilizada pelo scikit-learn, **silhouette_score**
  - O melhor valor para a métrica é 1 e o pior -1. Valores perto de 0 indicam sobreposição de grupos, enquanto valores negativos geralmente indicam que a amostra foi designada para um grupo errado, pois um grupo diferente era mais similar.
  - calcula usando a média da distância intra-grupos (**a**) e a média da distância dos grupos mais próximos (**b**) para cada amostra. Para uma amostra o coeficiente é **(b - a) / max(a, b).**

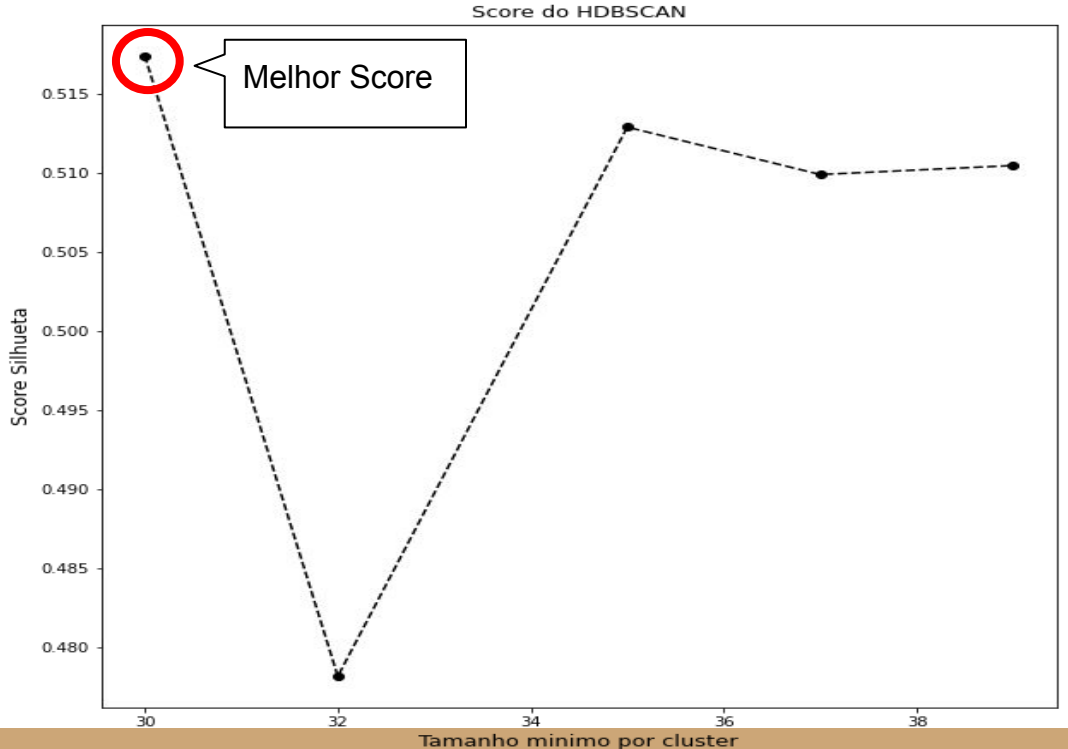
Resultados do Índice da Silhueta

- K-Means



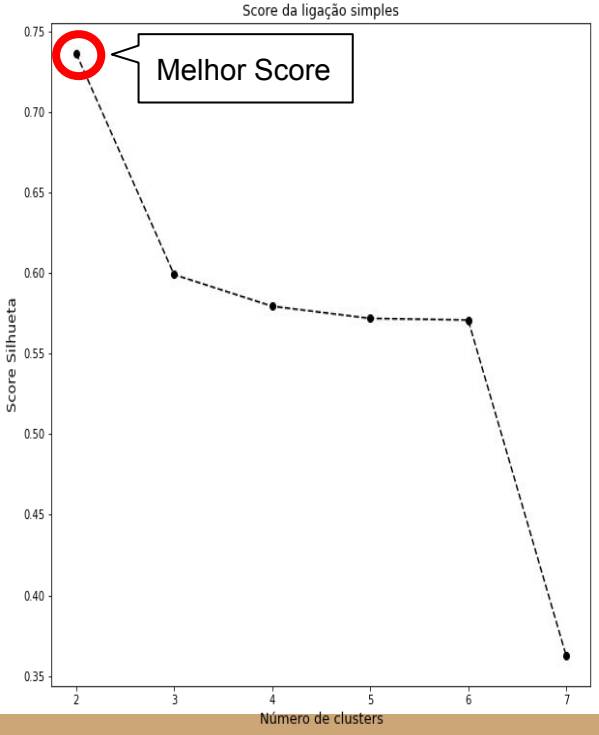
Resultados do Indice da Silhueta

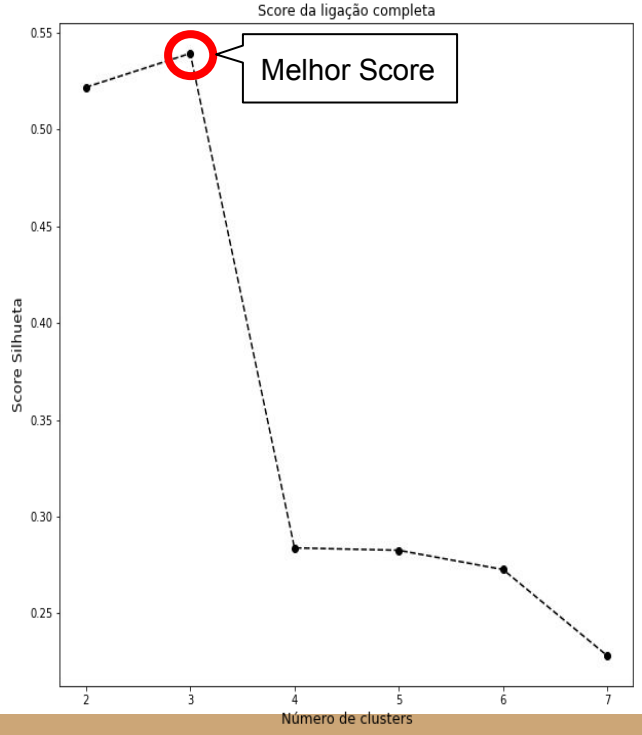
- HDBSCAN

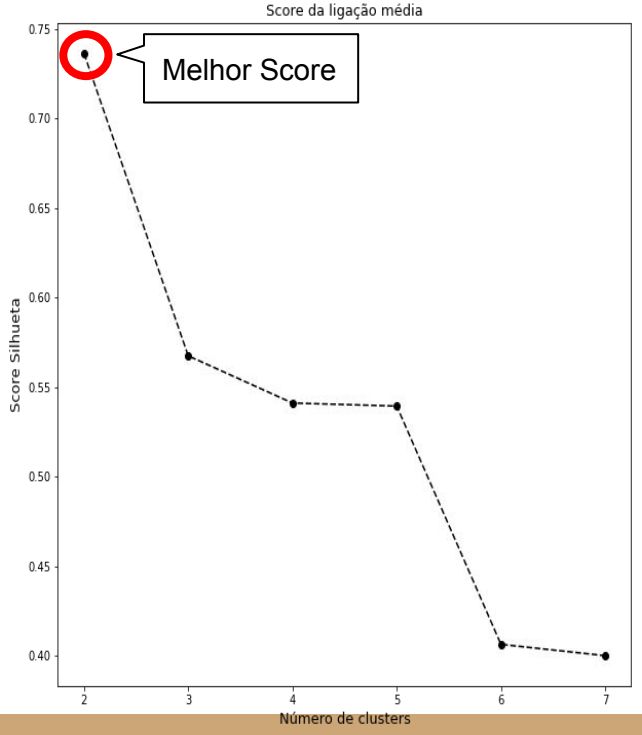


Resultados do Índice da Silhueta

- Algoritmo Aglomerativo







Comparação dos Resultados do Índice de Silhueta

Algoritmo Utilizado	Melhor Score	Score para n_clusters = 5
K-means	0.4047 / n_cluster = 3	0.3568
HDBSCAN	0.5173 / MinClusterSize = 30 / n_cluster = 2	—
Aglomerativo - Ligação Simples	0.7361 / n_cluster = 2	0.5717
Aglomerativo - Ligação Completa	0.5393 / n_cluster = 3	0.2826
Aglomerativo - Ligação Média	0.7361 / n_cluster = 2	0.5393

Conclusão

- Ao analisar a métrica de validação utilizada, o melhor método de agrupamento foi o Algoritmo Aglomerativo. Sendo que as medidas de ligação média e ligação simples empataram em 0.7361 com um total de 2 clusters.

- Apesar do número de clusters escolhidos com o SSE do k-means ser 5, ele não teve um bom desempenho com a métrica da silhueta, pois o SSE só considera a distância intra-grupos dos clusters formados, enquanto a silhueta considera a distância intra-grupos e a distância inter-grupos.

- O resultado do HDBSCAN evidencia a separação existente no conjunto de dados analisado, já que existem 2 conjuntos de pontos muito separados, sendo um muito grande e outro muito pequeno. No que contém muitos pontos, os subgrupos que podem ser gerados não têm densidade suficiente para gerar um novo grupo, e dessa forma o HDBSCAN identifica apenas 2 grupos.

Conclusão

- O resultado do Algoritmo Aglomerativo com ligação simples é explicável. Como este método é poderoso para identificar grupos densos, quando o número de clusters escolhido é 2 há uma maior densidade dos grupos, visto que há uma clara separação dos pontos do conjunto de dados analisado.

- Já o resultado do Algoritmo Aglomerativo com ligação completa não obteve resultados tão bons quanto ao algoritmo configurado com outros parâmetros, pois este método quebra os grupos grandes em menores, explicando o porquê quando o número de clusters escolhido foi 2, o algoritmo não obteve o melhor desempenho.

- Finalmente, o resultado do Algoritmo Aglomerativo com ligação média é muito semelhante ao método da ligação simples, visto que a ligação média tem características da ligação simples, como, por exemplo, a identificação de grupos densos, que é um atributo importante no conjunto de dados trabalhados.

- Por fim, como nosso dataset se trata de diversas características de feijões, é possível analisar que nossos dados podem ser divididos em 2 grupos de feijões que realmente se diferem em suas características, o resultado que adotamos, pois mais divisões nesse grupo, de acordo com nossos algoritmos, seriam possíveis, mas já acrescentaria sobreposições nos grupos e os feijões não teriam mais tantas diferenças e começariam a ser mais similares entre os grupos.