T2 - Clustering

Aprendizado de Máquina 1
2023-01

Geovanne Mansano Fritch da Silva	791072
Luana de Queiroz Garcia	740953
Matheus Bessa Coutinho Colombo	801839
Pedro Freire Baleeiro	790984
Thiago César Silva Barbieri	779807

Introdução

Heart Disease

Relevância do Tema

- O Infarto Agudo do Miocárdio é a maior causa de mortes no país.

- Estima-se que, no Brasil, ocorram de 300 mil a 400 mil casos anuais de infarto, e a cada 5 a 7 casos, ocorre um óbito.

- Devido à sua prevalência, impacto na saúde e custos associados, o estudo e a compreensão dos problemas cardíacos são fundamentais para avançar na prevenção, no diagnóstico e no tratamento.

https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto

Diferentes Tipos

- **Tipo 1** *(mais conhecido)***:** desencadeado pela obstrução da passagem do sangue por uma artéria.
- **Tipo 2:** ocasiona a diminuição da irrigação sanguínea no músculo cardíaco por causa de eventos graves. (ex.: crise hipertensiva, espasmos das artérias, pressão baixa e arritmias)
- **Tipo 3** *(infarto fulminante)*:geralmente, a falta de oxigênio e nutrientes mata a maioria das células cardíacas, e leva à morte súbita.
- **Tipo 4:** casos de infarto pós angioplastia das artérias coronárias, que podem ocorrer logo depois desse tratamento ou em decorrência de nova obstrução por cima do stent.
- **Tipo 5:** quando existe relação entre o infarto e a revascularização cardíaca (ponte de safena).

Análise exploratória dos Dados

Heart-Disease-patients

Conta ao total com 303 amostras e 14 dimensões.

Tipos de dados do dataset:

- 4 atributos numéricos contínuos
- 10 atributos numéricos discretos

Os atributos do dataset

- **age**: idade
- **sex**: sexo
- **cp** (chest pain): dor no peito
- **trestbps** (resting blood pressure): pressão arterial da pessoa em repouso ao chegar no hospital
- **chol** (cholesterol): colesterol
- **fbs** (fasting blood sugar): nível de açúcar no sangue em jejum > 120 mg/dl (1 = true; 0 = false)
- **restecg** (resting electrocardiographic results): resultado do eletrocardiograma em repouso

Os atributos do dataset

- **thalach** (maximum heart rate achieved): máximo de batimentos cardíacos alcançado
- **exang** (exercise induced angina): exercício induziu a angina (1 = yes; 0 = no)
- **oldpeak** (ST depression induced by exercise relative to rest): depressão de ST induzida por exercício em relação ao repouso
- **slope** (the slope of the peak exercise ST segment): inclinação do pico do segmento ST do exercício — 1: descendente; 2: plano; 3: ascendente
- **ca** (number of major vessels colored by flourosopy): número de vasos principais coloridos por fluoroscopia (0-3)
- **thal** (thalassemia): talassemia — 3 = normal; 6 = fixed defect; 7 = reversable defect
- **num** (diagnosis of heart disease): diagnóstico de doença cardíaca — valor 0: < 50% estreitamento do diâmetro da artéria (ausência de doença coronária); - Valores 1-4: > 50% estreitamento do diâmetro da artéria (presença de doença coronária)

As 10 primeiras amostras do dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
5	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0
6	62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0	0.0	3.6	3.0	2.0	3.0	3
7	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0
8	63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0	0.0	1.4	2.0	1.0	7.0	2
9	53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0	0.0	7.0	1

Atributos Numéricos

Dentre os atributos numéricos e discretos temos 3 que são binários:

- Sex (sexo)
- Fbs (nível de açúcar no sangue em jejum)
- Exang (exercício induziu a angina)

Atributos Numéricos

Dentre os atributos numéricos discretos restantes temos:

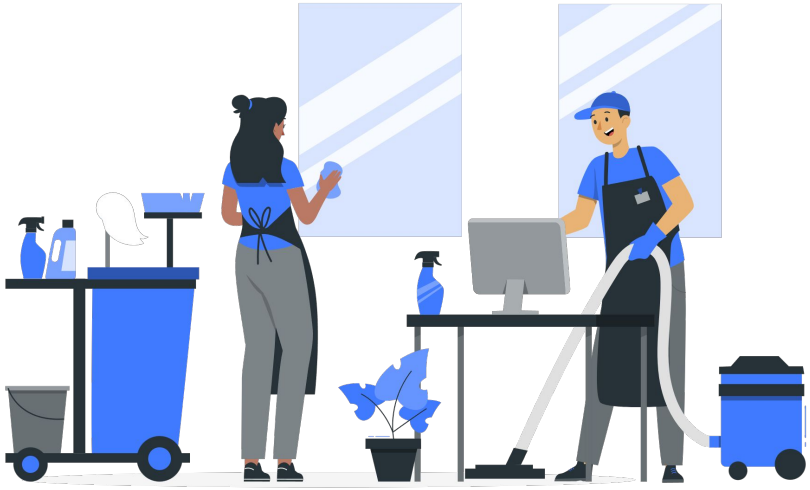
- Age
- Cp (dor no peito)
- Restecg
- Slope
- Ca
- Thal
- Num

Dentre os atributos numéricos contínuos temos:

- Trestbps
- Chol
- Thalach
- Oldpeak

Algumas estatísticas dos atributos

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.937294
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075	0.616226	1.228536
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	2.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	4.000000



Limpeza dos dados

Tratamento dos Atributos Numéricos

Dados Faltantes: não há

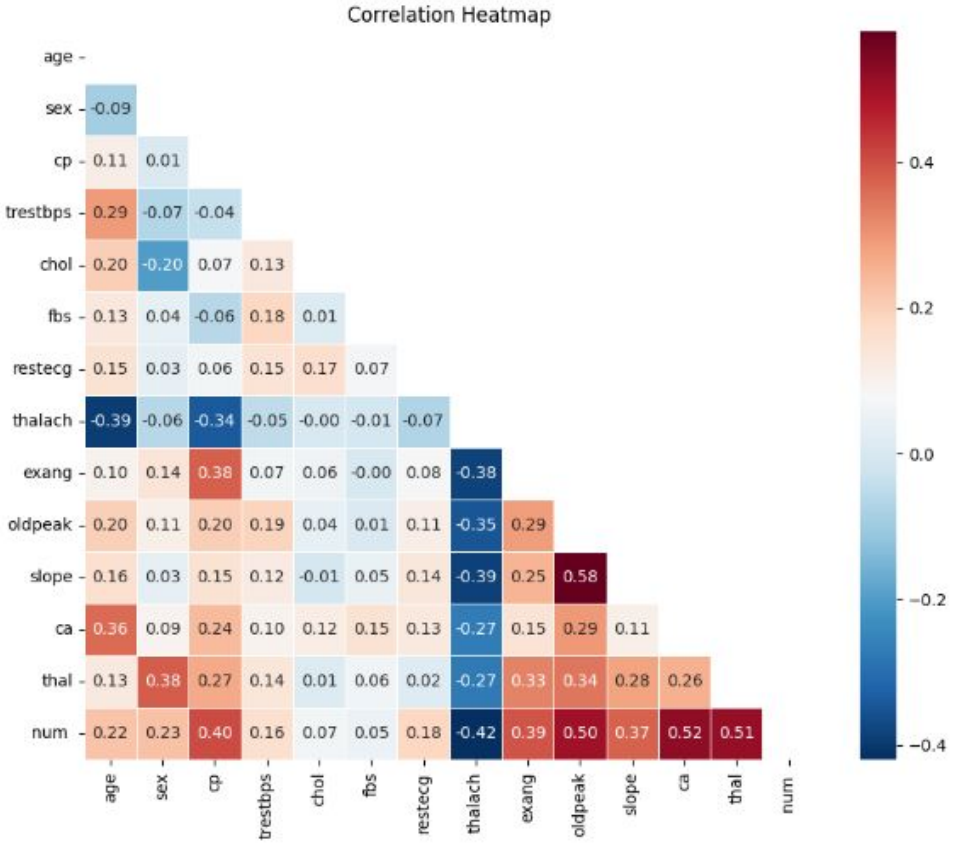
Dados Duplicados: não há

Dados Inconsistentes: foram encontrados 6 com campos ‘?’

- as amostras inconsistentes foram removidas

Transformação de dados: 2 atributos foram transformados de objeto para atributo numérico discreto (eles já eram numéricos discretos, foi só uma mudança de formato)

Correlação entre Atributos



- Menor correlação foi de -0.42
- Maior correlação de 0.58
- Sem correlações relevantes para remover alguma dimensão

Normalização dos dados

A normalização dos dados foi feita com a função *MinMaxScaler()*.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
292	57.0	0.0	4.0	140.0	241.0	0.0	0.0	123.0	1.0	0.2	2.0	0.0	7.0
293	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0	0.0	7.0
294	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0	2.0	7.0
295	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0	1.0	7.0
296	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0	1.0	3.0

Antes da Normalização

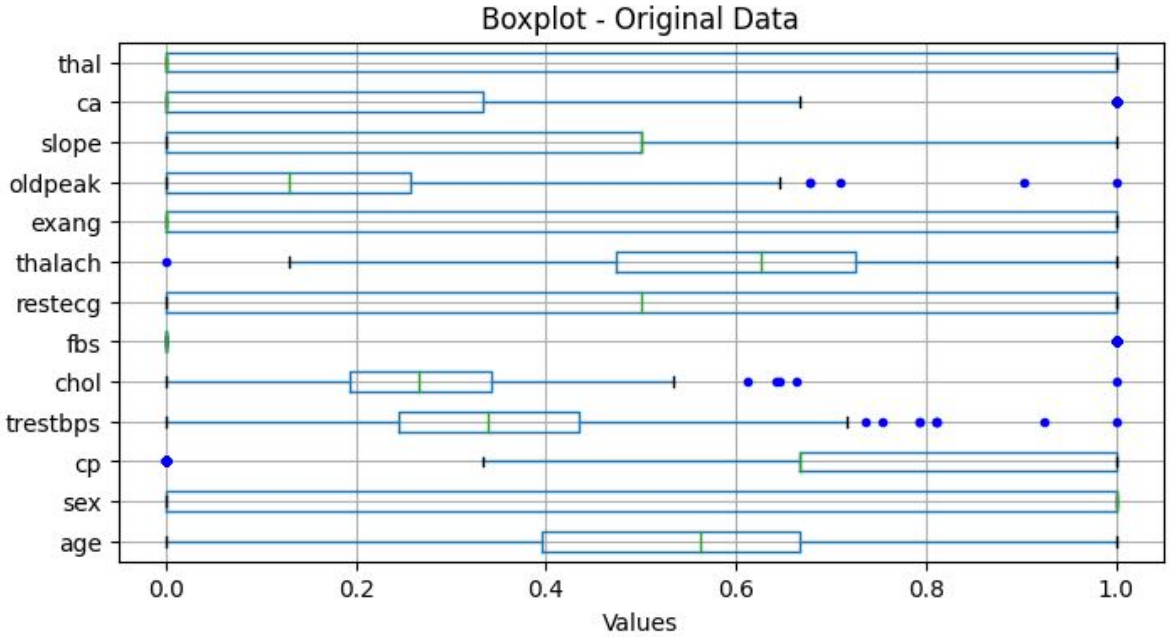
Normalização dos dados

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	0.708333	1.0	0.000000	0.481132	0.244292	1.0	1.0	0.603053	0.0	0.370968	1.0	0.000000	0.75
1	0.791667	1.0	1.000000	0.622642	0.365297	0.0	1.0	0.282443	1.0	0.241935	0.5	1.000000	0.00
2	0.791667	1.0	1.000000	0.245283	0.235160	0.0	1.0	0.442748	1.0	0.419355	0.5	0.666667	1.00
3	0.166667	1.0	0.666667	0.339623	0.283105	0.0	0.0	0.885496	0.0	0.564516	1.0	0.000000	0.00
4	0.250000	0.0	0.333333	0.339623	0.178082	0.0	1.0	0.770992	0.0	0.225806	0.0	0.000000	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
292	0.583333	0.0	1.000000	0.433962	0.262557	0.0	0.0	0.396947	1.0	0.032258	0.5	0.000000	1.00
293	0.333333	1.0	0.000000	0.150943	0.315068	0.0	0.0	0.465649	0.0	0.193548	0.5	0.000000	1.00
294	0.812500	1.0	1.000000	0.471698	0.152968	1.0	0.0	0.534351	0.0	0.548387	0.5	0.666667	1.00
295	0.583333	1.0	1.000000	0.339623	0.011416	0.0	0.0	0.335878	1.0	0.193548	0.5	0.333333	1.00
296	0.583333	0.0	0.333333	0.339623	0.251142	0.0	1.0	0.786260	0.0	0.000000	0.5	0.333333	0.00

Após a Normalização

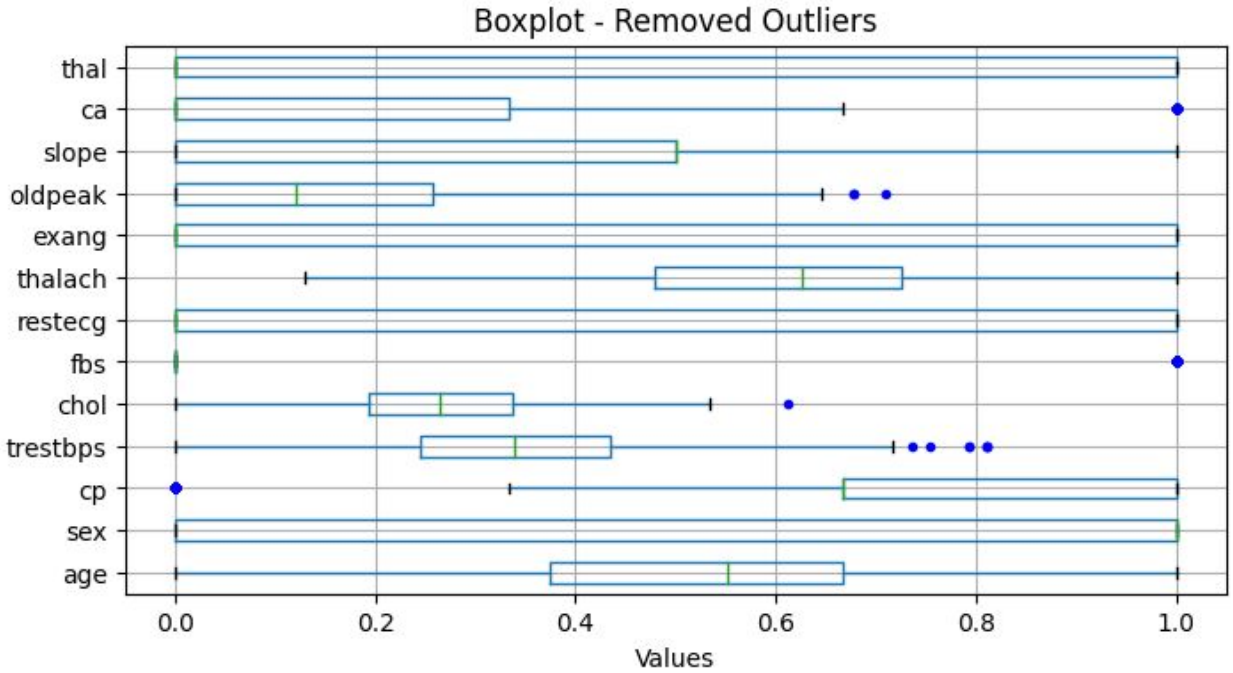
Remoção de Outliers

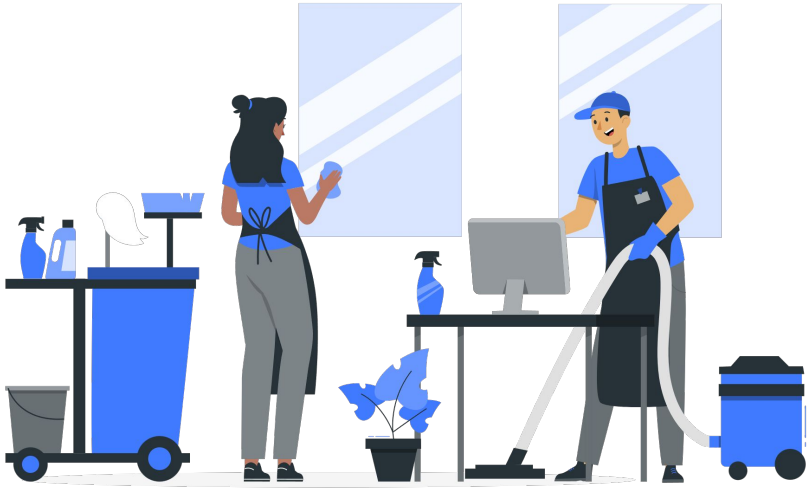
- Utilizado o padrão Z-score.
- Definido o limite absoluto em 3.



Remoção de Outliers

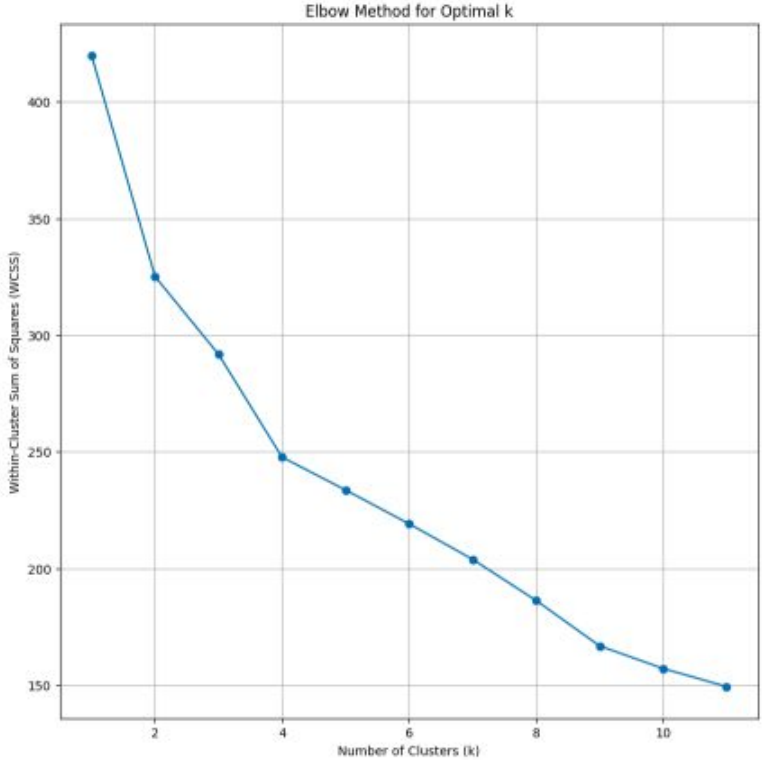
Foram encontrados ao todo 9 outliers, e foram removidos do dataset.

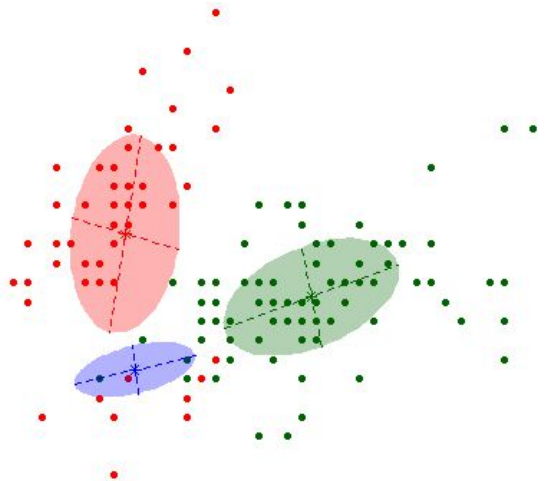




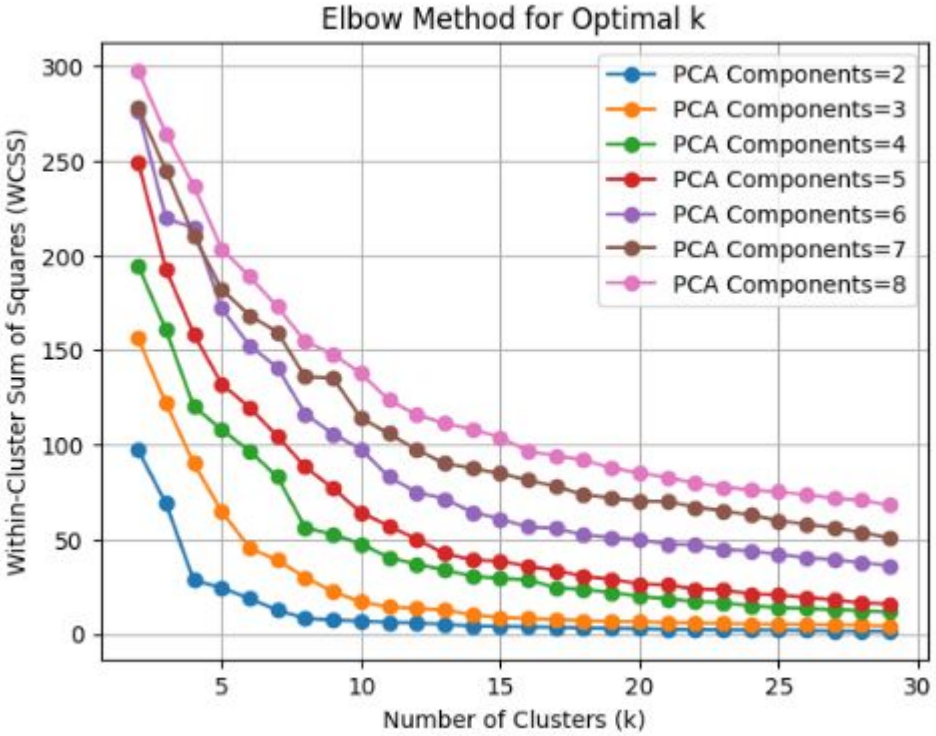
Classificadores

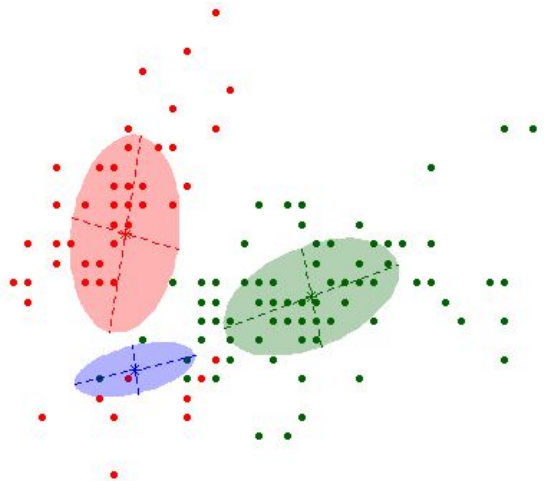
K-Means


Elbow Method for Optimal k
Within-Cluster Sum of Squares (WCSS)
Number of Clusters (k)
400
350
300
250
200
150
2
4
6
8
10

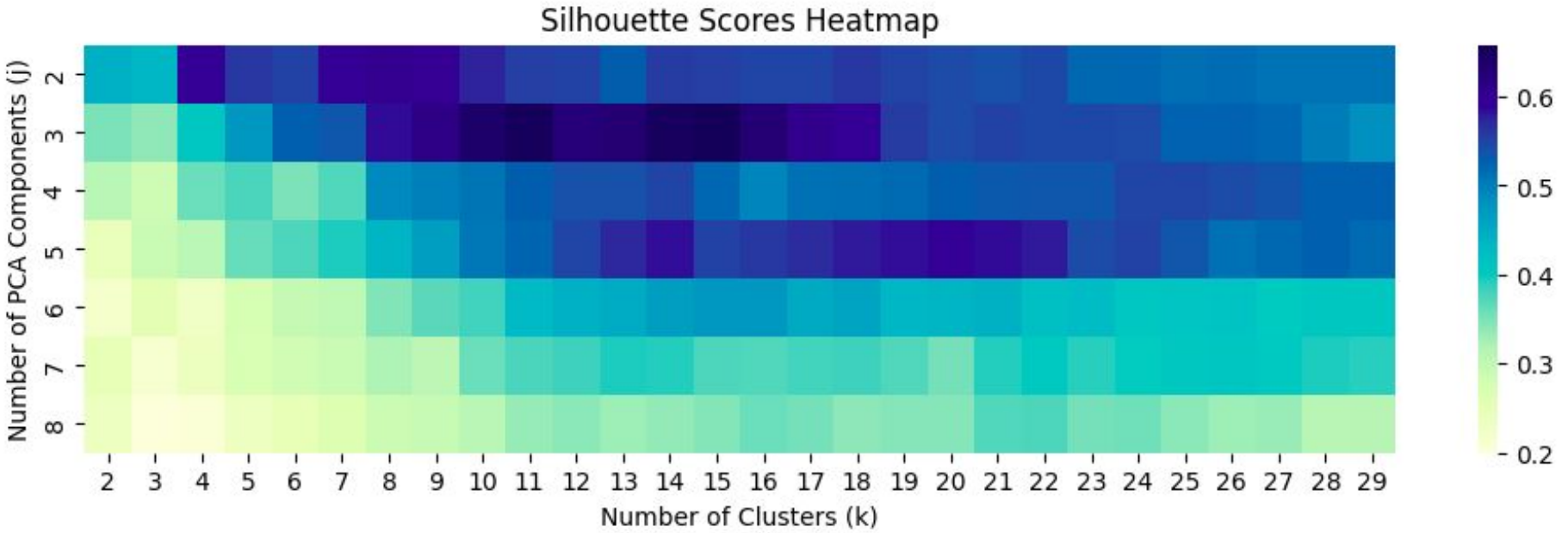


K-Means

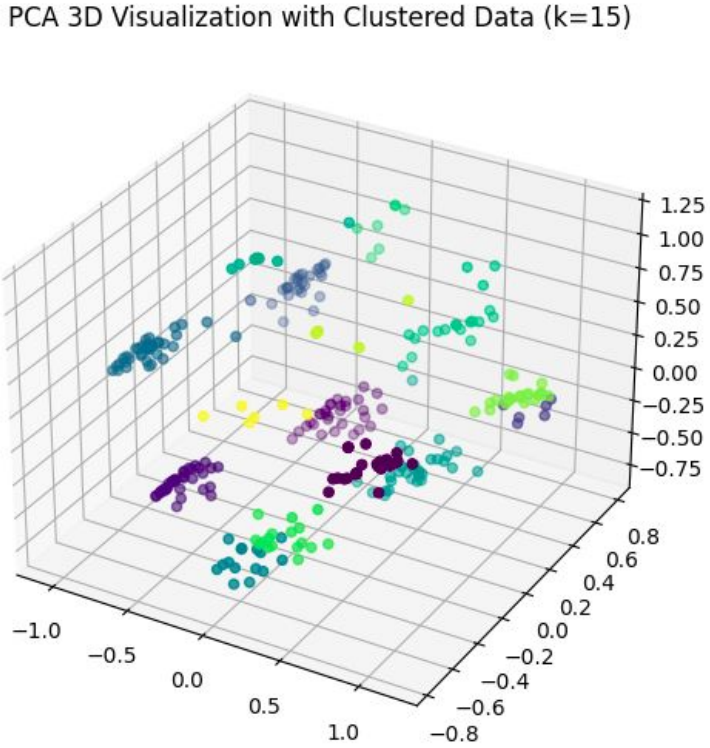

Elbow Method for Optimal k
PCA Components=2
PCA Components=3
PCA Components=4
PCA Components=5
PCA Components=6
PCA Components=7
PCA Components=8
Within-Cluster Sum of Squares (WCSS)
Number of Clusters (k)
0
50
100
150
200
250
300
5
10
15
20
25
30

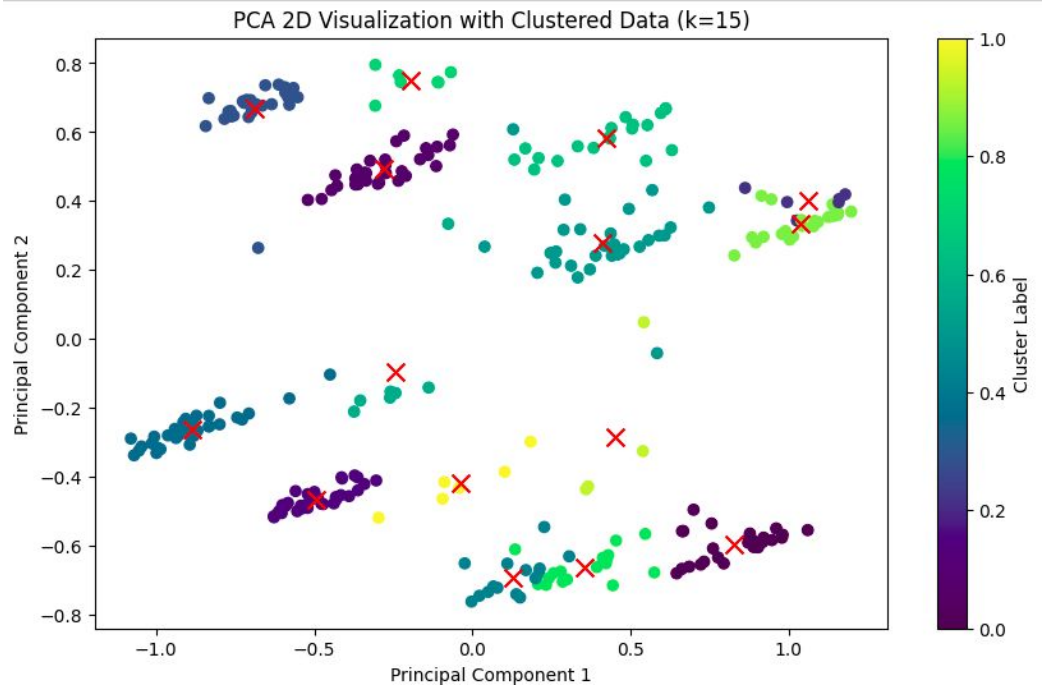


K-Means

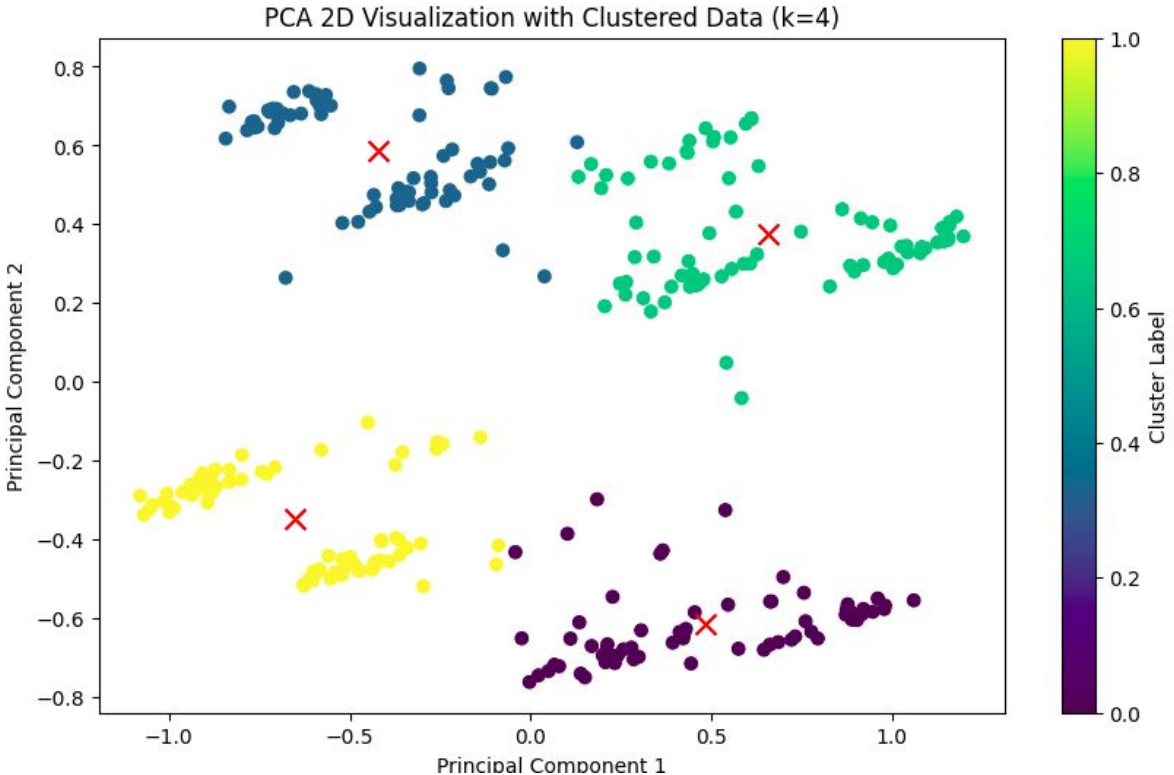

Silhouette Scores Heatmap
Number of PCA Components (j)
2
3
4
5
6
7
8
Number of Clusters (k)
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
0.6
0.5
0.4
0.3
0.2

K-Means (k = 15), PCA = 3





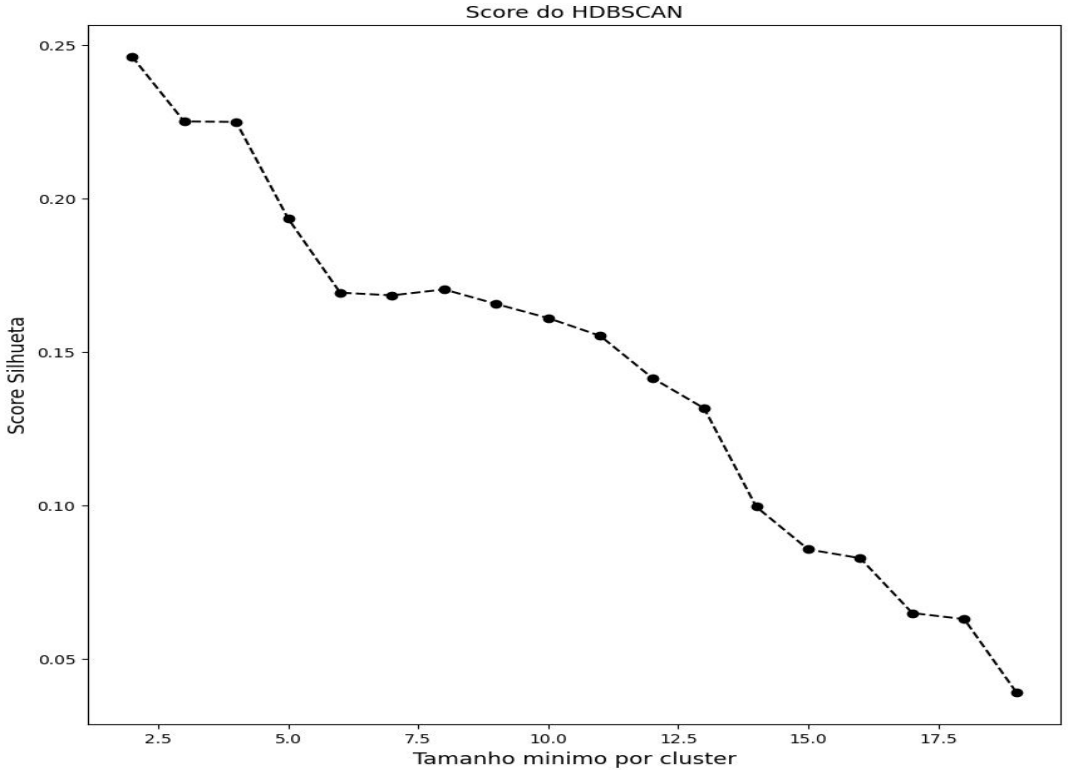
K-Means (k = 4), PCA = 2

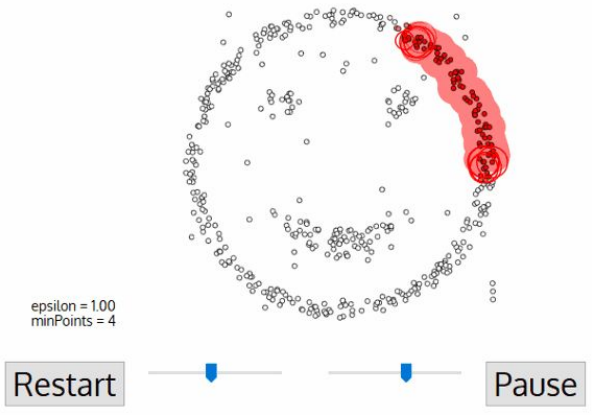


K-means

- O K-means demonstrou desempenho variável, dependendo das configurações de PCA e número de clusters.
- A configuração com dois componentes principais (PCA) e quatro clusters foi especialmente relevante, possivelmente refletindo subdivisões coerentes dos grupos com falha cardíaca.

HDBSCAN

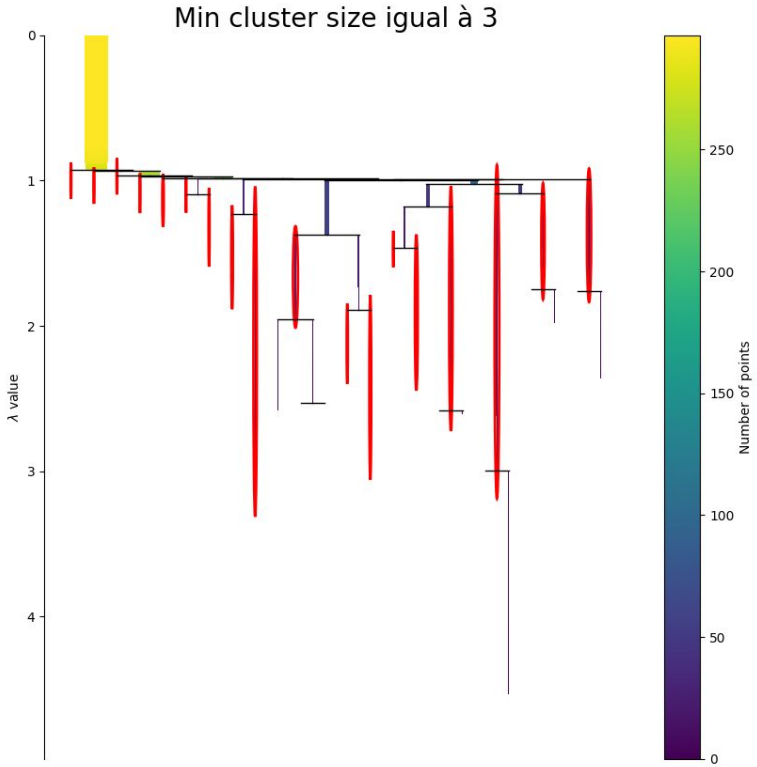

Score do HDBSCAN
Score Silhueta
0.25
0.20
0.15
0.10
0.05
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tamanho minimo por cluster


epsilon = 1.00
minPoints = 4
Restart
Pause

HDBSCAN, tamanho mínimo de cluster = 3

- Muitas ramificações
- Estrutura de cluster mais complexa nos dados
- Dados bastante heterogêneos e várias densidades diferentes

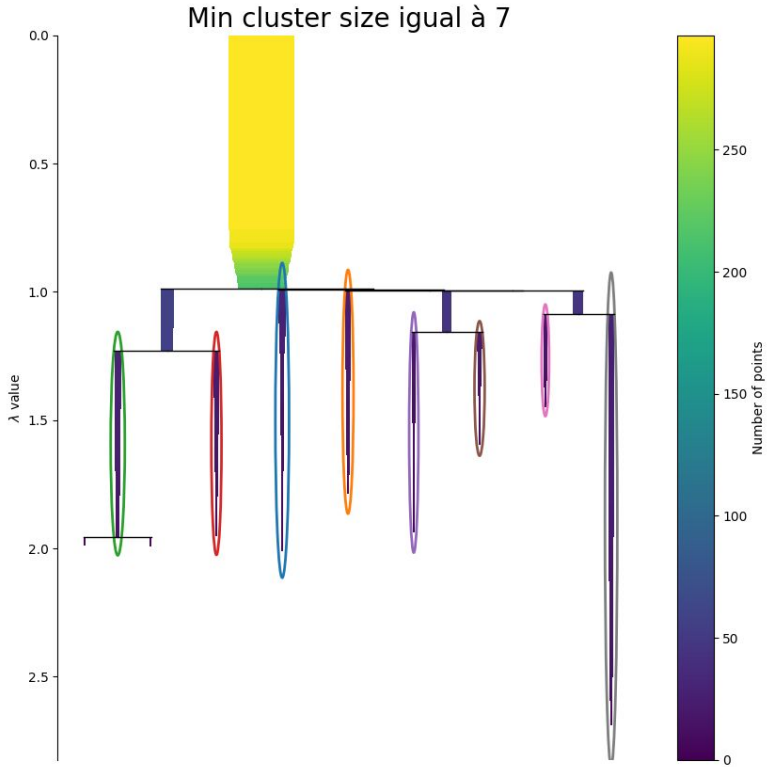
Silhouette Score = 0.24



HDBSCAN, tamanho mínimo do cluster = 7

- Menos ramificações
- Dados mais uniformes e menos subestruturas
- Clusters principais mais densos
- Sem a necessidade de muitas subdivisões

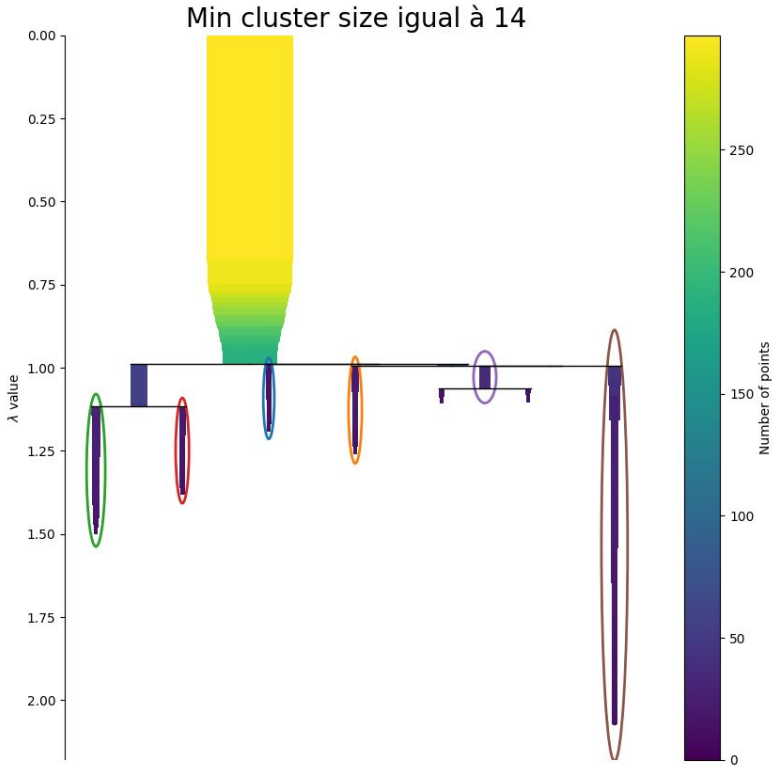
Silhouette Score = 0.16



HDBSCAN, tamanho mínimo do cluster = 14

- Clusters coesos e compactos
- Interessante para identificar grupos distintos e bem definidos
- Sem necessidade de subdivisão em mais clusters

Silhouette Score = 0.09



HDBSCAN

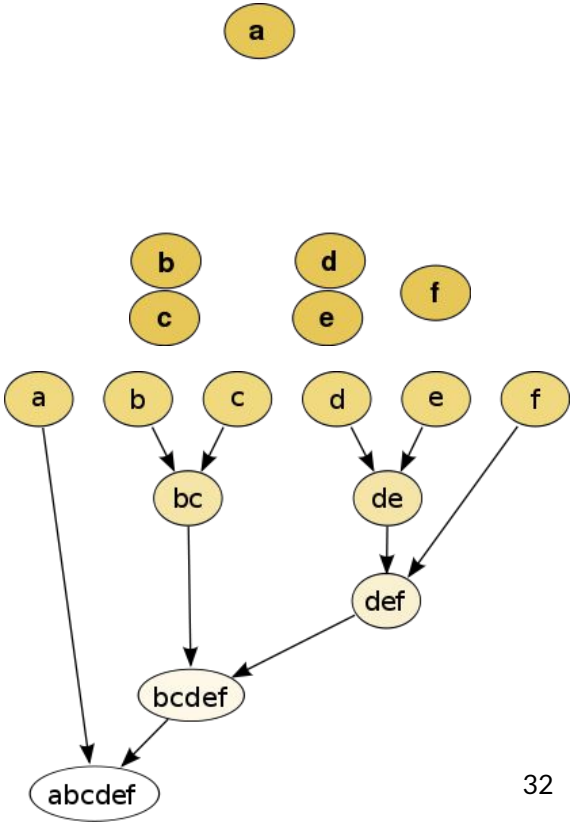
- O HDBSCAN não obteve resultados expressivos, mesmo com diferentes configurações de tamanho mínimo de grupos.

- A baixa quantidade de amostras pode ter contribuído para a falta de desempenho do algoritmo.

Aglomerativo

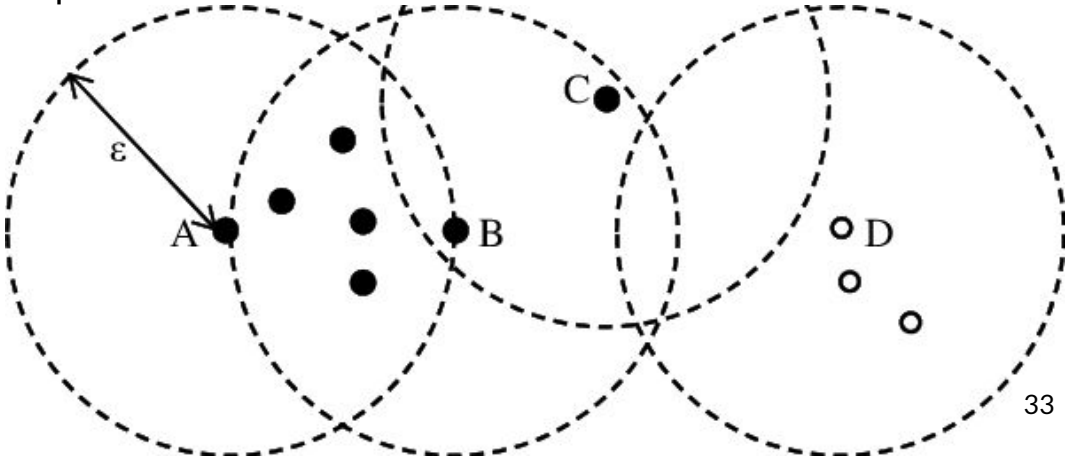
- Algoritmo aglomerativo é capaz de realizar um aninhamento de partições bottom up (começa a partir dos pontos únicos)
- A aglomeração ocorre conforme os grupos são mais próximos, caso haja empate os grupos serão conectados na mesma altura
- É bom para apresentar visualmente dados de dimensões maiores
- Sugere outliers.

As diferenças dos algoritmos está na forma que calcula a dissimilaridade de grupos já formados para decidir quais clusters devem ser combinados.



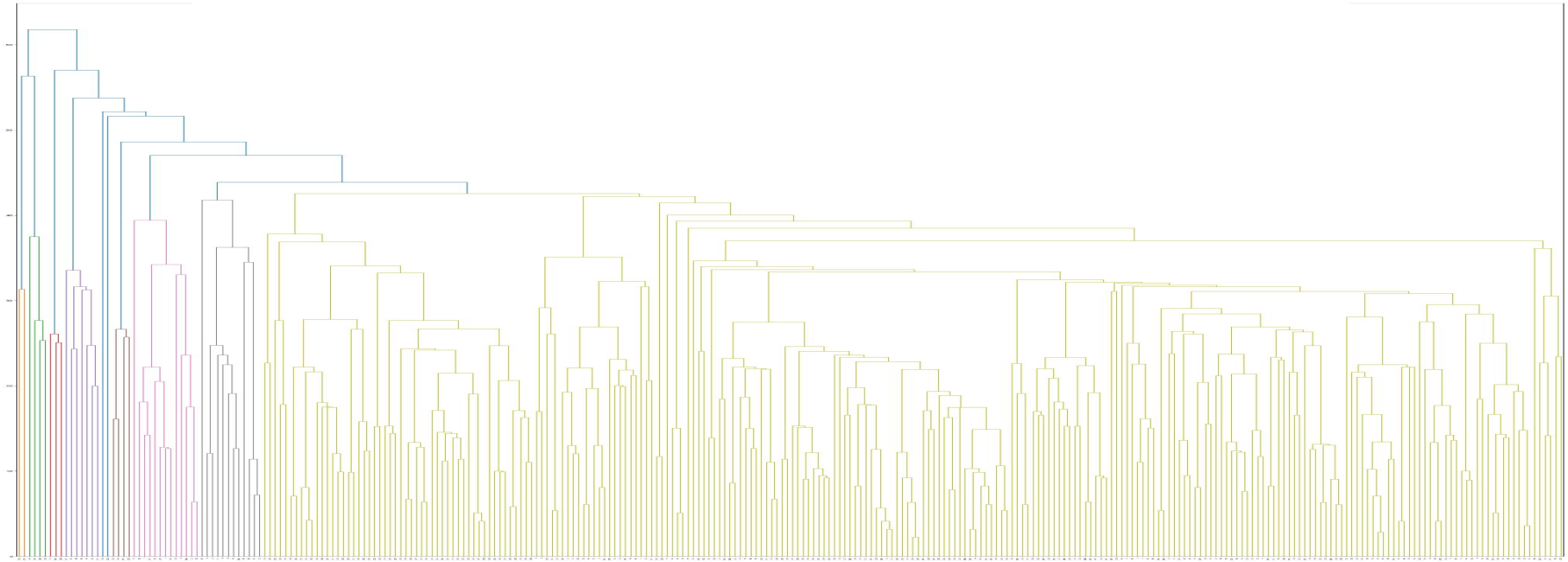
Aglomerativo - Ligação Simples

- Funciona com base na menor distância entre todos os pontos dos clusters.
- Este método tende a produzir clusters longos e finos nos quais os elementos próximos do mesmo cluster têm pequenas distâncias
- Os elementos nas extremidades opostas de um cluster podem estar muito mais distantes um do outro do que dois elementos de outros clusters.
- É sensível a ruídos contínuos



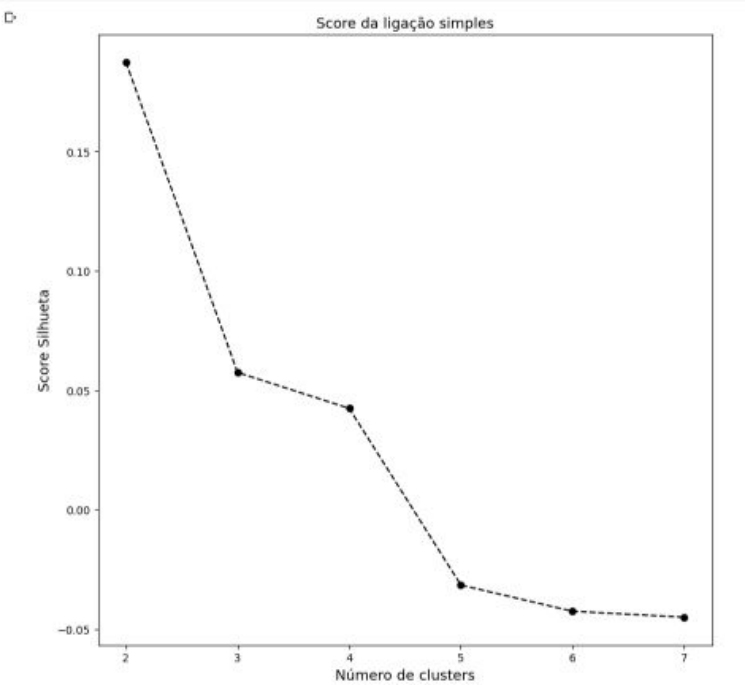
Ligação Simples - Com dados normalizados

Dendograma Ligação Simples, com n_cluster = 2 e linkage = single



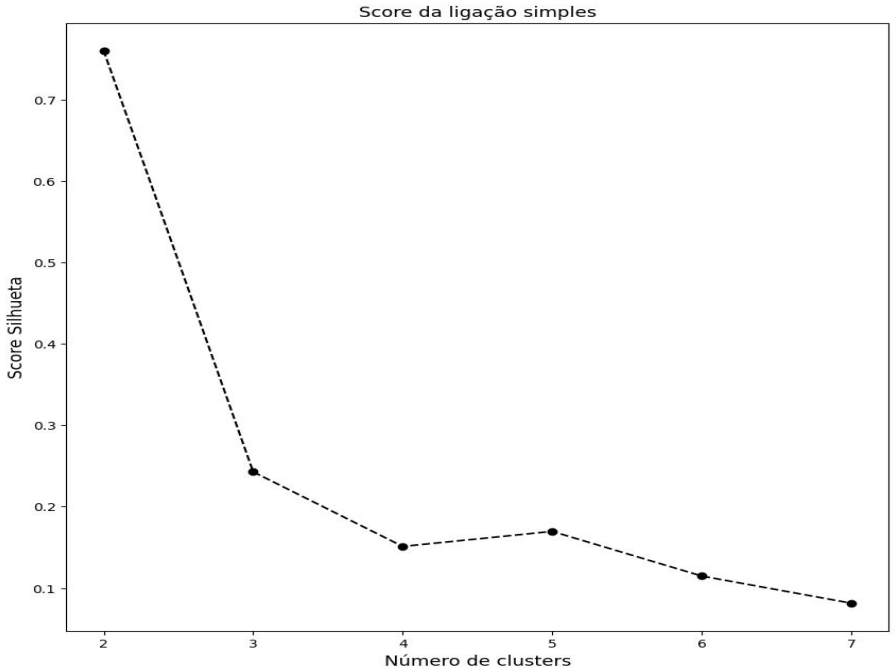
Ligação Simples - Com dados normalizados

- Obtivemos melhor Score de Silhueta de 0.18 para n_clusters = 2



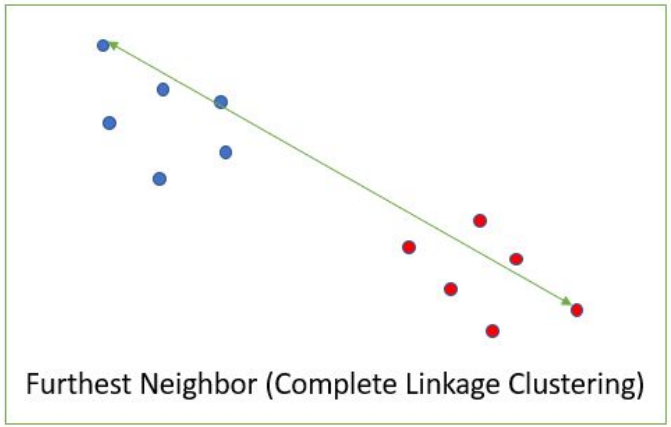
Ligação Simples - Dados não normalizados e c/ outliers

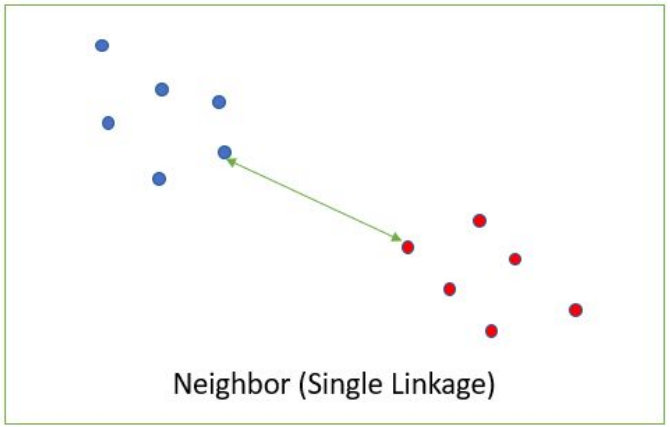
- Obtivemos melhor Score de Silhueta de 0.75 para n_clusters = 2



Aglomerativo - Ligação Completa

- Avalia a maior distância entre os pontos de dois grupos e escolhe o resultado menor desse tipo.
- É uma técnica mais robusta a ruído e outliers, pois são absorvidos logo no começo.



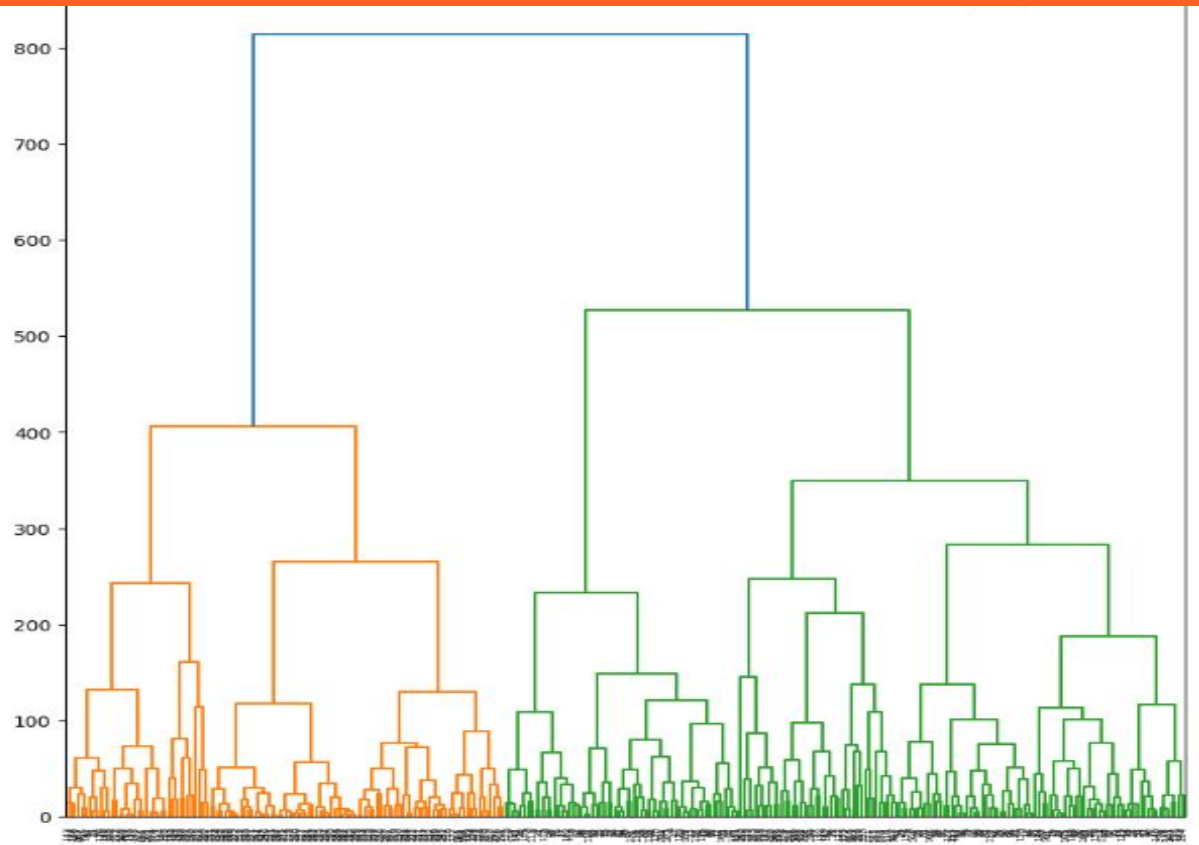


Ligação Completa - Dados não normalizados e c/ outliers

Dendograma Ligação Completa,

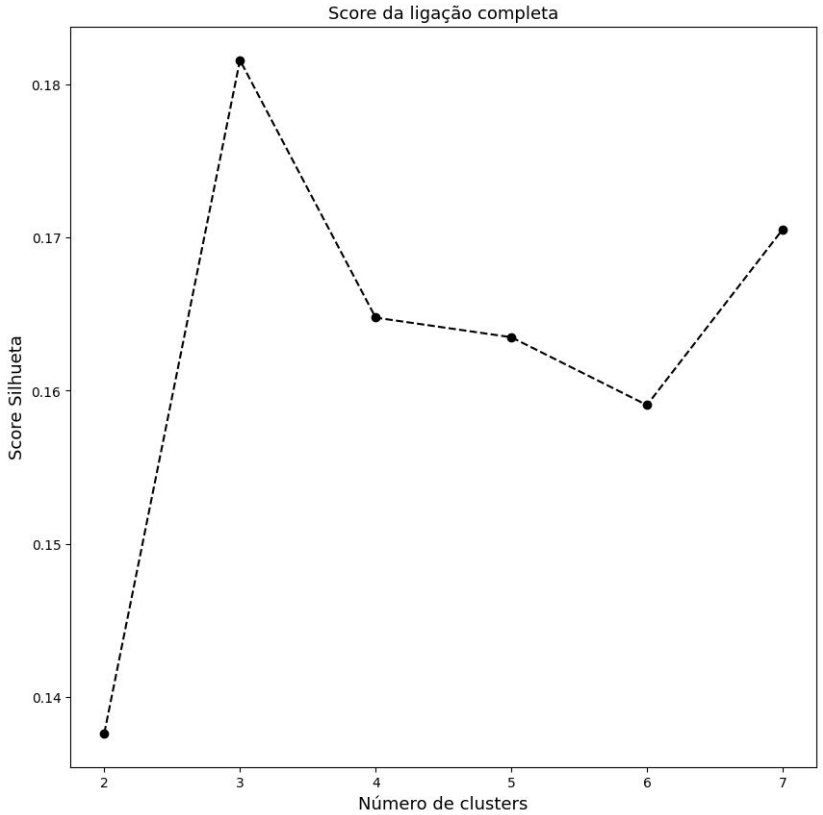
com n_cluster = 2

linkage = complete



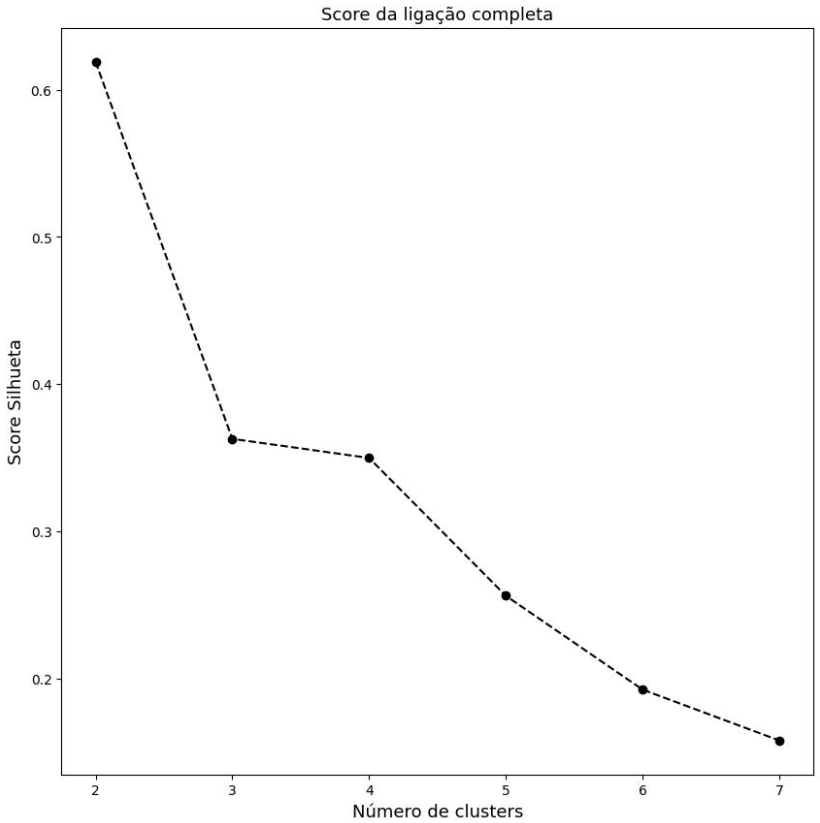
Ligação Completa - Com dados normalizados

- Obtivemos melhor Score de Silhueta de 0.182 para n_clusters = 3



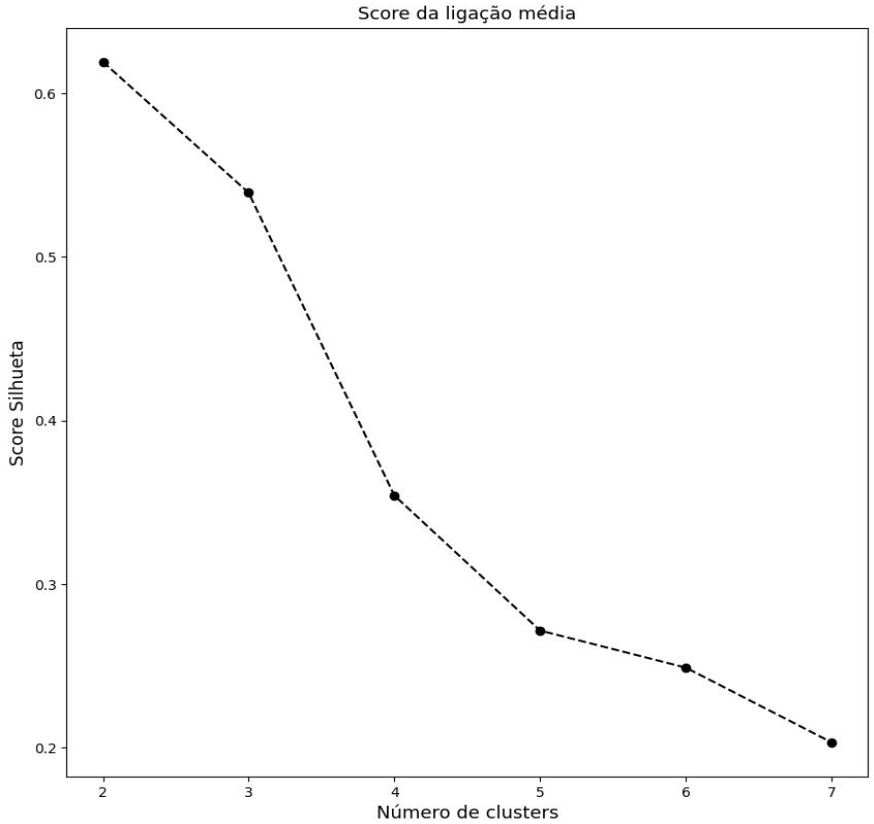
Ligação Completa - Dados não normalizados e c/ outliers

- Obtivemos melhor Score de Silhueta de 0.62 para n_clusters = 2



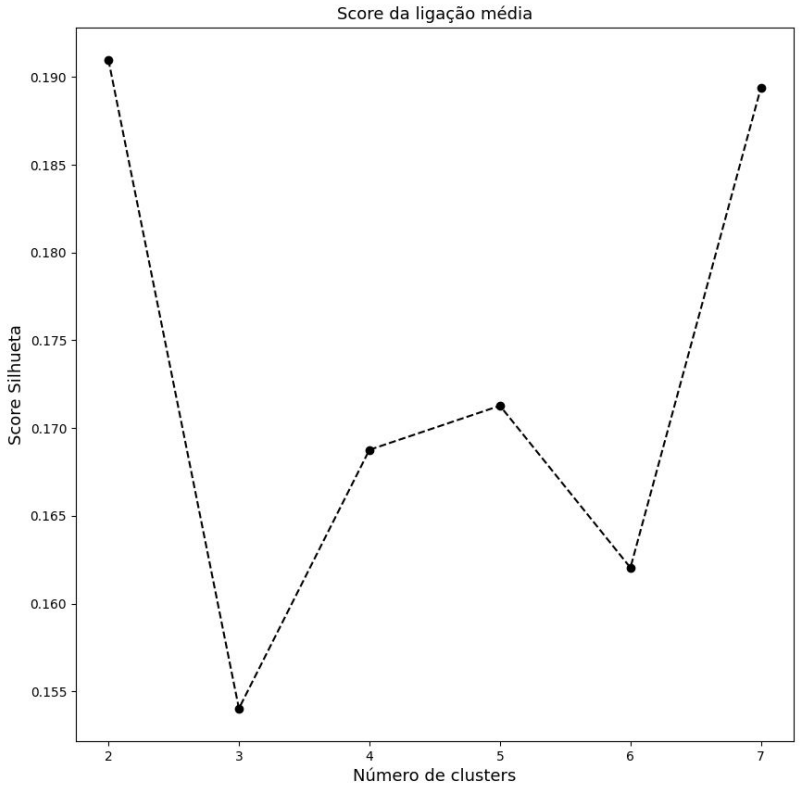
Ligação Média com dados não normalizados e c/ outliers

- A distância entre dois clusters é considerada a distância média entre os elementos de cada cluster.
- Resultados dados não normalizados e com outliers:
- Score de Silhueta de 0.62 para n_clusters = 2



Ligação Média - Com dados normalizados

- Obtivemos melhor Score de Silhueta de 0.19 para n_clusters = 2



Algoritmos de Agrupamento:

- O algoritmo aglomerativo se destacou ao considerar o índice de silhueta como métrica de validação.
- Diversas abordagens de ligação foram testadas, e a ligação simples obteve o melhor desempenho, seguida pelas ligações completa e média, que tiveram um desempenho igual para n_clusters = 2.
- Ligação completa apresentou um decaimento maior de score de silhueta para n_clusters = 3
- Esses resultados foram consistentes mesmo em cenários com presença de outliers e dados não normalizados, indicando a robustez do método aglomerativo.

Conclusão

Algoritmo Aglomerativo (melhor)

De acordo com o índice (*score*) de silhueta - o melhor método de agrupamento foi o **algoritmo aglomerativo**.

- Ligação simples: se destacou com 0.73 de *score*
  - nº cluster = 2
- Ligação completa: 0.619 de *score*
  - nº cluster = 2
- Ligação média: 0.619 de *score*
  - nº cluster = 2
- Resultados foram obtidos com o conjunto de dados contendo outliers e sem os dados estarem normalizados

K-Means

- Melhor *score* de silhueta: 0.65
  - PCA = 3 e nº clusters = 15
  - nº de grupos muito elevado para o conjunto de dados em questão, subdividindo muito as amostras e aparentemente sem resultados significativos para inferir sobre os grupos esperados (0-4).
- Outro resultado interessante: 0.59
  - PCA = 2 e nº clusters = 4
  - em que o número de grupos formados ficam mais coerentes com o esperado para análise

HDBScan

Não mostrou um desempenho expressivo para o agrupamento do nosso conjunto de dados.

Score de silhueta máximo: 0.24

- tamanho mínimo dos grupos = 2 e 18 grupos formados
- o que gerou uma estratificação demasiada dos dados

Conforme o tamanho mínimo de elementos dos grupos foi aumentando

- formação de cerca de 6 grupos
- em conformidade com o esperado, contudo apresentado índices de silhueta muito baixos e consequentemente uma baixa separação dos grupos formados

Obs.: resultados ruins podem ter sido causados pela baixa quantidade de amostras (elas podem não se apresentar em um formato denso e assim prejudicar seu desempenho)

Sobre os atributos

- Nosso dataset apresenta diversos atributos relevantes para formação de grupos de pacientes que podem apresentar falha cardíaca
  - Os métodos com melhores resultados apresentaram a formação de 2 grupos, muito próximo às interpretações do atributo alvo (“num”)
  - O atributo alvo possui 4 classificações, mas no nosso trabalho fizemos agrupamento por quem tinha ou não falha cardíaca
- Além da apuração obtida com boa pontuação para a formação de 4 grupos com k-means, o que pode-se explicar pela subdivisão do grupo dos pacientes com falha cardíaca.

Dúvidas?