Análise de Big Data via Machine Learning

LABDATA

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Machine Learning

Tema da Aula: Cluster

Prof. Anderson França

Introdução

O que significa fazer um agrupamento ?

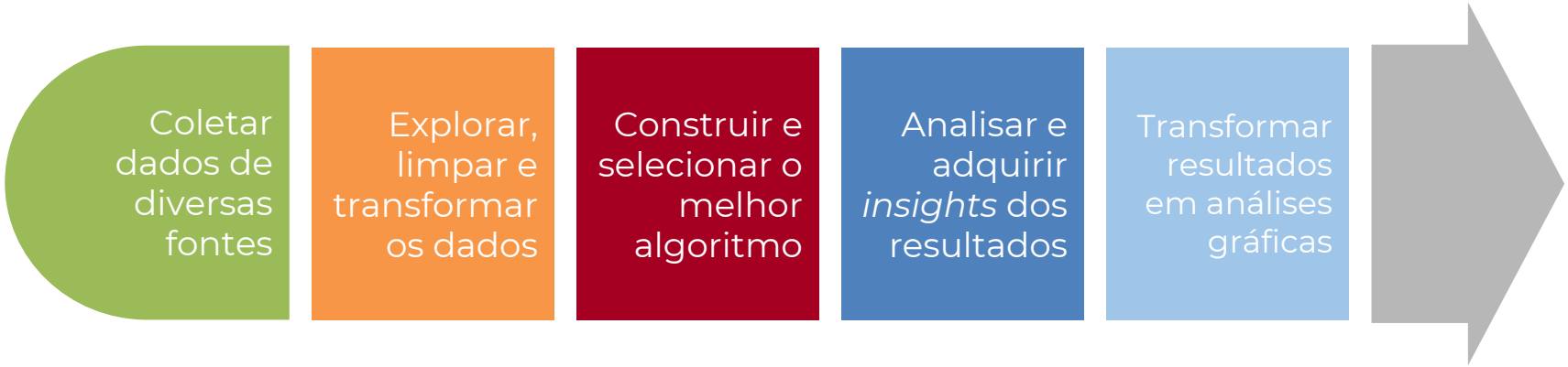
Por que fazer agrupamento ?

Quem já se deparou com uma situação em que seu Diretor de Marketing lhe diz:

"Ajude-me a entender melhor nossos clientes para que possamos comercializar nossos produtos para eles de uma maneira melhor!"

Machine Learning e Clustering

Aprendizado de Máquina (Machine Learning) é um campo de estudo que fornece a capacidade de uma Máquina de **entender dados** e **aprender com os dados**. O ML não é apenas sobre modelagem analítica, mas é uma modelagem de ponta a ponta que envolve as seguintes etapas:



Fonte: R-Bloggers

Objetivo

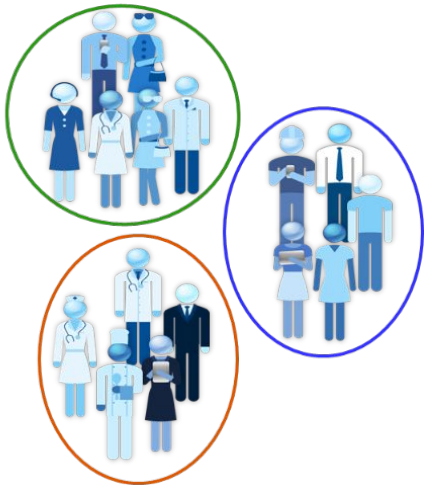
O objetivo da análise de cluster é agrupar as observações em grupos de tal forma que dentro de cada grupo as observações são semelhantes e distintas entre os grupos.

Dentro de cada grupo a variabilidade deve ser mínima e a variabilidade entre os grupos deve ser máxima.

Aplicações

Clustering tem um grande número de aplicações espalhados por vários áreas. Algumas das aplicações mais populares de clustering são:

- Sistemas de Recomendação
- Segmentação de Mercado
- Social network analysis
- Agrupamento de resultado de busca
- Imagens Médicas
- Segmentação de Imagens
- Detecção de Anomalias

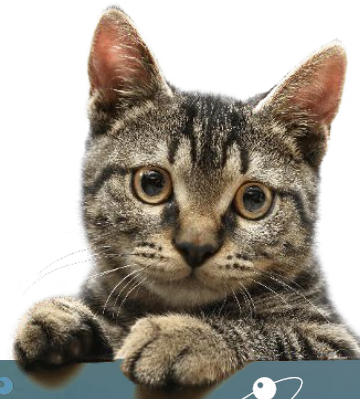


Medidas de Similaridade e Dissimilaridade

Na análise de cluster as observações são agrupadas de acordo com medidas de similaridade ou dissimilaridade.

Existem várias formas de medir similaridade ou dissimilaridade depende do critério a ser considerado.

A leoa é mais parecida com a gata ou com a cachorrinha?



A leoa é mais parecida com a gata ou com o cachorro?

Para determinar se a leoa é mais parecida com a gata ou com a cadela é necessário definir um critério de similaridade.

Considere como critério de similaridade o porte do animal. Neste caso a leoa será mais parecida com o cachorro

A leoa é mais parecida com a gata ou com o cachorro?

Considere agora como critério de similaridade o formato do fucinho. Neste caso a leoa será mais parecida com a gata.

Medidas de Similaridade: Quanto maior for a medida de similaridade maior será a semelhança entre os elementos. O coeficiente de correlação linear de Pearson é uma medida de similaridade.

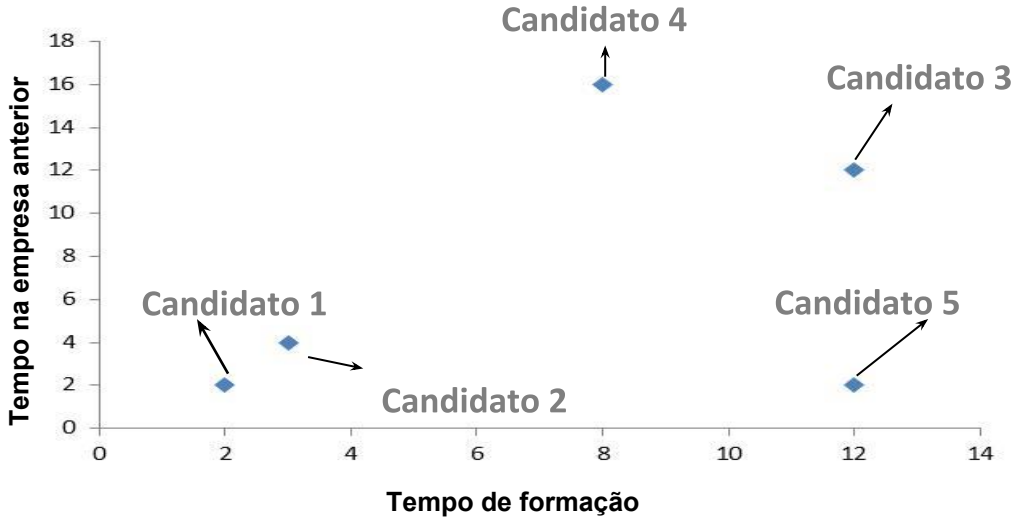
Medidas de Dissimilaridade: Quanto maior for a medida de dissimilaridade menor será a semelhança entre os elementos. A distância euclidiana e a distância euclidiana ao quadrado são medidas de dissimilaridade.

Exemplo 1

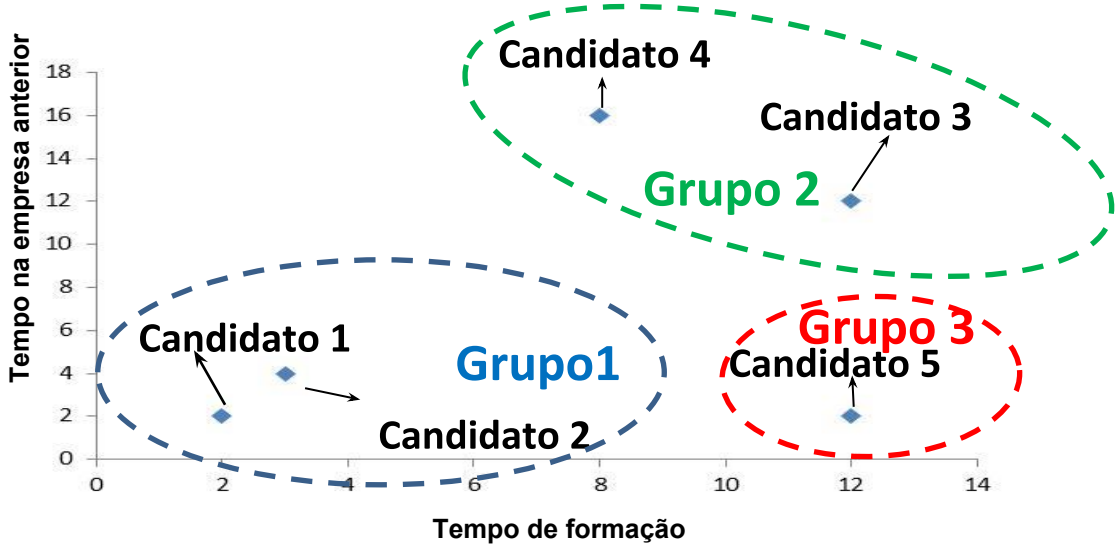
Considere o exemplo de uma analista de gestão de pessoas que deseja agrupar os candidatos em três grupos considerando duas variáveis: o **tempo de formação do candidato** e o **tempo que o candidato permaneceu na empresa anterior**. A Tabela apresenta os valores das variáveis para os cinco candidatos.

Candidato	Tempo de Formação	Tempo Empresa Anterior
1	2	2
2	3	4
3	12	12
4	8	16
5	12	2

O Gráfico de dispersão apresenta os valores das variáveis para os cinco candidatos.



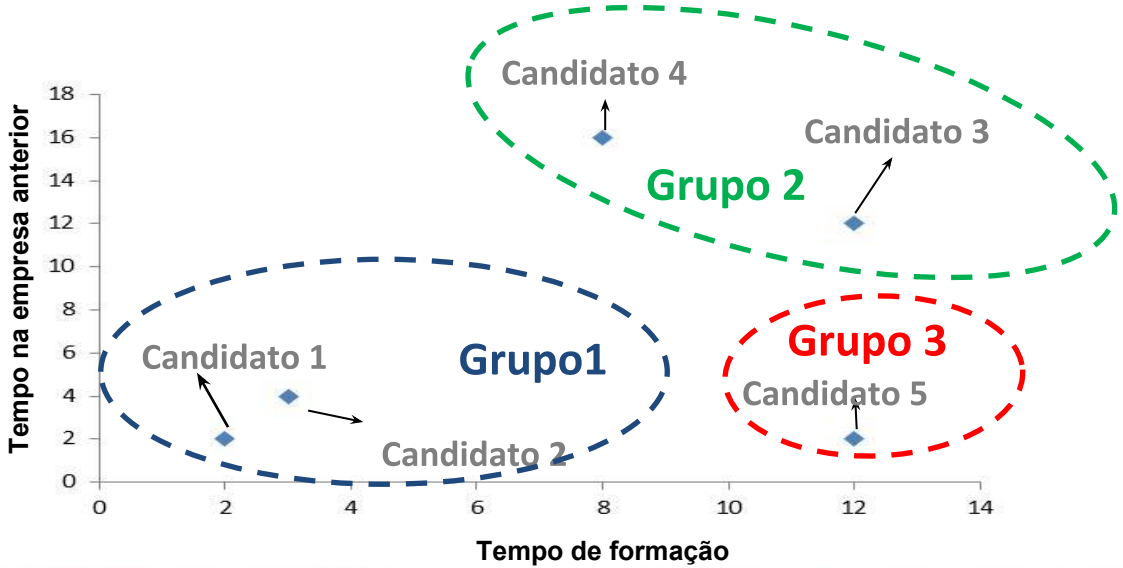
Como a analista de gestão de pessoas deseja agrupar os candidatos em três grupos, considerando duas variáveis, o gráfico apresenta uma sugestão de agrupamento. Os candidatos foram agrupados de acordo com um critério.



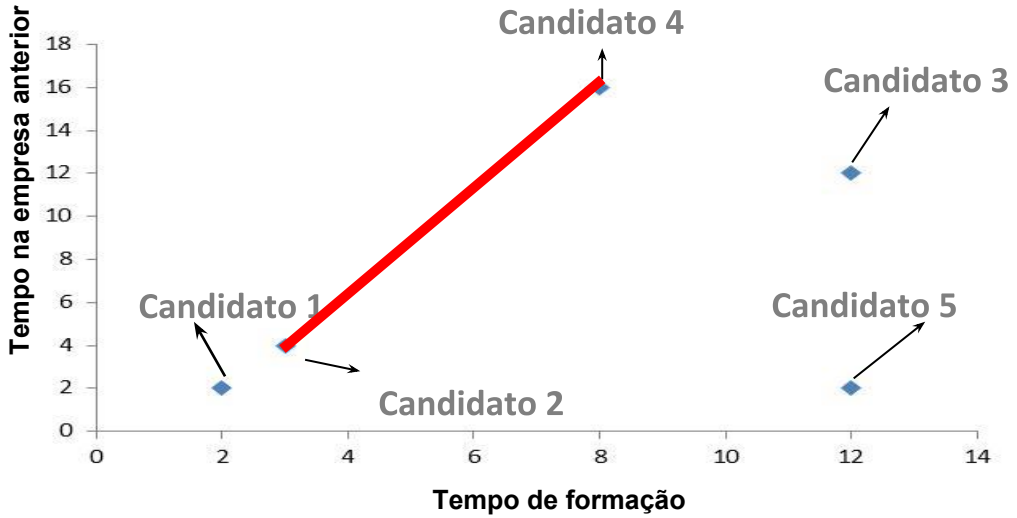
Grupo 1 → formado por candidatos com pouco tempo de formação e pouco tempo na empresa anterior.

Grupo 2 → formado por candidatos com tempo de formação superior a 7 anos e com tempo na empresa anterior superior a 11 anos.

Grupo 3 → formado por um candidato com 12 anos de formação e 2 anos na empresa anterior.



Um critério de **dissimilaridade** que pode ser considerado para agrupar observações é a **distância Euclidiana**. A distância Euclidiana entre os candidatos 2 e 4 é dada pela reta vermelha.



A **distância Euclidiana ao Quadrado** entre os candidatos 2 e 4 é dada por:

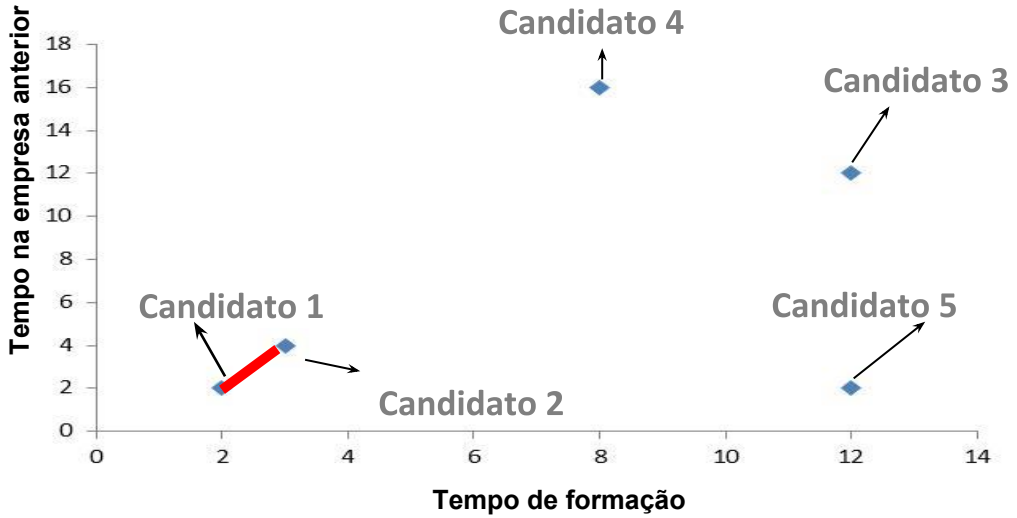
$$D^2 = (8-3)^2 + (16-4)^2 = 5^2 + 12^2 = 169$$

Candidato	Tempo de Formação	Tempo Empresa Anterior
1	2	2
2	3	4
3	12	12
4	8	16
5	12	2

A **distância Euclidiana** entre os candidatos 2 e 4 é obtida por meio da raiz quadrada positiva da distância Euclidiana ao Quadrado .

$$\mathrm{D} = \sqrt{169} = 13$$

A distância Euclidiana entre os candidatos 1 e 2 é dada pela reta vermelha.



A **distância Euclidiana ao Quadrado** entre os candidatos 1 e 2 é dada por:

$$D^2 = (3-2)^2 + (4-2)^2 = 1^2 + 2^2 = 5$$

Candidato	Tempo de Formação	Tempo Empresa Anterior
1	2	2
2	3	4
3	12	12
4	8	16
5	12	2

A **distância Euclidiana** entre os candidatos 1 e 2 é obtida por meio da raiz quadrada positiva da distância Euclidiana ao Quadrado.

$$D = \sqrt{5} = 2{,}23$$

A matriz de distância **Euclidiana ao Quadrado** é uma matriz simétrica.

As distâncias Euclidianas ao Quadrado, entre todos os elementos, localizadas acima da diagonal principal são apresentadas na matriz.

	1	2	3	4	5
1		5	200	232	100
2			145	169	85
3				32	100
4					212
5					

A **distância Euclidiana** é obtida por meio da raiz quadrada da distância Euclidiana ao quadrado.

A matriz de **distância Euclidiana** é uma matriz simétrica.

As distâncias Euclidianas, entre todos os elementos, localizadas acima da diagonal principal são apresentadas na matriz.

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Métodos de Agrupamento

Métodos de Agrupamento

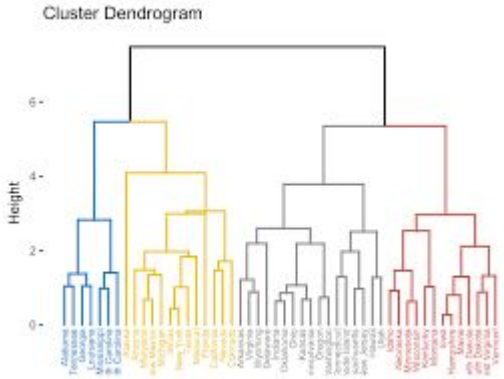
De um modo geral, o agrupamento pode ser dividido em dois subgrupos:

- **Hard Clustering:** No agrupamento, cada ponto de dados ou pertence a um cluster completamente ou não. Por exemplo, cada cliente é colocado em um grupo dos 10 grupos.
- **Soft Clustering:** Em vez de colocar cada ponto de dados em um cluster separado, uma probabilidade ou verossimilhança de que o ponto de dados esteja nesses clusters é atribuída. Por exemplo, a cada cliente é atribuída uma probabilidade de estar em qualquer um dos 10 clusters da loja de varejo.

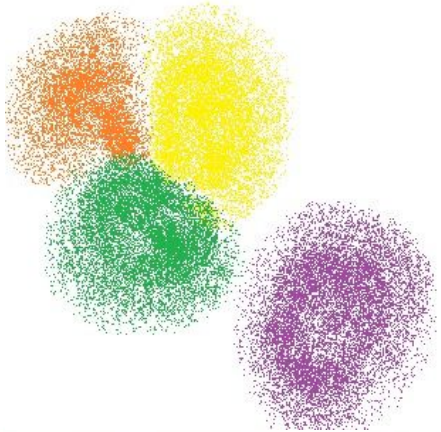
Métodos de Agrupamentos

De fato, existem mais de 100 algoritmos de clustering conhecidos. Mas poucos algoritmos são usados popularmente, vamos destacar os dois principais:

Método Hierárquico



Método das K médias



Método Hierárquico

Técnicas de Agrupamento

Vizinho mais Próximo

(Nearest Neighbor)

Vizinho mais Próximo (Nearest Neighbor)

Para a realização dos agrupamentos considerando como técnica de agrupamento o vizinho mais próximo pode-se partir da matriz de distância Euclidiana entre todos os candidatos.

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1

a) Agrupar as observações com a menor distância:

Como as observações 1 e 2 possuem as menores distâncias, elas serão agrupadas no primeiro passo

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1

b) Calcular a distância entre as observações 1 e 2 e as demais observações:

Distância entre 1 e 3 = 14,14

Distância entre 2 e 3 = 12,04

A **menor** distância é 12,04.

Distância entre 1 e 4 = 15,23

Distância entre 2 e 4 = 13,0

A **menor** distância é 13,0.

Distância entre 1 e 5 = 10,0

Distância entre 2 e 5 = 9,22

A **menor** distância é 9,22.

c) Elaborar uma nova matriz de distância com as observações 1 e 2 grupadas:

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

	1+2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 2

a) Agrupar as observações com a menor distância:

Como as observações 3 e 4 possuem as menores distâncias, elas serão agrupadas no passo 2

	1+2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 2

b) Calcular a distância entre as observações 3 e 4 e as demais observações:

Distância entre 3 e (1+2) = 12,04

Distância entre 4 e (1+2) = 13,00

A **menor** distância é 12,04.

Distância entre 3 e 5 = 10,00

Distância entre 4 e 5 = 14,56

A **menor** distância é 10,00

c) Elaborar uma nova matriz de distância com as observações 3 e 4 agrupadas:

	1+2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

	1+2	3+4	5
1 + 2		12,04	9,22
3+4			10,00
5			

Passo 3

a) Agrupar as observações com a menor distância:

Como as observações (1+2) e 5 possuem as menores distâncias, elas serão agrupadas no passo 3

	1+2	3+4	5
1 + 2		12,04	9,22
3+4			10,00
5			

Passo 3

b) Calcular a distância entre as observações (1+2) e 5 e as demais observações:

Distância entre (1+2) e (3+4)= 12,04

Distância entre 5 e (3+4) = 10,00

A **menor** distância é 10,00.

c) Elaborar uma nova matriz de distância com as observações (1+2) e 5 agrupadas:

	1+2	3+4	5
1 + 2		12,04	9,22
3+4			10,00
5			

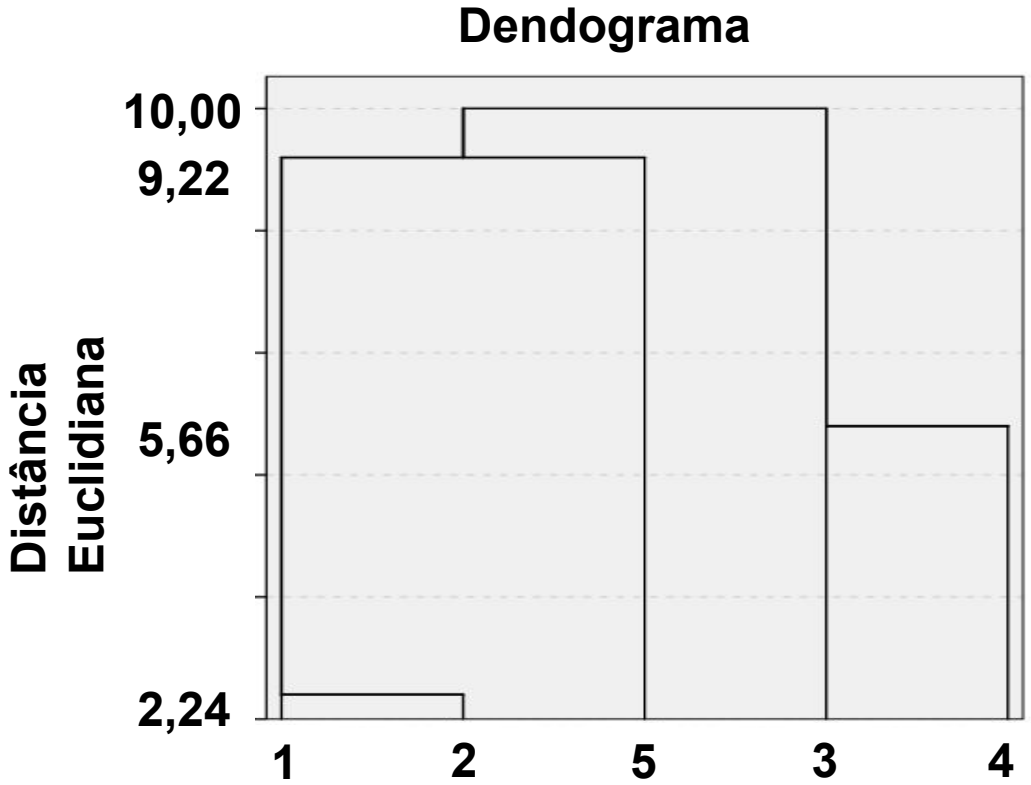
	1+2+5	3+4
1 + 2 + 5		10,00
3+4		

Dendrograma

O dendograma é um gráfico que tem como objetivo representar graficamente os passos realizados em um agrupamento feito por um método hierárquico.

Com base na análise do dendrograma é possível determinar o número de grupos para o conjunto de observações.

Este é o Dendograma gerado a partir dos agrupamentos realizados nos passos de 1 a 3

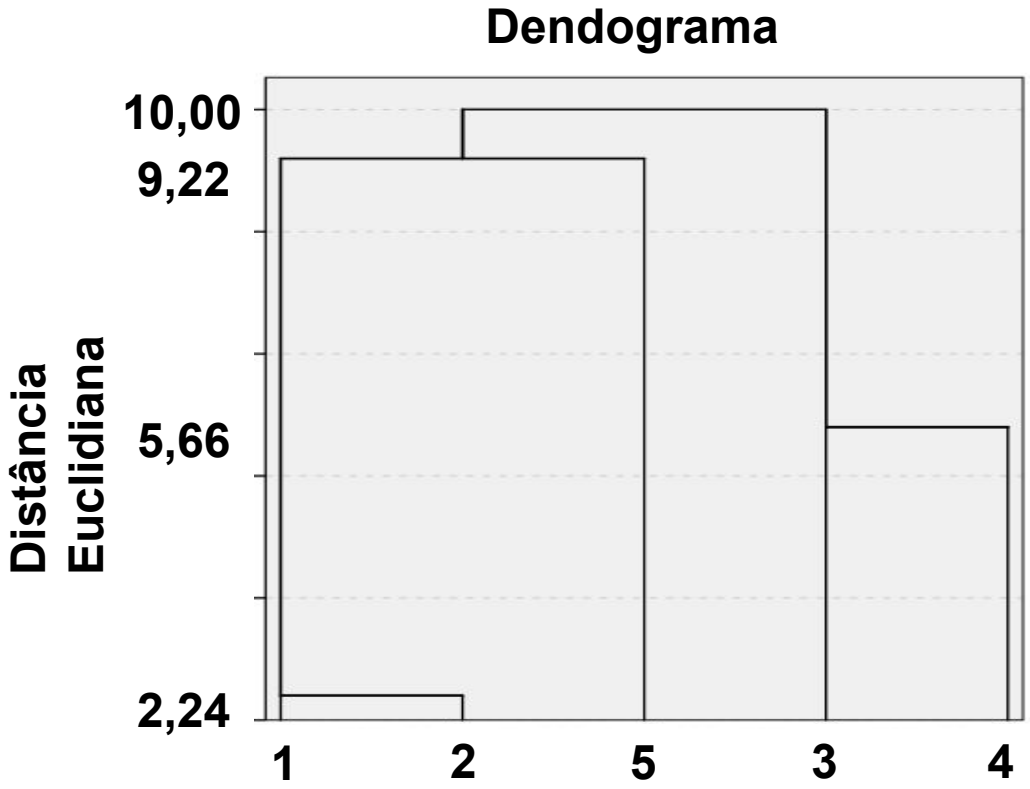


O elemento 1 foi agrupado ao elemento 2 na distância 2,24.

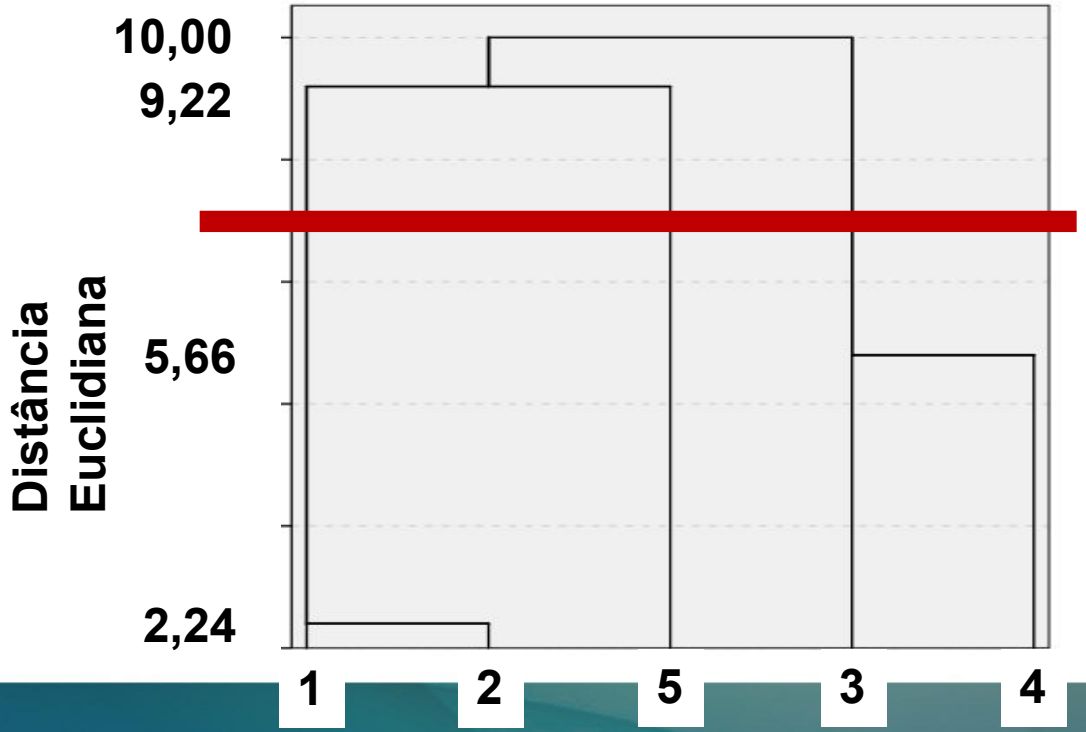
O elemento 3 foi agrupado ao elemento 4 na distância 5,66.

O grupo (1+2) foi agrupado ao elemento 5 na distância 9,22.

O grupo (1+2+5) foi agrupado ao grupo (3+4) na distância 10,00.

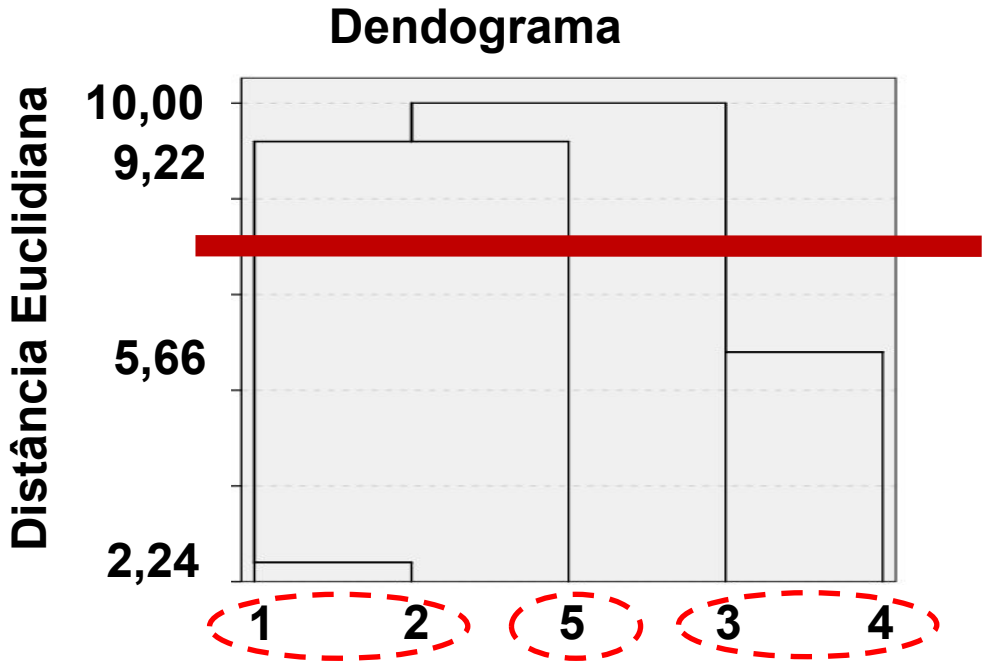


Por meio do dendograma pode-se sugerir o número de grupos a serem considerados. Em geral, observa-se quando o próximo agrupamento é realizado em uma distância muito superior ao agrupamento anterior.



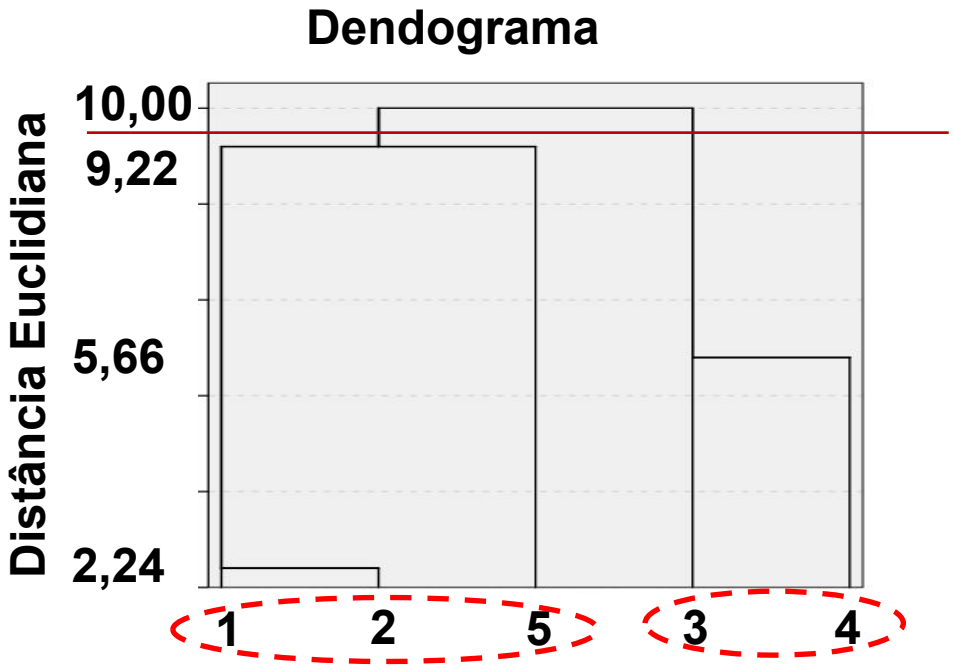
Como a distância entre 9,22 e 5,66 é grande, sugere-se separar os grupos em uma distância superior a 5,66 e inferior a 9,22

Considerando a linha vermelha como a separação dos grupos, nota-se que os elementos 1 e 2 formam um grupo, o elemento 5 forma um grupo e os elementos 3 e 4 formam um grupo.



Caso o objetivo do problema seja separar os elementos em 2 grupos, pode-se considerar a linha vermelha como a separação.

Nota-se que os elementos 1, 2 e 5 formam um grupo e os elementos 3 e 4 formam o outro grupo.



Exemplo

Banco de Dados: MCDONALDS.xls

Neste exemplo, pretende-se agrupar os lanches do Mcdonalds de acordo com as variáveis apresentadas.

	Valor Energético	Carboidratos	Proteínas	Gorduras Totais	Gorduras Saturadas	Gorduras Trans	Colesterol	Fibra Alimentar	Sódio	Calcio	Ferro
Angus Deluxe	863	56	51	49	21	1,7	139	5	1716	197	3,8
Angus Bacon	861	57	54	46	21	1,7	145	5	1917	193	4
Big Tasty	843	45	41	55	24	1,7	104	5,1	1511	381	8,2
CBO	643	56	27	35	11	0,7	71	4	1220	236	6,9
Mcnifico Bacon	625	38	34	37	16	1,3	95	3,9	1255	209	11
Chicken Club Crispy	610	54	31	30	9,3	0,5	58	6,7	1831	115	3
Quarterão	558	36	31	32	16	1,3	86	3,6	1216	275	10
Chicken Club Grill	545	46	41	22	7	0,4	93	5,5	1542	121	3
Chicken Bacon Crispy	515	45	27	25	10	0,4	47	3	1487	165	4,4
Cheddar McMelt	507	33	29	29	14	1,2	81	2,9	819	199	10
Big Mac	504	41	25	27	12	0,5	54	3,5	960	162	6,5
Chicken Classic Crispy	490	52	24	20	4,4	0,2	33	6,7	1361	38	2,5
McChicken	454	40	18	25	6	0,3	49	3,1	1068	67	4
Chicken Lemon Crispy	454	54	24	16	7	0,2	28	6,7	1290	37	2,5
Chicken Classic Grill	425	44	34	12	2	0,2	68	5,5	1072	44	2,6
Chicken Bacon Grill	425	31	39	16	5	0,4	83	3	1151	161	3,9
Chicken Lemon Grill	389	46	34	7,8	0,8	0,2	63	5,5	1001	43	2,5
McFish	373	38	18	16	5,8	0,4	43	2,2	773	168	3,8
Wrap Crispy Maionese	372	33	14	20	4,6	0,2	25	1,8	1032	172	0,81
Wrap Grill Maionese	340	29	19	16	3,4	0,1	42	1,2	888	175	0,83
Wrap Crispy Lemon	338	34	14	16	3,3	0,1	21	1,8	974	172	0,77
McChicken Jr	337	33	13	17	4,5	0,4	11	3,3	633	58	3,1
Cheeseburger	310	32	15	14	7	0,4	31	1,9	781	146	4,2
Wrap Grill Lemon	306	30	19	12	2,1	0,1	38	1,1	829	175	0,78
Hamburger	257	31	13	9,3	4,1	0,3	22	1,9	542	70	4,2

Para padronizar uma variável deve-se subtrair da variável original o valor da média e dividir o resultado pelo desvio padrão.

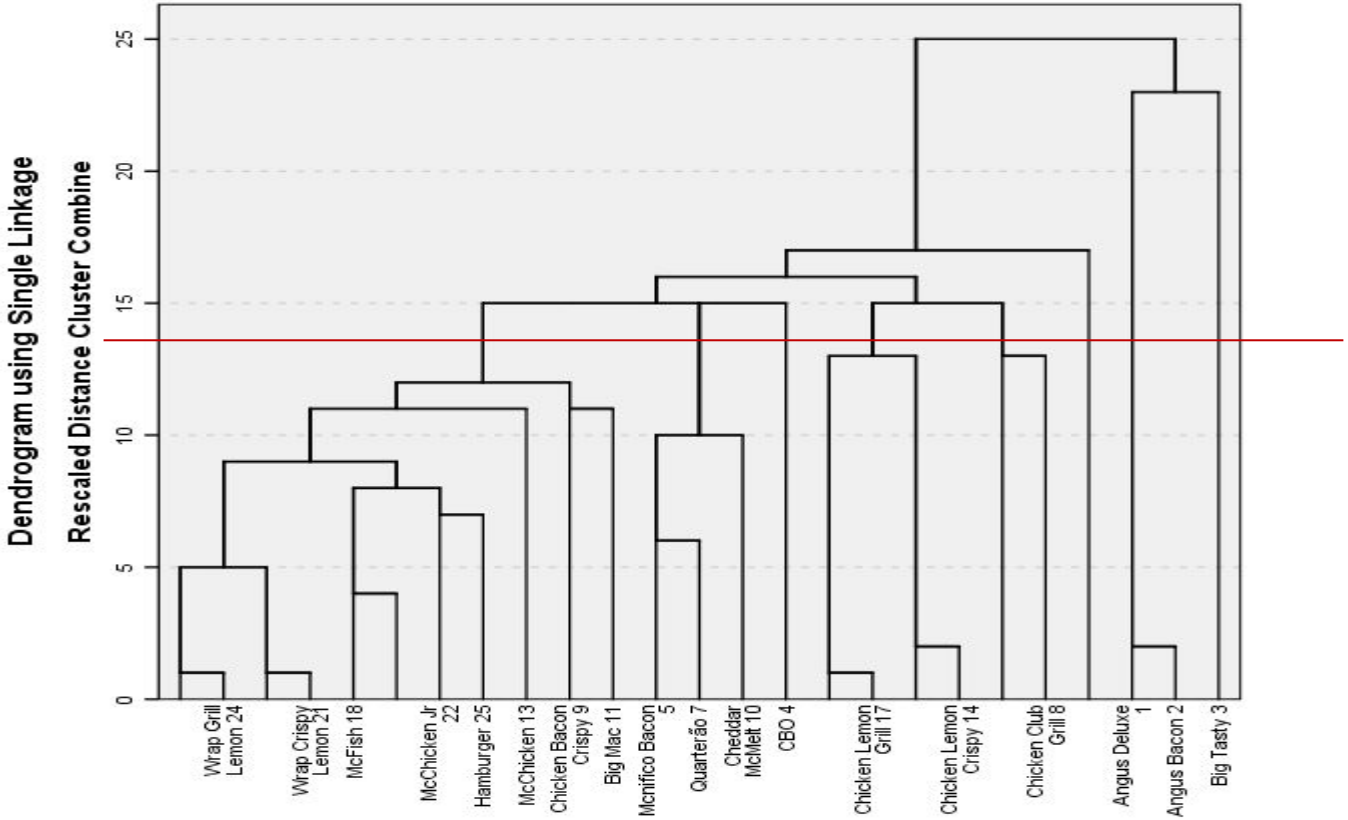
A variável padronizada é denominada Z.

$$Z = \frac{(X - \overline{X})}{S}$$

X: variável aleatória com média $\overline{X}$ e desvio padrão S

Z: variável aleatória padronizada com média 0 e variância 1.

Dendrograma



Esta tabela gerada faz uma comparação das médias das variáveis entre os grupos.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: Valor Energético	10,640	2	,124	22	86,075	,000
Zscore(Carboidratos)	3,559	2	,767	22	4,638	,021
Zscore(Proteínas)	7,355	2	,422	22	17,420	,000
Zscore: Gorduras Totais	10,482	2	,138	22	75,939	,000
Zscore: Gorduras Saturadas	10,528	2	,134	22	78,672	,000
Zscore: Gorduras Trans	9,686	2	,210	22	46,053	,000
Zscore(Colesterol)	8,556	2	,313	22	27,327	,000
Zscore: Fibra Alimentar	1,547	2	,950	22	1,628	,219
Zscore(Sódio)	6,209	2	,526	22	11,793	,000
Zscore(Calcio)	5,048	2	,632	22	7,987	,002
Zscore(Ferro)	5,609	2	,581	22	9,654	,001

Teste F para comparação de médias

H_0: as médias da variável são iguais para todos os grupos;

H_1: as médias da variável são diferentes em pelo menos um grupo;

Zscore:Valor energético = Valor energético padronizada (com média zero e desvio padrão 1).

A hipótese testa se a média dessa variável para o **grupo 1** é igual a média dessa variável para o **grupo 2** e é igual a média dessa variável para o **grupo 3**.

Teste F para comparação de médias

H_0: as médias da variável são iguais para todos os grupos;

H_1: as médias da variável são diferentes em pelo menos um grupo;

Regra de decisão: Quando o Sig (Nível descritivo do teste) for menor do que α (0,10) rejeitamos H_0, ou seja, há evidência de que as médias da variável são diferentes em pelo menos um grupo

Como o Sig associado à variável Zscore: Fibra Alimentar é maior do que 0.10, há evidência de que as médias dessa variável são iguais para todos os grupos. Desta forma, esta variável não é importante para a formação dos grupos. Todas as outras variáveis são importantes.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: Valor Energético	10,640	2	,124	22	86,075	,000
Zscore(Carboidratos)	3,559	2	,767	22	4,638	,021
Zscore(Proteínas)	7,355	2	,422	22	17,420	,000
Zscore: Gorduras Totais	10,482	2	,138	22	75,939	,000
Zscore: Gorduras Saturadas	10,528	2	,134	22	78,672	,000
Zscore: Gorduras Trans	9,686	2	,210	22	46,053	,000
Zscore(Colesterol)	8,556	2	,313	22	27,327	,000
Zscore: Fibra Alimentar	1,547	2	,950	22	1,628	,219
Zscore(Sódio)	6,209	2	,526	22	11,793	,000
Zscore(Calcio)	5,048	2	,632	22	7,987	,002
Zscore(Ferro)	5,609	2	,581	22	9,654	,001

Ajusta-se novamente a análise de cluster pelo método das k médias sem a variável Zscore: Fibra Alimentar.

Como o Sig associado as variáveis são inferiores a 0.10, todas essas variáveis são importantes.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: Valor Energético	9,276	2	,248	22	37,464	,000
Zscore(Carboidratos)	2,546	2	,859	22	2,962	,073
Zscore(Proteínas)	6,186	2	,529	22	11,705	,000
Zscore: Gorduras Totais	9,909	2	,190	22	52,125	,000
Zscore: Gorduras Saturadas	10,503	2	,136	22	77,173	,000
Zscore: Gorduras Trans	10,580	2	,129	22	81,961	,000
Zscore(Colesterol)	7,977	2	,366	22	21,811	,000
Zscore(Sódio)	4,056	2	,722	22	5,617	,011
Zscore(Calcio)	5,887	2	,556	22	10,593	,001
Zscore(Ferro)	8,733	2	,297	22	29,401	,000

Cluster Hierarquico - R

Cluster K-means - R

Obrigado!

Anderson França
Email: anderson.frca@gmail.com
LinkedIn: /andersonfranca1/