THE DEVELOPER'S CONFERENCE

Trilha – Learning Machine Cluster Analysis em 4 passos

Marco Siqueira Campos

Globalcode

THE DEVELOPER'S CONFERENCE

Marco Siqueira Campos

Sócio fundador Siqueira Campos Associados e sos-stat

SIQUEIRA CAMPOS

sos stat

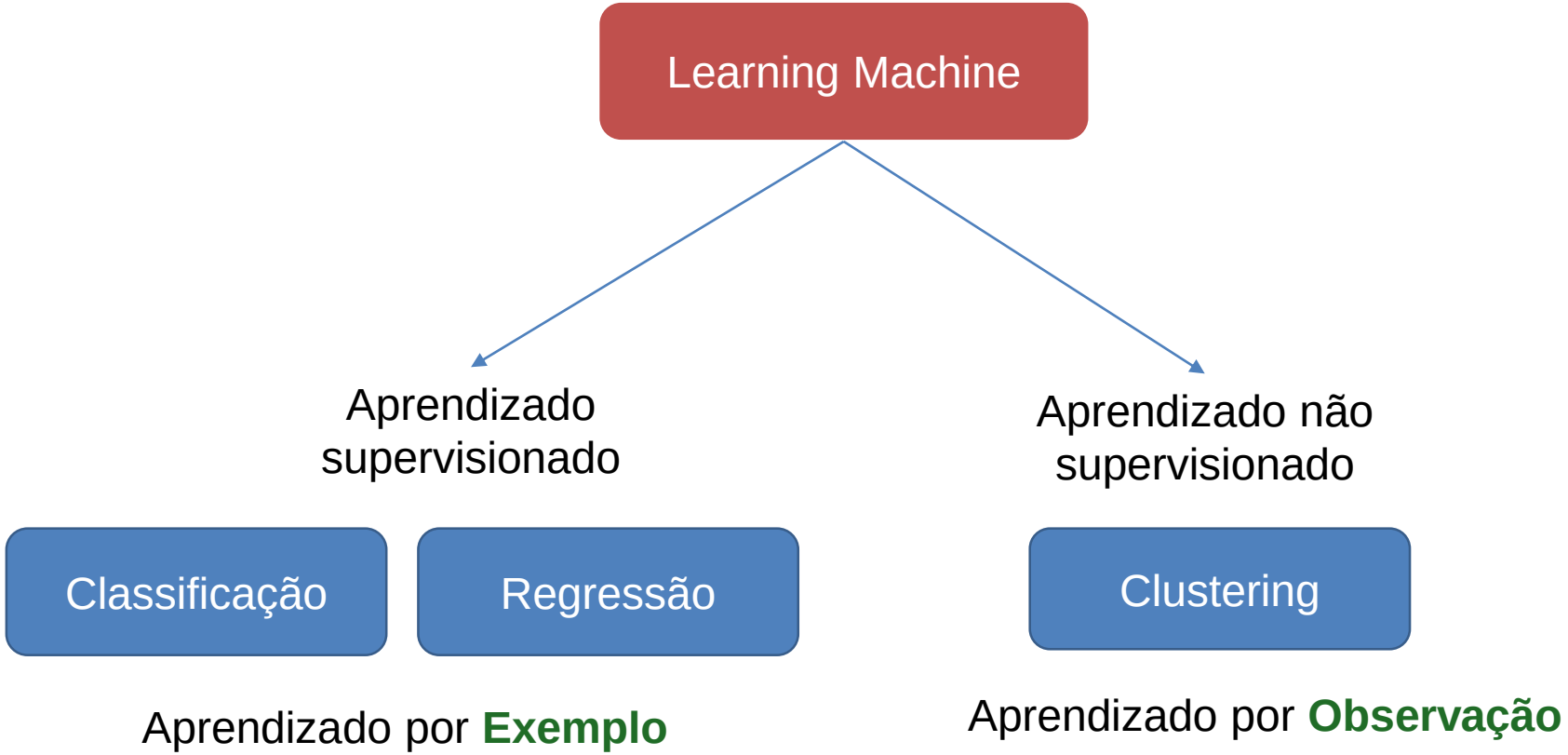
Estatístico – UFRGS

Certificado Data Science Specialization – Johns Hopkins-Coursera

Mestrando Data Science CUNY The City University of New York

Globalcode

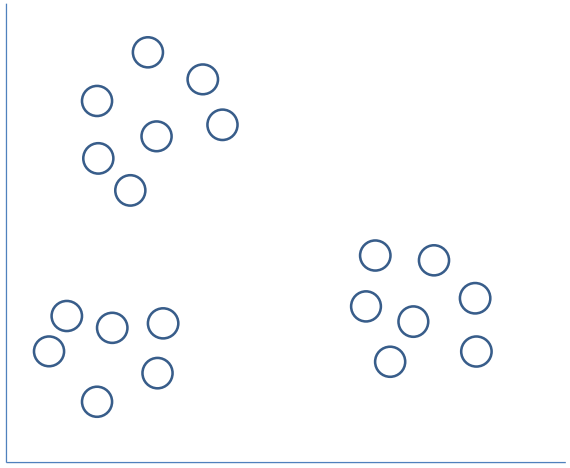
Modelo


Learning Machine
Aprendizado supervisionado
Aprendizado não supervisionado
Classificação
Regressão
Clustering
Aprendizado por **Exemplo**
Aprendizado por **Observação**

O que é clustering?

THE DEVELOPER'S CONFERENCE

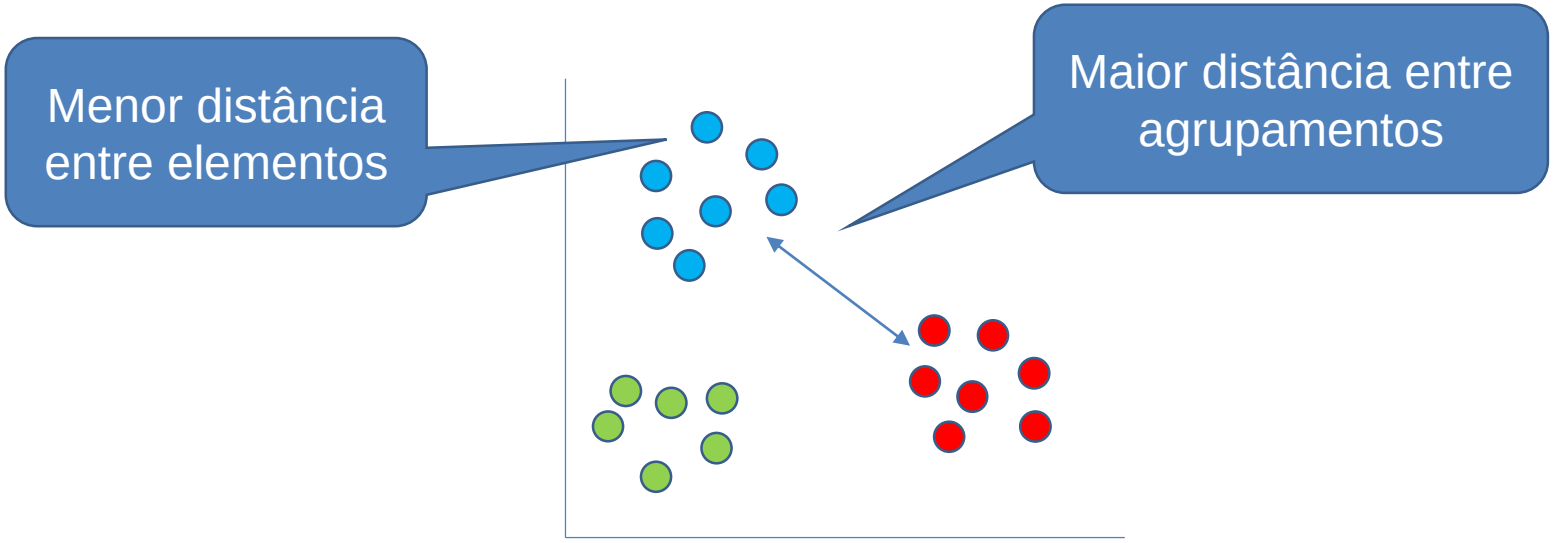
É uma técnica estatística multivariada para identificar agrupamentos dos dados de acordo com o grau de semelhança. Queremos achar um grupo de objetos, que são similares entre si e diferentes de outros grupos.



Globalcode

O que é clustering?

É uma técnica estatística multivariada para identificar agrupamentos dos dados de acordo com o grau de semelhança. Queremos achar um grupo de objetos, que são similares entre si e diferentes de outros grupos.



Que perguntas o cluster pode responder?

THE DEVELOPER'S CONFERENCE

- Como agrupo os clientes que compram no site?
- Onde coloco uma antena de operadora de celular?
- Onde deixo uma ambulância estacionada para atender a uma emergência?
- Como classifico cervejas em categorias distintas?

Globalcode

Utilização mais usual de cluster

THE DEVELOPER'S CONFERENCE

- Organizar dados dentro do cluster para mostrar a estrutura dos dados. Ex. Cluster em genes.
- Particionar os dados. Ex. Segmentação de clientes.
- Preparar para outras técnicas. Ex. Sistemas de recomendação.
- Descoberta nos dados. Ex. Descoberta de padrões.

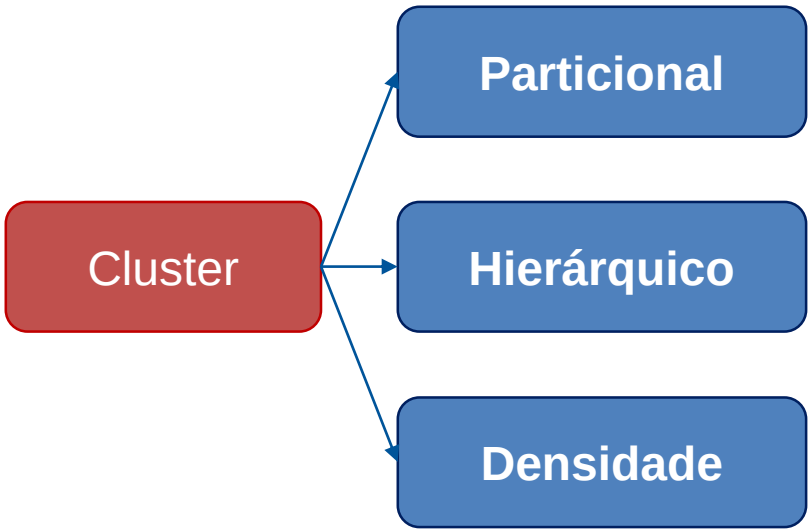
Globalcode

THE DEVELOPER'S CONFERENCE

Preparar para outras técnicas: **Sistemas de recomendação**

Globalcode

Tipos de cluster



Agrupamento Particional
Uma divisão de objetos em subconjuntos não sobrepostos (clusters), de modo que cada objeto esteja exatamente em um subconjunto. Ex. Segmentação de clientes.

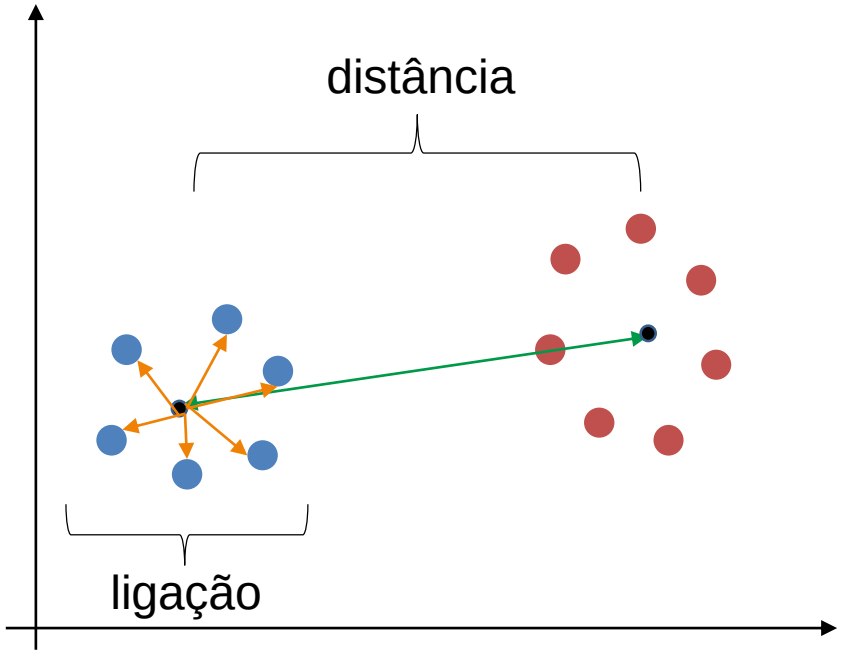
Agrupamento hierárquico
Um conjunto de clusters aninhados organizados como uma árvore hierárquica.

Agrupamento por Densidade
Baseado na existência de regiões densas de dados, separadas por regiões com baixa densidade de dados.

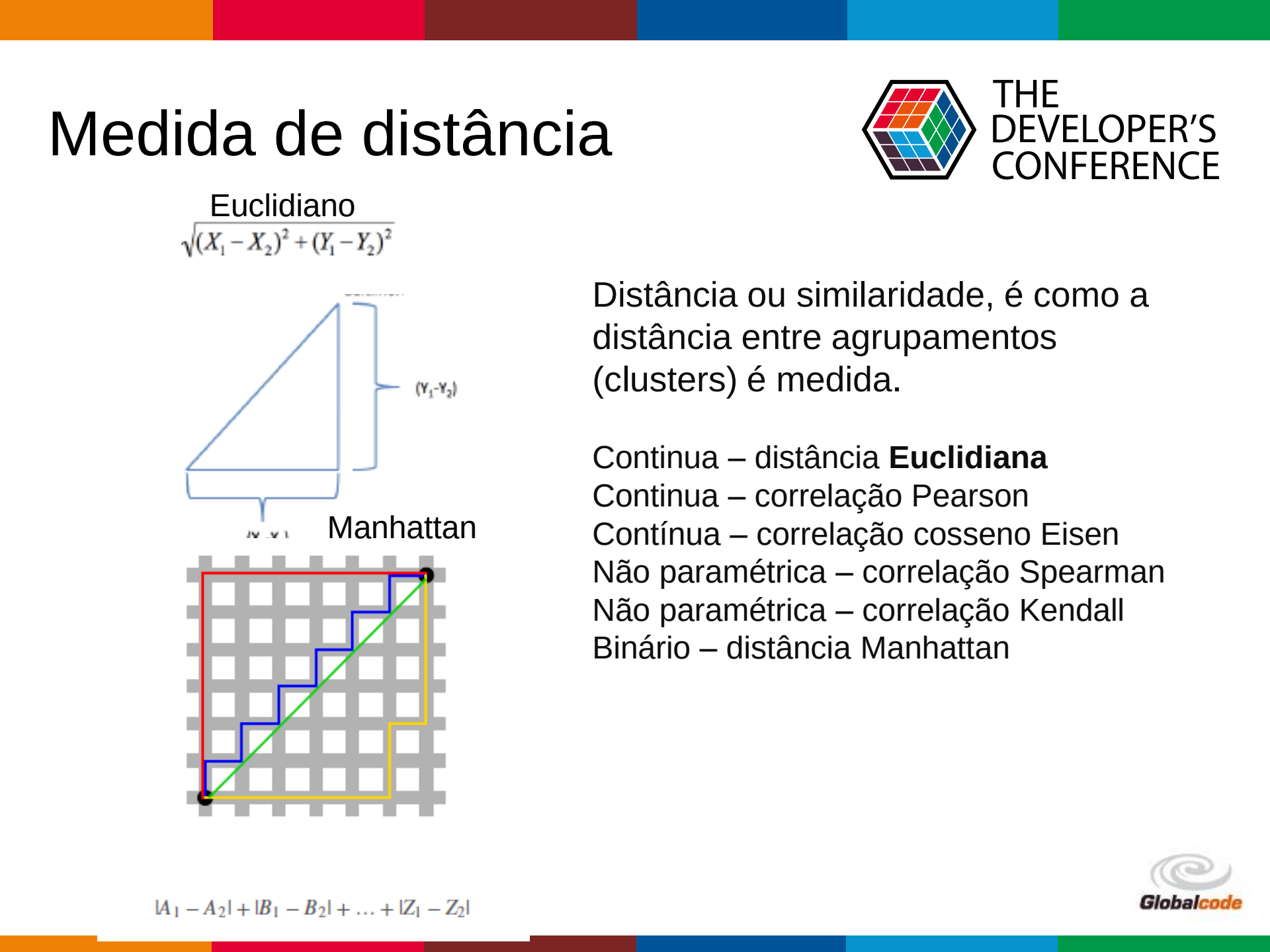
Método de ligação

Utiliza as informações de distância para agrupar pares de objetos em clusters com base na sua similaridade.

- Completo
- **Centroide**
- **Ward**
- Média
- Mediana



Medida de distância



Euclidiano

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

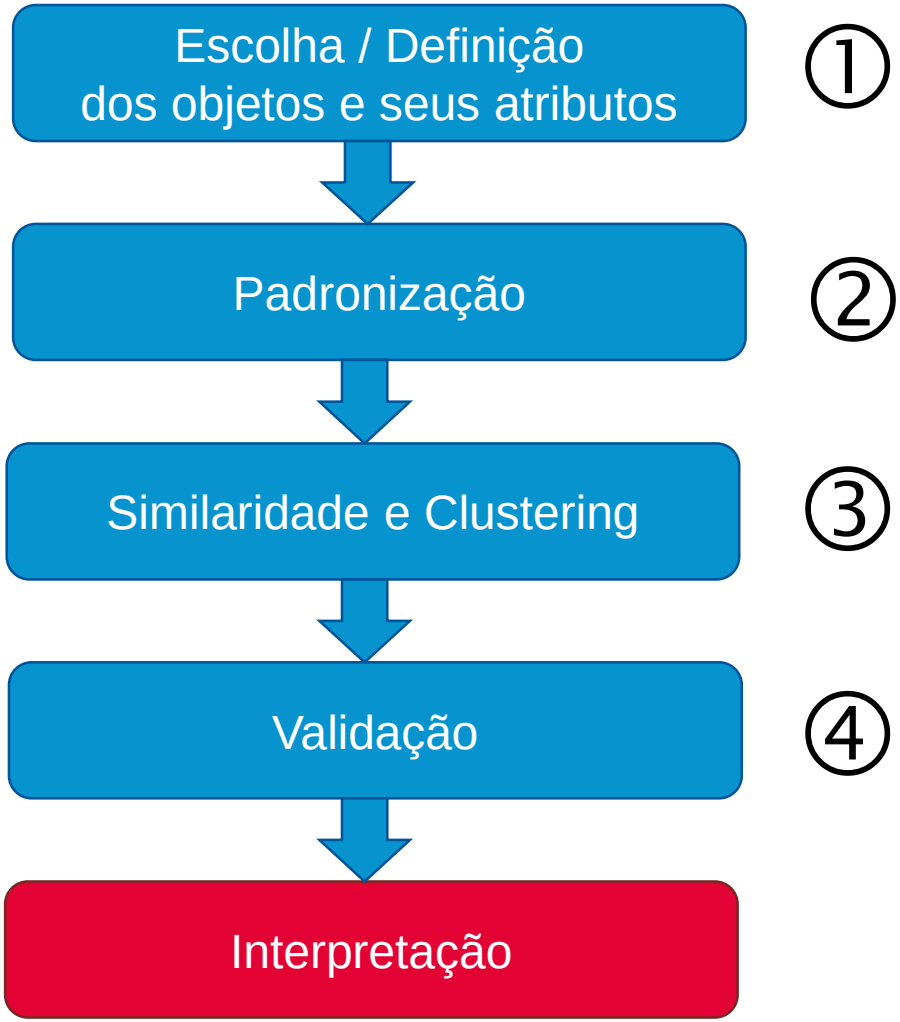
Manhattan

$$|A_1 - A_2| + |B_1 - B_2| + \ldots + |Z_1 - Z_2|$$

Distância ou similaridade, é como a distância entre agrupamentos (clusters) é medida.

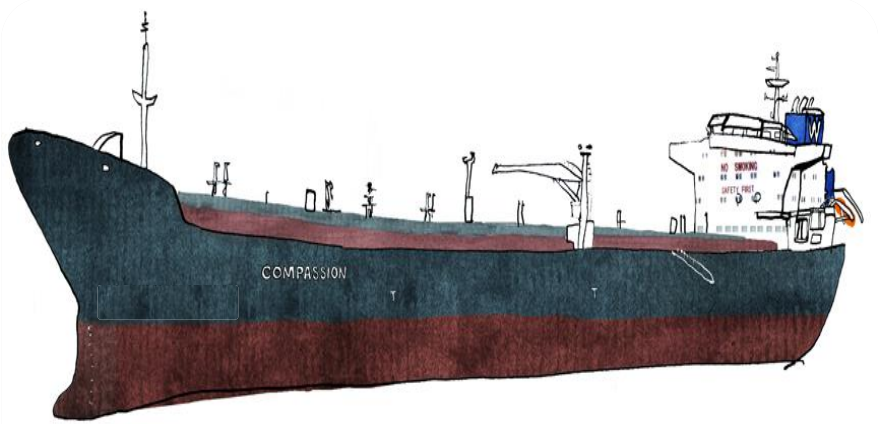
Continua – distância **Euclidiana**
Continua – correlação Pearson
Contínua – correlação cosseno Eisen
Não paramétrica – correlação Spearman
Não paramétrica – correlação Kendall
Binário – distância Manhattan

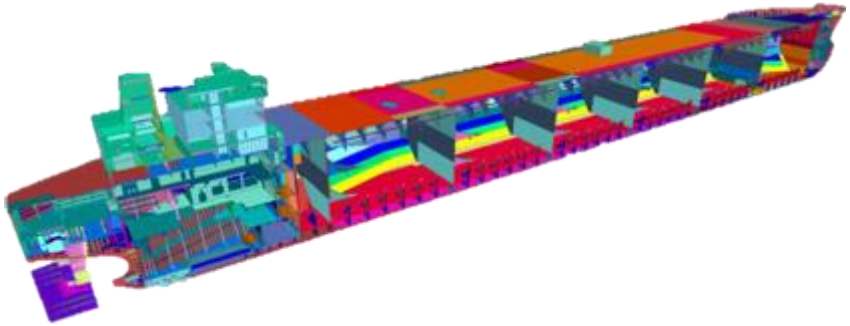
Passos para o cluster



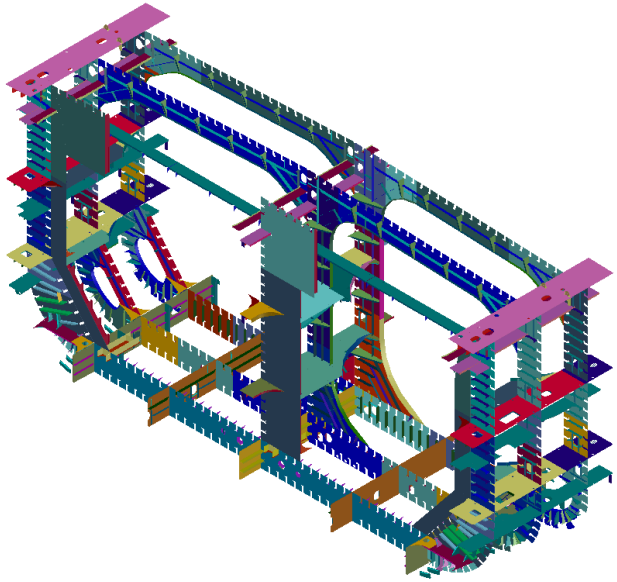
Projeto Estaleiro Atlântico Sul

THE DEVELOPER'S CONFERENCE





SUBMONTAGEM



SUEZMAX ESTRUTURA
≈ 21.500 t

SUEZMAX SUBMONTAGENS
≈ 7.100 t

= 1/3

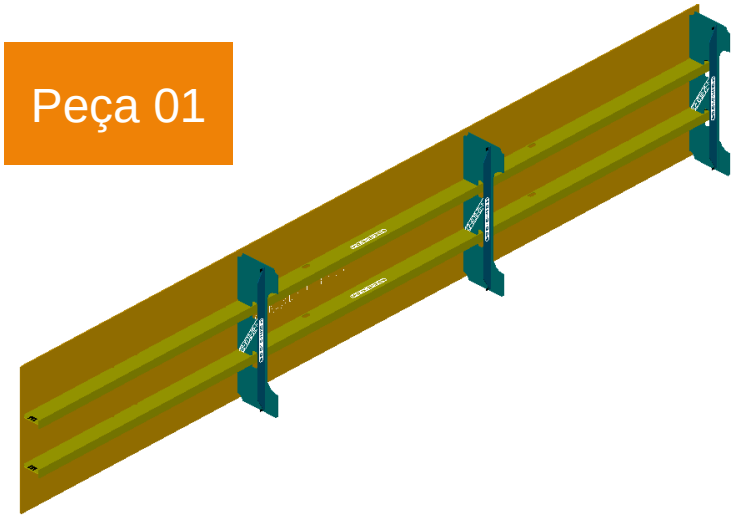
Globalcode

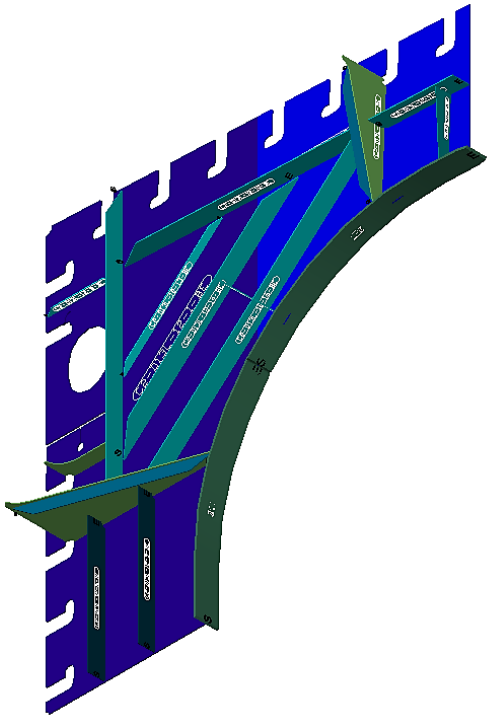
Projeto Estaleiro Atlântico Sul

Submontagem - Situação Atual

THE DEVELOPER'S CONFERENCE

Peça 01





Peça 02

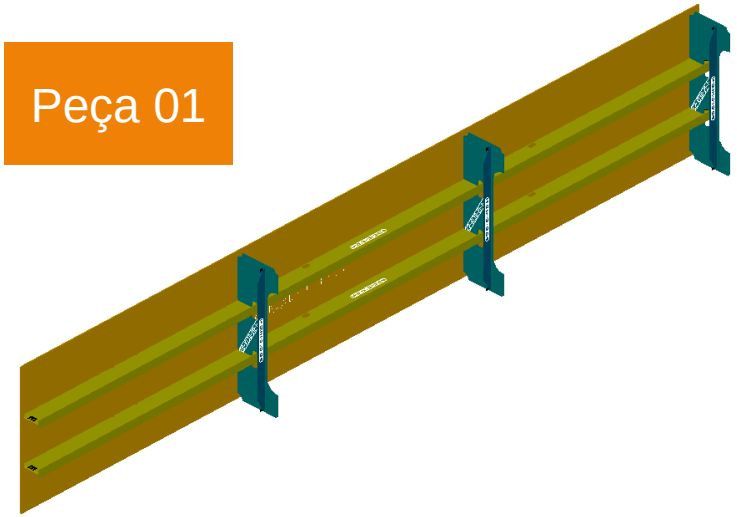
	peso	compr. de solda	EMENDA		ITENS SECUNDÁRIOS				
			PU	unilateral	quantidade	diversidade	inclinação	outro lado	barra face
Peça 01	7,5 t	83 m	não	não	5	não	não	não	não
Peça 02	4,1 t	96 m	sim	não	20	sim	não	sim	sim

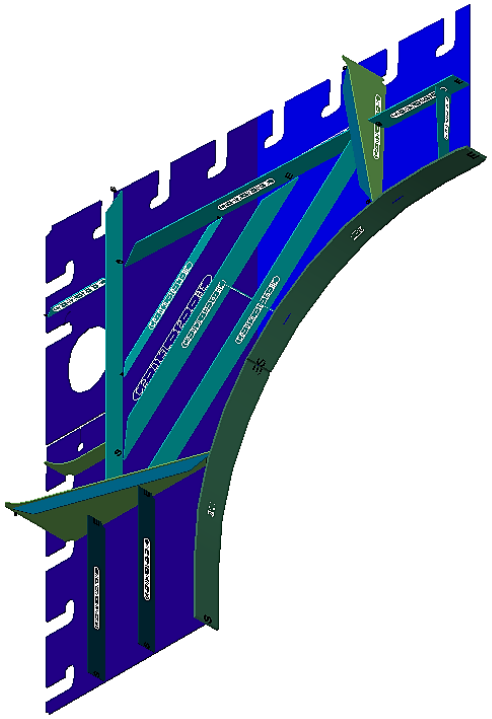
Projeto Estaleiro Atlântico Sul

Submontagem - Situação Atual

THE DEVELOPER'S CONFERENCE

Peça 01





Peça 02

	Comprimento de solda (m)	Tempo estimado (info pela produção) (h)
Peça 1	83	13
Peça 2	96	24
Diferença %	16%	85%

① Escolha / Definição dos objetos e seus atributos

THE DEVELOPER'S CONFERENCE

Caso tenhamos muitas variáveis, é importante reduzir a dimensionalidade, utilizando técnicas como a análise de componentes principais - PCA

COMPONENTE	NÚM.PEÇ	NÚM.FILH	COMP.SOI	ESPESS.PC	ÁREA BAS	PESO PEÇ	PESO FILH	NÚM.CHA	FLANGE	PEÇAS OU	INCLINAD
ANEL 01/A11/C/P	8	0	27,5	19,3	6,3	1939,8	0,0	1	0	0	0
ANEL 01/A11/C/P	12	0	2,9	22,8	3,0	709,2	0,0	2	1	0	0
ANEL 01/A11/C/P	0	5	6,4	20,9	0,4	2267,7	2267,7	1	0	1	1

....

Matriz de 11 x 6.528

11 variáveis e 6.528 subconjuntos (componentes)

Globalcode

② Padronização

THE DEVELOPER'S CONFERENCE

- Objetivo:
  - Tornar as variáveis comparáveis.
- Fórmula:
  - $\frac{x-\mu_x}{\sigma_x}$

Globalcode

③ Similaridade e Clusterização

THE DEVELOPER'S CONFERENCE

- Utilizando distância Euclidiana

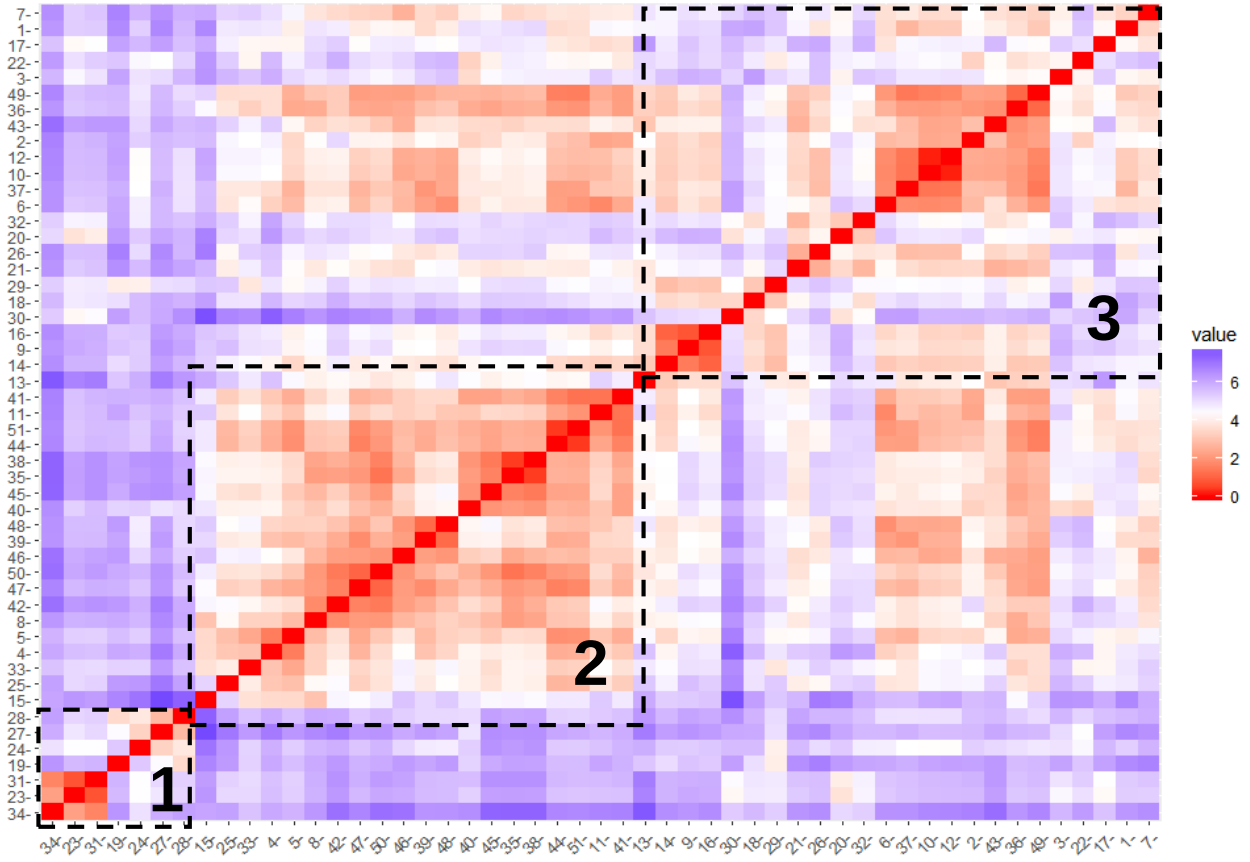
	1	2	3	4	5	6
1	0.00	4.32	4.23	4.83	4.04	3.63
2	4.32	0.00	4.90	4.69	3.50	2.82
3	4.23	4.90	0.00	5.95	5.16	5.33
4	4.83	4.69	5.95	0.00	1.50	3.34
5	4.04	3.50	5.16	1.50	0.00	2.85
6	3.63	2.82	5.33	3.34	2.85	0.00

Globalcode

③ Similaridade e Clusterização

THE DEVELOPER'S CONFERENCE

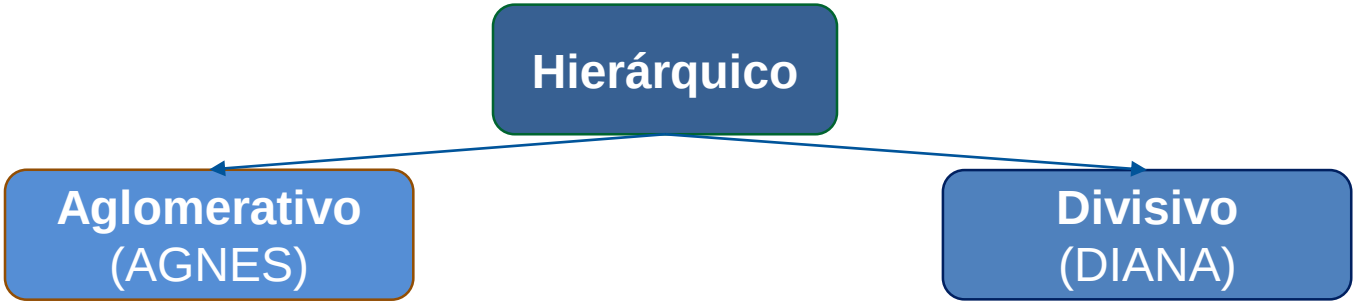
Matriz de dissimilaridade



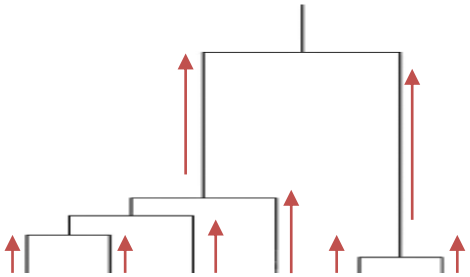
Globalcode

③ Similaridade e Clusterização

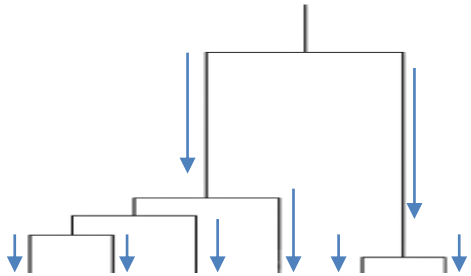
THE DEVELOPER'S CONFERENCE



Inicia com cada observação em um cluster. Ao final há apenas um cluster com todas observações.



Ao contrário do aglomerativo, inicia com um cluster único e ao final cada observação está em um cluster.

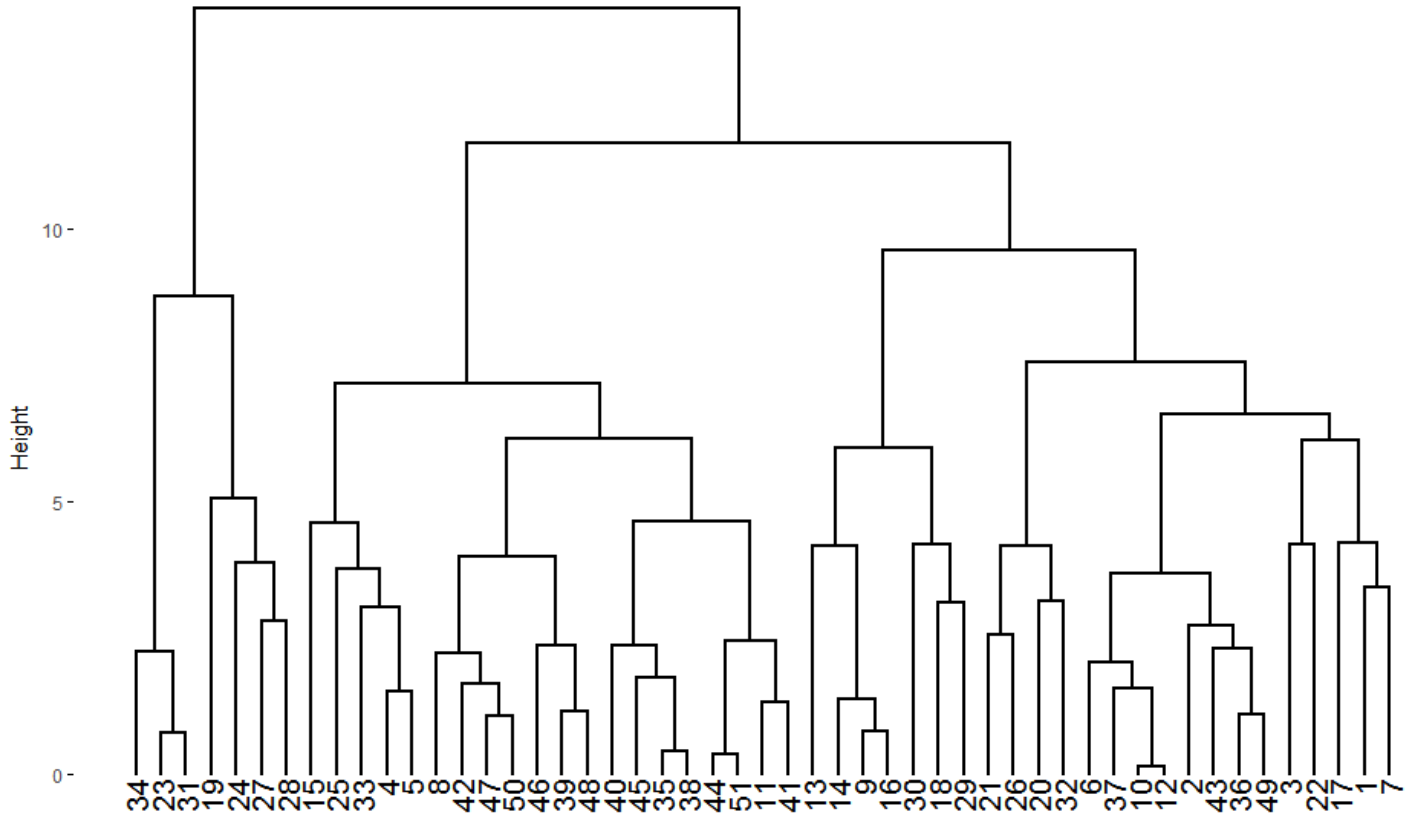


Globalcode

③ Similaridade e Clusterização

THE DEVELOPER'S CONFERENCE

Dendrograma



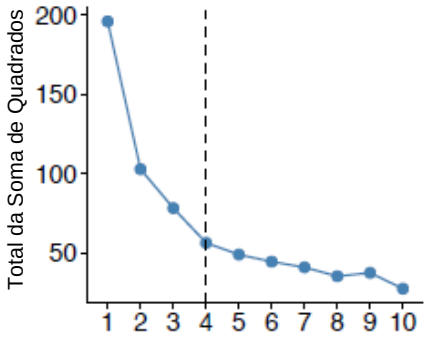
Globalcode

③ Determinação do número de clusters

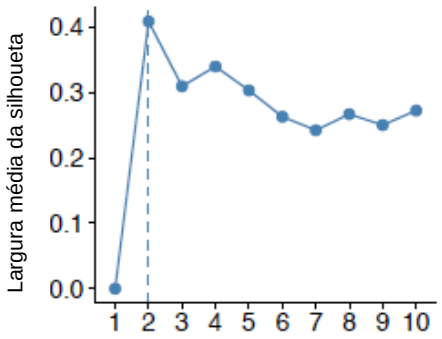
THE DEVELOPER'S CONFERENCE

- Para **cluster particional**:
  - Métodos diretos
    - Elbow
    - Silhouette
  - Métodos estatísticos
    - Gap statistic

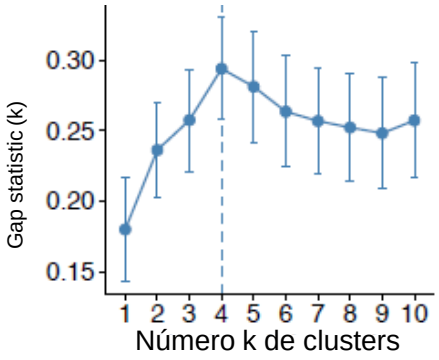
Método Elbow



Método Silhouette



Gap statistic

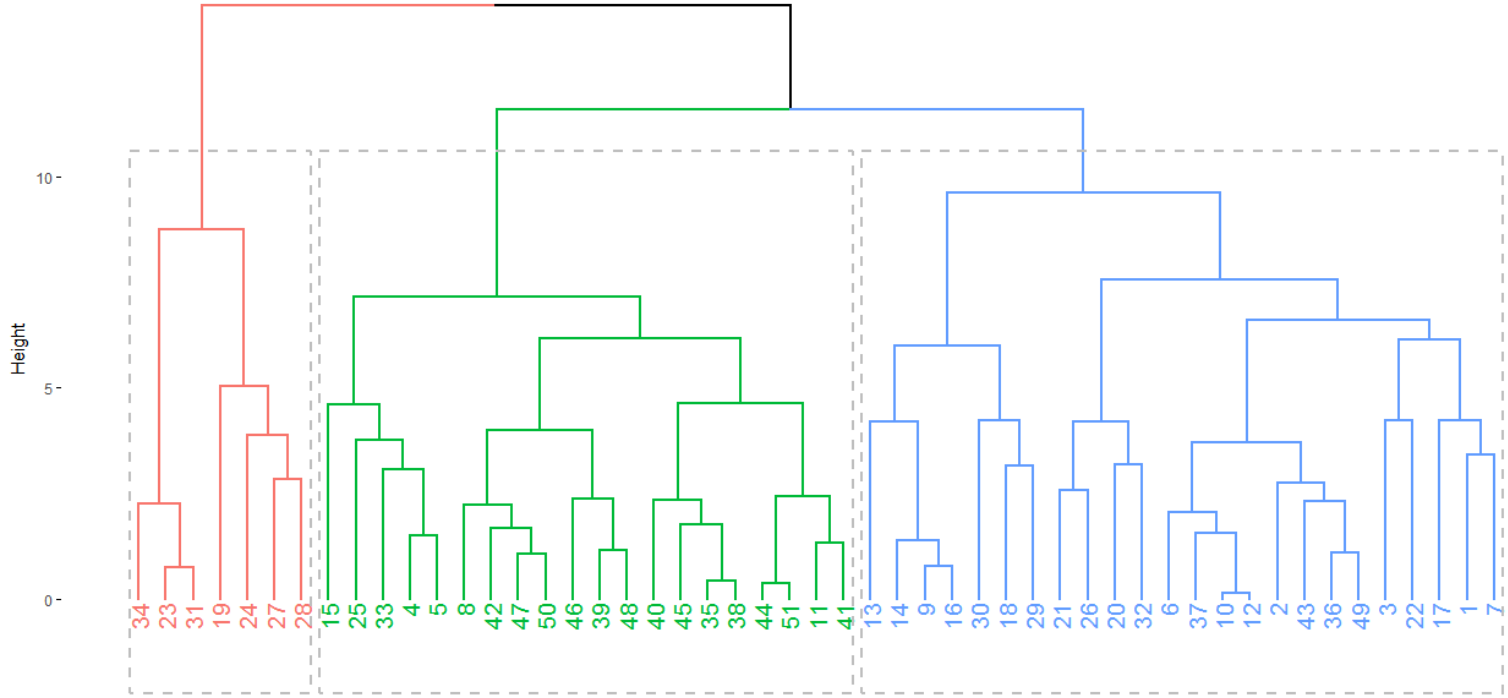


Globalcode

③ Similaridade e Clusterização

THE DEVELOPER'S CONFERENCE

Dendrograma para 3 clusters



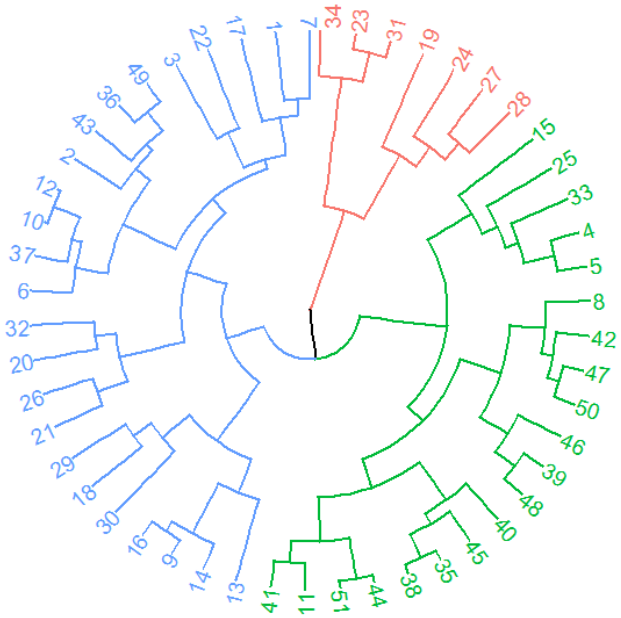
Globalcode

③ Similaridade e Clusterização

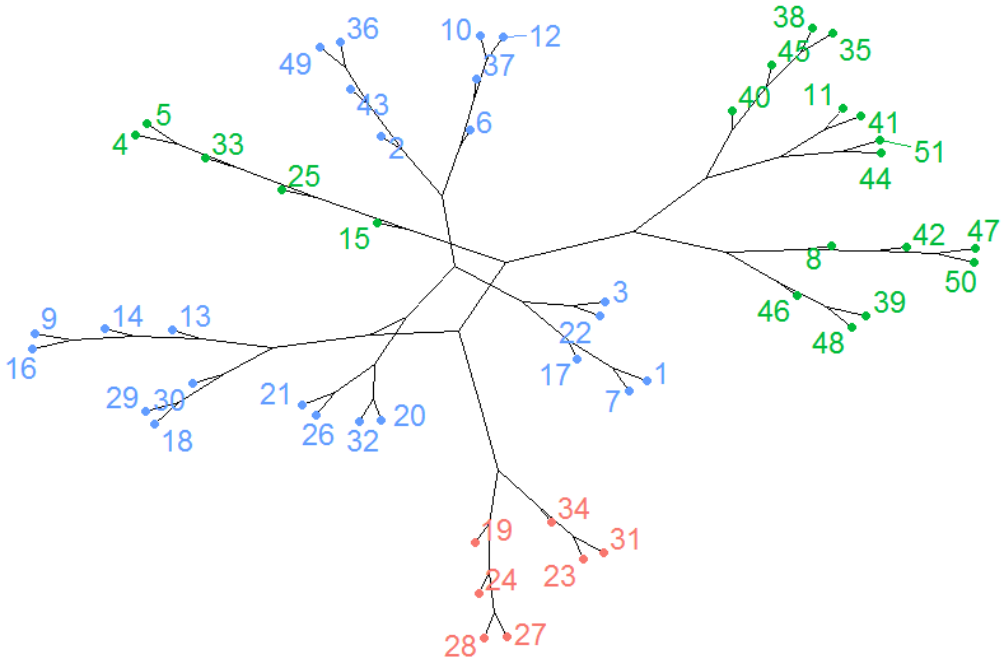
THE DEVELOPER'S CONFERENCE

Outras visualizações:

Dendrograma circular



Dendrograma filogenético



Globalcode

④ Validação

THE DEVELOPER'S CONFERENCE

- Tendência de agrupamento
  - Método estatístico: *estatística de Hopkins*
  - Método visual: *Avaliação visual da tendência do agrupamento (VAT)*

Globalcode

④ Validação

THE DEVELOPER'S CONFERENCE

Estatística de Hopkins

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Sendo os dados originais $p_1, p_2, \ldots, p_n$

$$x_i = dist(p_i, p_j)$$

Sendo amostra aleatória $q_1, q_2, \ldots, q_n$

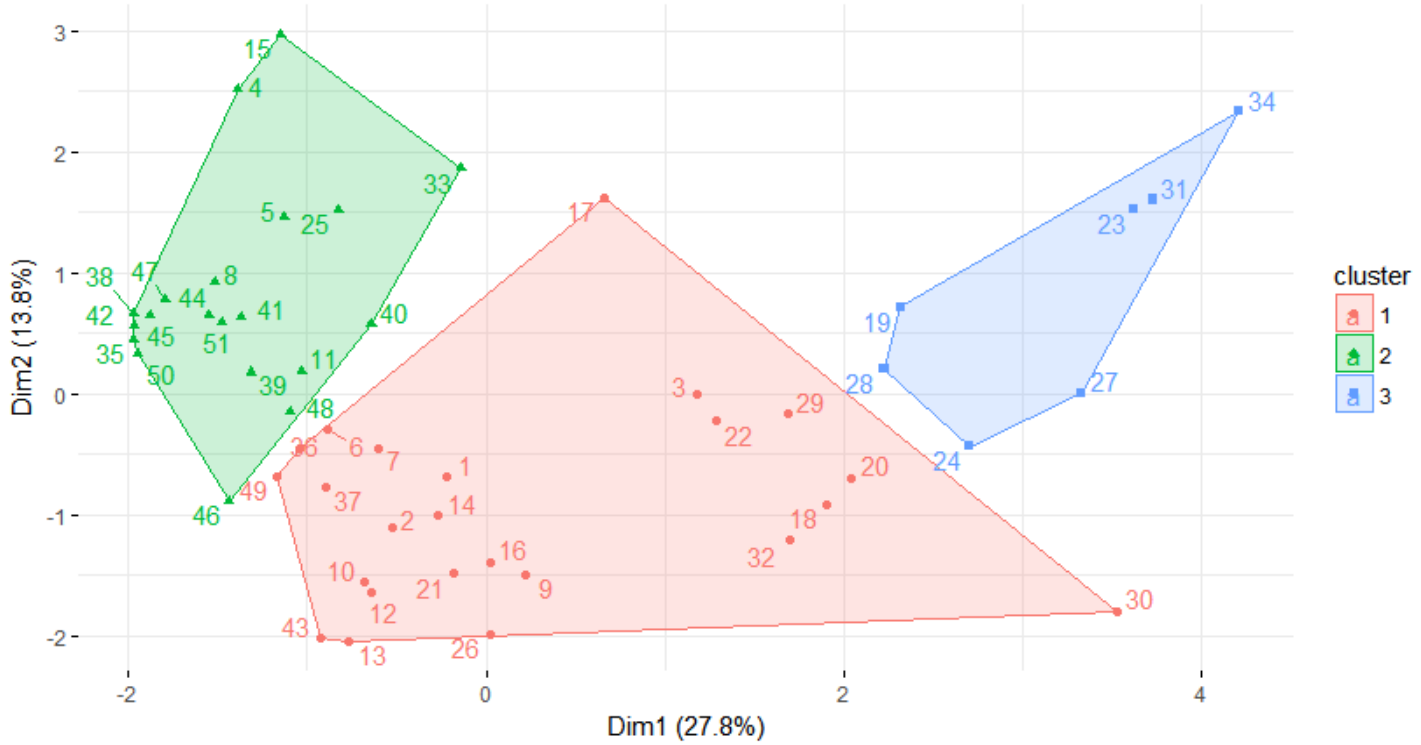
$$y_i = dist(q_i, q_j)$$

Um valor de H distante de 0,5 indica que os dados originais são diferentes dos dados aleatórios, portanto os clusters são significativos.

```
$H
[1] 0.3178
```

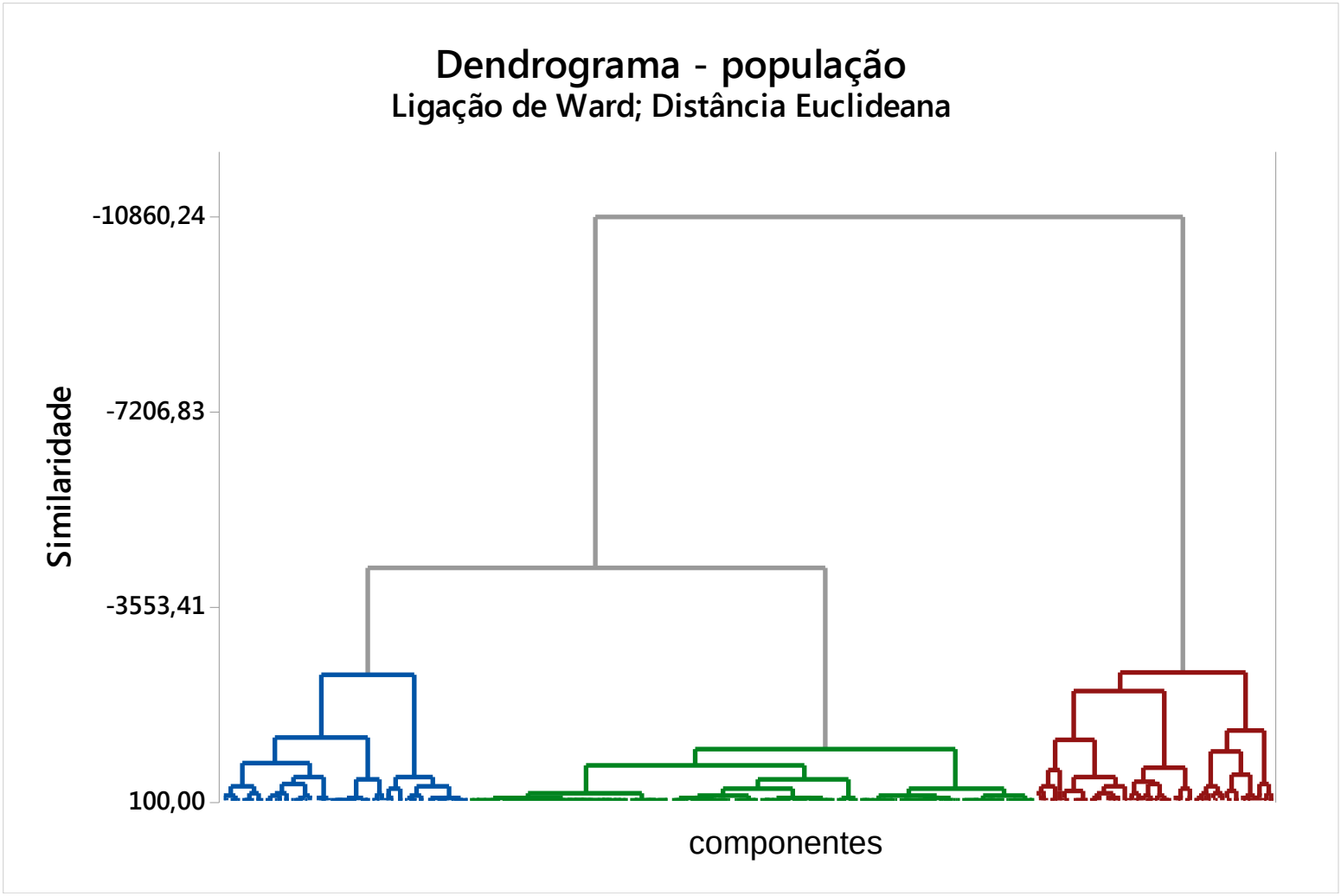
Globalcode

④ Validação



Interpretação

Dendograma final com 6.528 componentes



THE DEVELOPER'S CONFERENCE

Nós fazemos a revolução digital.
Vamos juntos!

www.siqueiracampos.com
marco@siqueiracampos.com

Globalcode