

MBA
USP
ESALQ

Unsupervised Machine Learning: Clustering I

Profa. Adriana Silva

Supervisionado x Não Supervisionado



VS



Supervisionado x Não Supervisionado



“Aprendendo sem conhecimento prévio da classificação da amostra, aprendendo sem um professor.”

Kohonen (1995), *“Self-Organizing Maps”*

Análise de Cluster

Aplicações

- Marketing
- Vendas
- Fraude

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

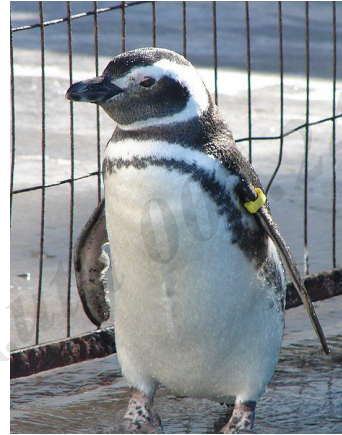
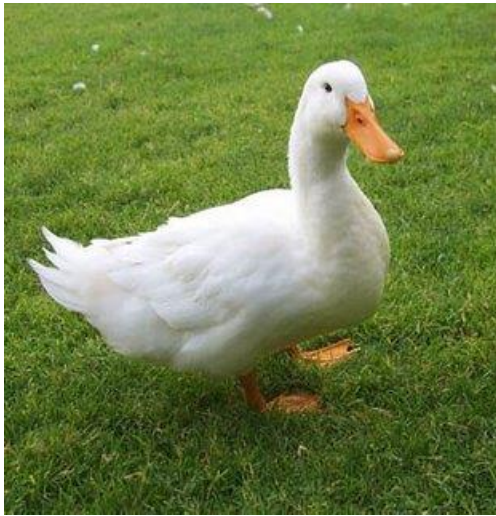
Métodos de Agrupamento

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.

Análise de Cluster

O que é similaridade?



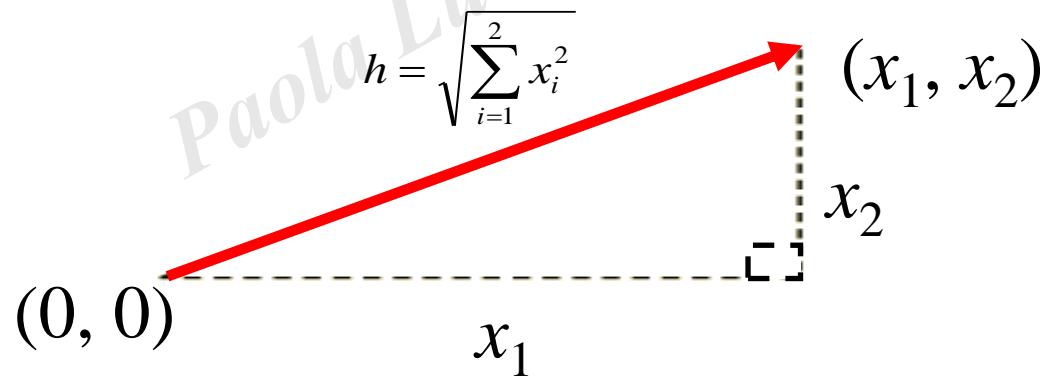
????

Análise de Cluster

Distância Euclidiana

$$D_E = \|\mathbf{x} - \mathbf{w}\| = \sqrt{\sum_{i=1}^k (x_i - w_i)^2}$$

- A distância Euclidiana gera a distância linear entre quaisquer dois pontos em um campo com k dimensões.
- É uma generalização do Teorema de Pitágoras



Análise de Cluster

Distância Minkowsky

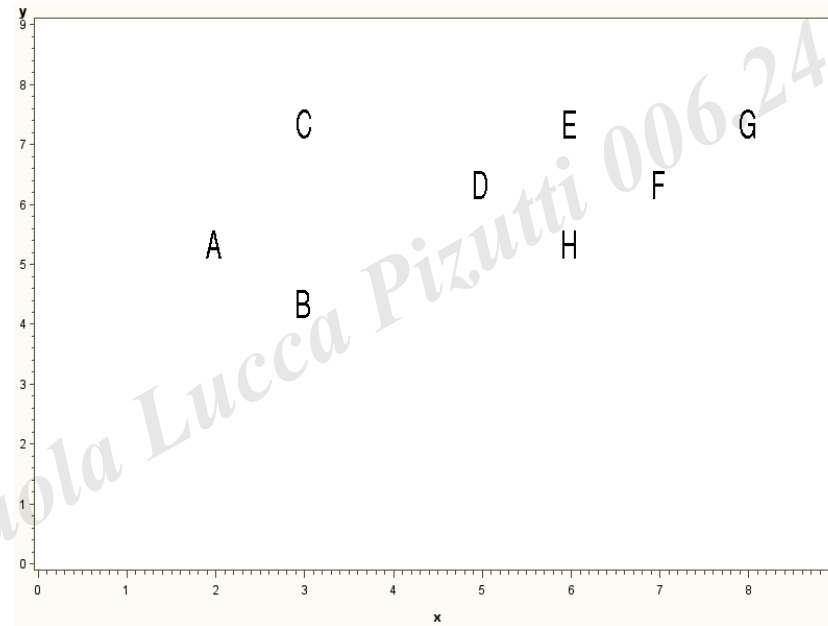
$$d_p(x_i, x_j) = \left(\sum_{k=1}^d |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$$

- Pode criar pesos para cada variável, quando necessário.
- É uma medida menos afetada pela presença de outliers (comparada a distância euclidiana).

Análise de Cluster

Exemplo Distâncias

ID	X	Y
A	2	5
B	3	4
C	3	7
D	5	6
E	6	8
F	7	6
G	8	8
H	6	5



Análise de Cluster

Exemplo Distâncias

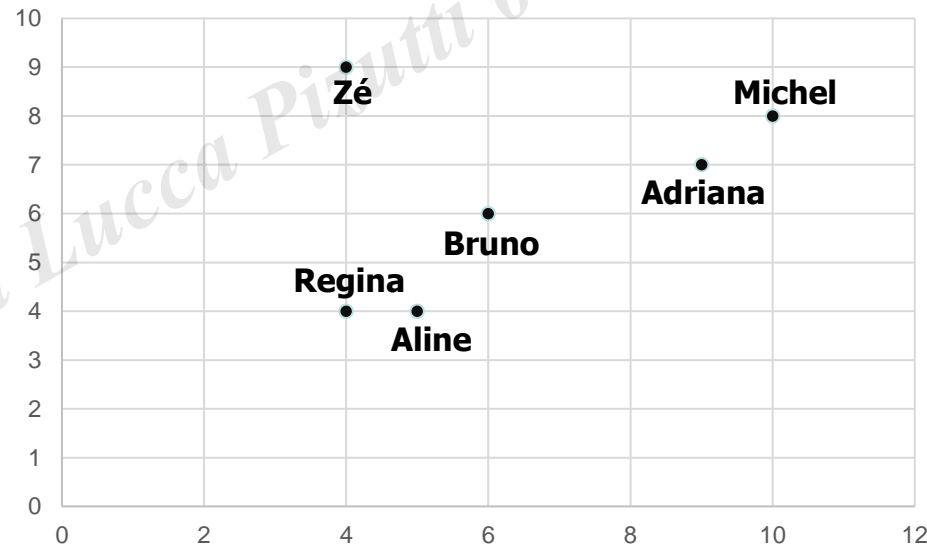
- Agrupar alunos que sejam parecidos, com relação as notas em matemática e português.

Aluno	Matemática	Português
Adriana	9	7
Aline	5	4
Bruno	6	6
Michel	10	8
Regina	4	4
Zé	4	9

Análise de Cluster

Exemplo Distâncias

Aluno	Matemática	Português
Adriana	9	7
Aline	5	4
Bruno	6	6
Michel	10	8
Regina	4	4
Zé	4	9



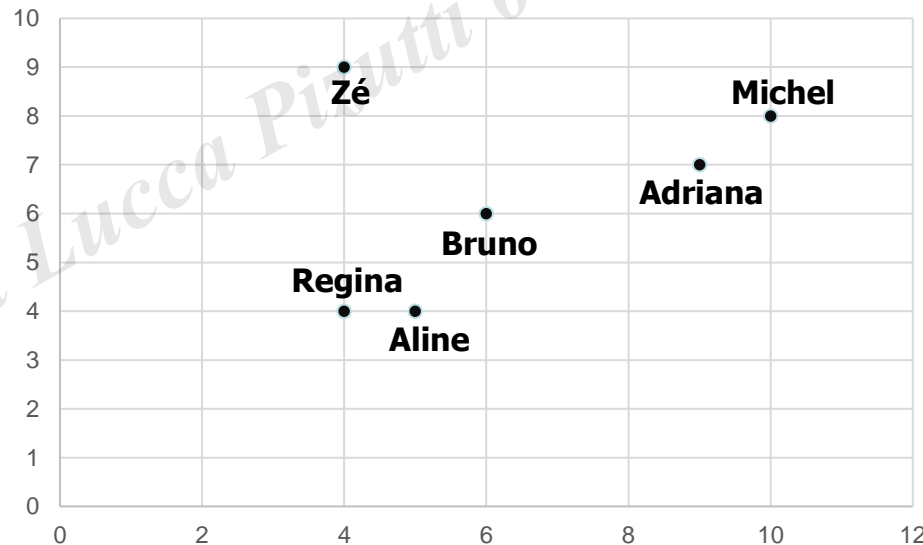
Análise de Cluster

Exemplo Distâncias Euclidiana

Qual a distância entre os Aline e Michel?

$$D^2 = (\underline{x_{14}} - \underline{x_{12}})^2 + (x_{24} - x_{22})^2$$

Aluno	Matemática	Português
Adriana	9	7
Aline	5	4
Bruno	6	6
Michel	10	8
Regina	4	4
Zé	4	9



Análise de Cluster

Distância Euclidiana

Distância Euclidiana ao Quadrado

$$D^2 = (x_{11} - x_{12})^2 + (x_{21} - x_{22})^2$$

Distância Euclidiana

$$D = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

Aluno	Matemática	Português
Adriana	9	7
Aline	5	4
Bruno	6	6
Michel	10	8
Regina	4	4
Zé	4	9

Aluno	Matemática	Português
Adriana	x11	x21
Aline	x12	x22
Bruno	x13	x23
Michel	x14	x24
Regina	x15	x25
Zé	x16	x26

Qual a distância (D) entre Aline e Michel?

$$D^2 = (x_{14} - x_{12})^2 + (x_{24} - x_{22})^2$$

$$D^2 = (5 - 10)^2 + (4 - 8)^2 = 5^2 + 4^2 = 41$$

D = 6,40 → Distância Euclidiana

Análise de Cluster

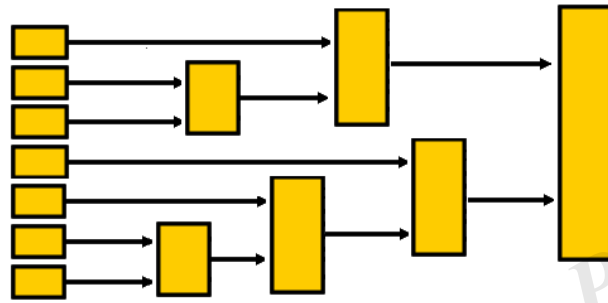
Distância Euclidiana

	Dr i	Li	Bru	Mi	Re
Li	5.000000				
Bru	3.162278	2.236068			
Mi	1.414214	6.403124	4.472136		
Re	5.830952	1.000000	2.828427	7.211103	
zž	5.385165	5.099020	3.605551	6.082763	5.000000

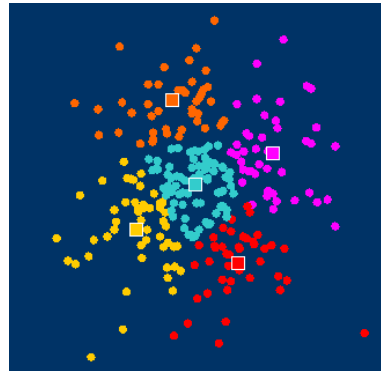
Análise de Cluster

Métodos de Agrupamento

- Hierárquico



- Cluster Não Hierárquico



Análise de Cluster

Técnicas de Agrupamentos - Hierárquico

- Single Linkage - Vizinho mais próximo
- Complete Linkage - Vizinho mais longe
- Average Linkage - Média
- Centroid Method – Centroíde
- Ward's Method

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Single Linkage – Vizinho mais próximo

1. Calcula-se a distância de todos os alunos contra todos os alunos
2. Agrupa os alunos mais próximos (menor distância)
3. Define a distância do primeiro grupo contra os demais alunos baseado na **menor** distância entre cada integrante do grupo com os demais
4. Etapa 3 até ter apenas um único grupo
5. Desenha-se o dendograma baseado na distância encontrada

Análise de Cluster

Técnicas de Agrupamentos – Single Linkage – Vizinho mais próximo

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Single Linkage – Vizinho mais próximo

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Complete Linkage – Vizinho mais longe

1. Calcula-se a distância de todos os alunos contra todos os alunos
2. Agrupa os alunos mais próximos (menor distância)
3. Define a distância do primeiro grupo contra os demais alunos baseado na **maior** distância entre cada integrante do grupo com os demais
4. Etapa 3 até ter apenas um único grupo
5. Desenha-se o dendograma baseado na distância encontrada

Análise de Cluster

Técnicas de Agrupamentos – Complete Linkage – Vizinho mais longe

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Complete Linkage – Vizinho mais longe

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Complete Linkage – Vizinho mais longe

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Average Linkage

1. Calcula-se a distância de todos os alunos contra todos os alunos
2. Agrupa os alunos mais próximos (menor distância)
3. Define a distância do primeiro grupo contra os demais alunos baseado na **média** da distância entre cada integrante do grupo com os demais
4. Etapa 3 até ter apenas um único grupo
5. Desenha-se o dendograma baseado na distância encontrada

Análise de Cluster

Técnicas de Agrupamentos – Centroid Method – Centróide

1. Calcula-se a distância de todos os alunos contra todos os alunos
2. Agrupa os alunos mais próximos (menor distância)
3. Define a distância do primeiro grupo contra os demais alunos através da fórmula (cálculo do centroide)
4. Etapa 3 até ter apenas um único grupo
5. Desenha-se o dendograma baseado na distância encontrada

Análise de Cluster

Técnicas de Agrupamentos – Centróide

Os elementos que serão agrupados são denominados **p e q**

A novo grupo (**p+q**) é denominado **t**.

A distância entre o novo grupo t e o elemento/grupo r é dada por

$$S_{t,r} = \frac{N_p}{N_p + N_q} S_{p,r} + \frac{N_q}{N_p + N_q} S_{q,r} - \frac{N_p * N_q}{(N_p + N_q)^2} S_{p,q}$$

Em que,

N_p é o número de observações no grupo p,

N_q é o número de observações no grupo q,

$S_{p,r}$ é a distância entre p e r,

$S_{q,r}$ é a distância entre q e r,

$S_{p,q}$ é a distância entre p e q.

Em outra notação:

$$d_{(UV)W} = (\bar{X}_{UV} - \bar{X}_W)'(\bar{X}_{UV} - \bar{X}_W)$$

Análise de Cluster

Técnicas de Agrupamentos – Centroid Method – Centróide

Os elementos que serão agrupados são denominados **p e q**

A novo grupo (**p+q**) é denominado **t**.

A distância entre o novo grupo t e o elemento/grupo r é dada por

$$S_{t,r} = \frac{N_p}{N_p + N_q} S_{p,r} + \frac{N_q}{N_p + N_q} S_{q,r} - \frac{N_p * N_q}{(N_p + N_q)^2} S_{p,q}$$

Em que,

N_p é o número de observações no grupo p,

N_q é o número de observações no grupo q,

$S_{p,r}$ é a distância entre p e r,

$S_{q,r}$ é a distância entre q e r,

$S_{p,q}$ é a distância entre p e q.

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Centroid Method – Centróide

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Técnicas de Agrupamentos – Ward Method

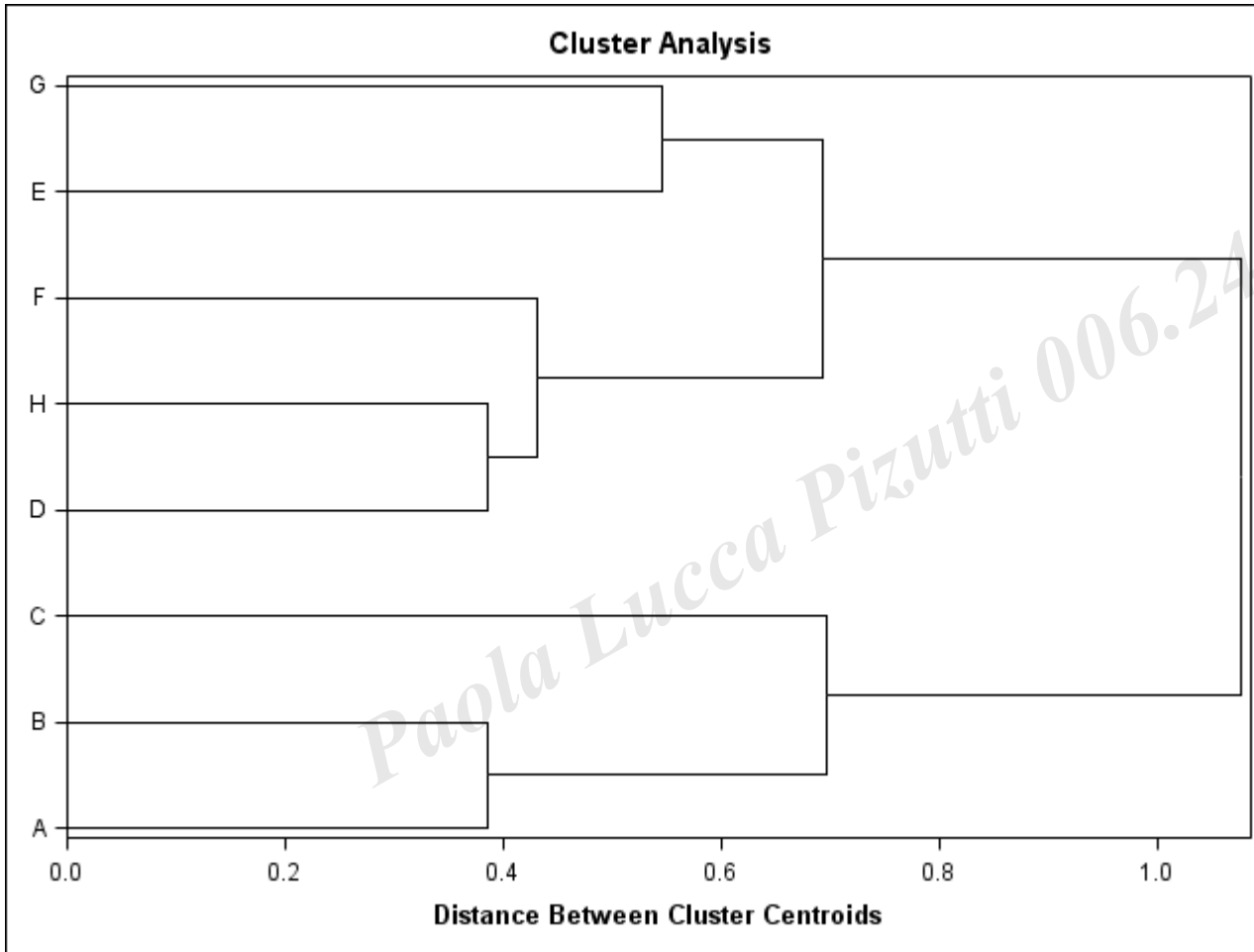
Também conhecido como método do incremento das somas de quadrados, é baseado na análise de variância. Neste método as somas de quadrados entre e dentro grupos, em relação as p variáveis, são utilizadas como critério de agrupamento. A ideia do método de Ward é aglomerar os grupos R e S que minimizam a soma de quadrados dentro dos grupos, ou seja, a soma de quadrado dos erros.

1. Calcula-se a distância de todos os alunos contra todos os alunos
2. Agrupa os alunos mais próximos (menor distância)
3. Define a distância do primeiro grupo contra os demais alunos através da fórmula de Wald
4. Etapa 3 até ter apenas um único grupo
5. Desenha-se o dendograma baseado na distância encontrada

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i} \right] (\bar{X}_l - \bar{X}_i)(\bar{X}_l - \bar{X}_i)$$

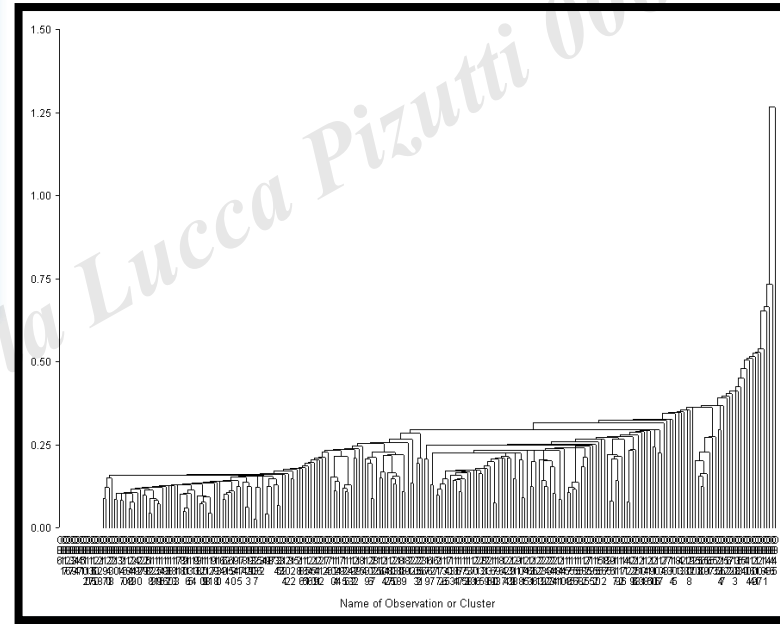
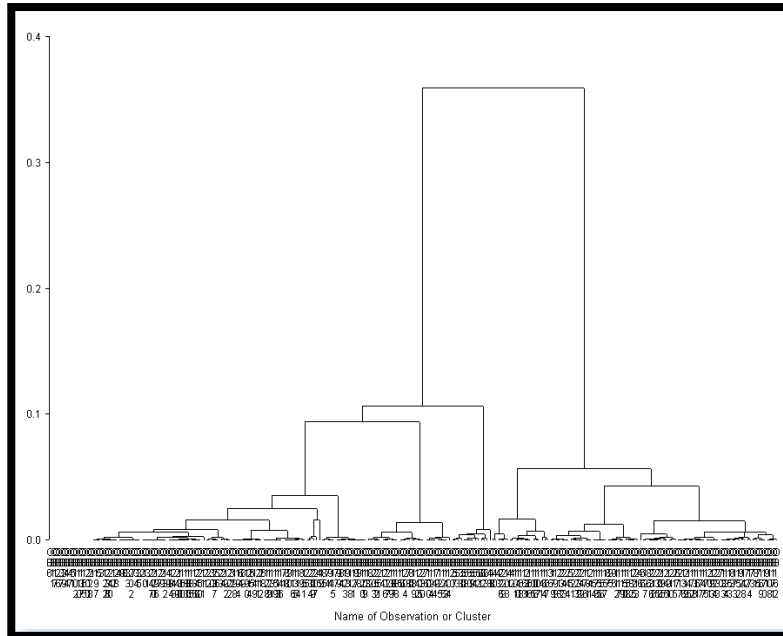
Análise de Cluster

Dendograma

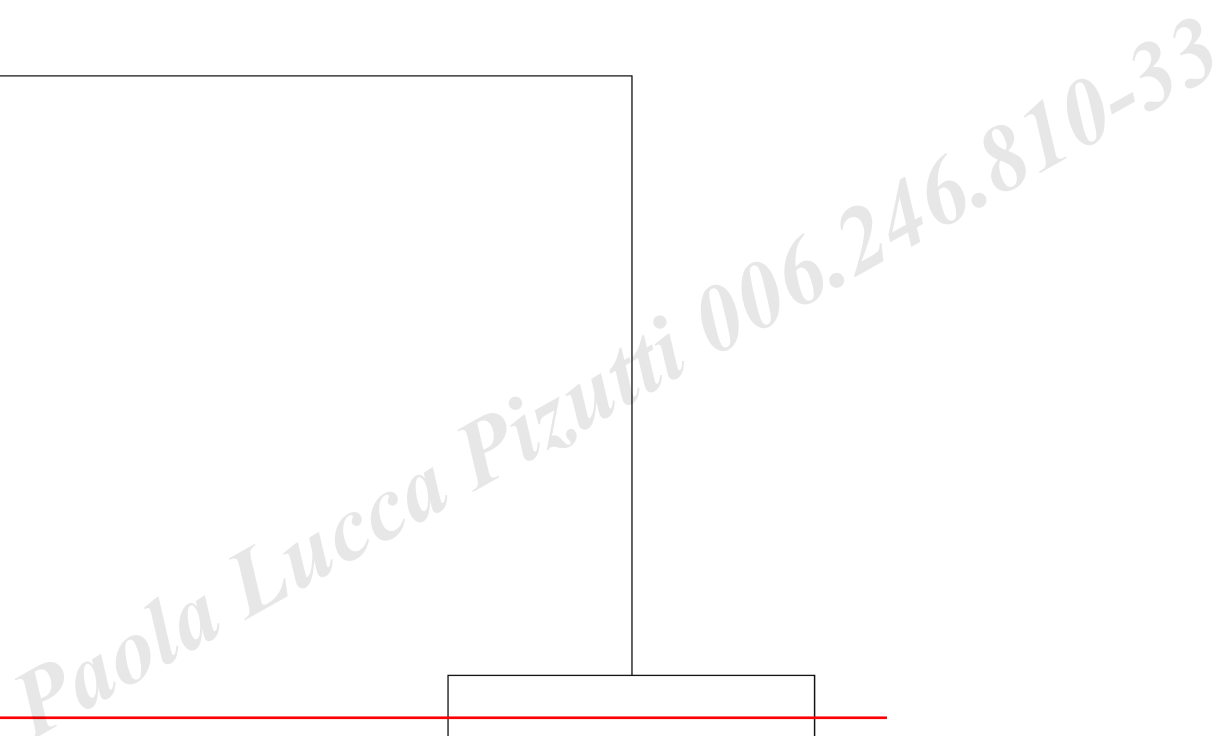


Análise de Cluster

Dendrograma



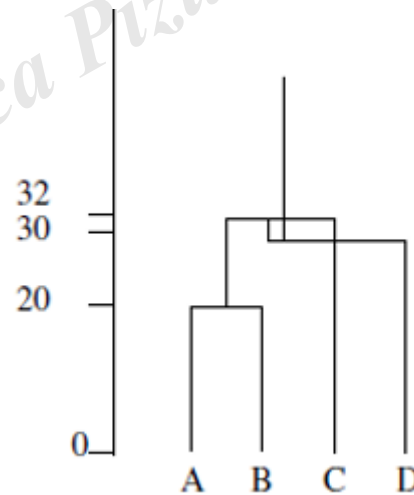
Dendrograma



Análise de Cluster

Técnicas de Agrupamentos – Hierárquico

- Podem provocar inversões. Ocorrem quando inexiste uma estrutura de cluster clara.
- Neste caso D é adicionado ao grupo (ABC), a uma distância de 30, inferior à distância a qual se juntou C(AB).



Análise de Cluster

Técnicas de Agrupamentos – Hierárquico – Comparação

- Single Linkage - Vizinho mais próximo
 - Estruturas geométricas diferentes, mas é incapaz de delinear grupos pouco separados.
- Complete Linkage - Vizinho mais longe
 - Clusters de mesmo diâmetro e isolam os outliers nos primeiros passos.
- Average Linkage – Média
 - Clusters de mesma variância interna, produzindo melhores partições.
- Ward's Method
 - Cluster com o mesmo n° de itens, baseado nos princípios de análises de variâncias.

Análise de Cluster

Técnicas para escolha do número de clusters – Hierárquico

- Coeficiente R^2

a) Soma de quadrados total : $SSTc = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})$

b) Soma de quadrados total intragrupo : $SSR = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(X_{ij} - \bar{X}_{i.})$

c) Soma de quadrados total intergrupos : $SSB = \sum_{i=1}^{g^*} n_i (\bar{X}_{i.} - \bar{X})(\bar{X}_{i.} - \bar{X})$

$$R^2 = \frac{SSB}{SSTc}$$

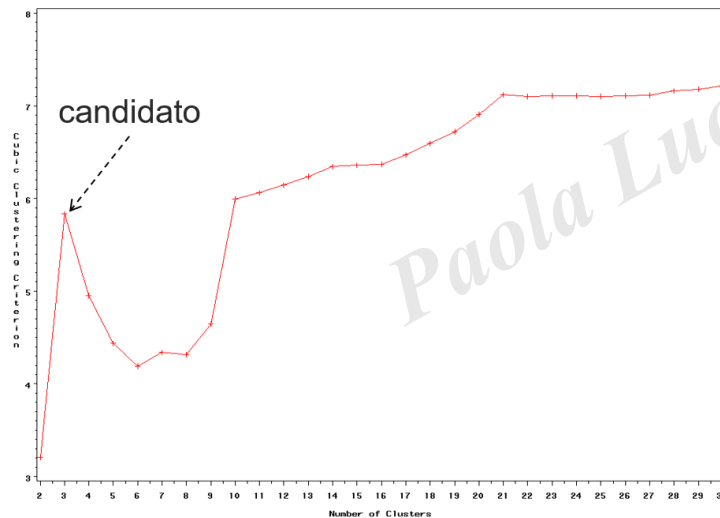
Quanto maior o R^2 , maior o SSB, consequentemente menor o SSR (que é o que nos importa, pouca variabilidade dentro dos clusters).

Procuramos sempre por algum salto! É possível notar que quanto menor o número de grupos (maior a variabilidade entre), menor o R^2 .

Análise de Cluster

Técnicas para escolha do número de clusters – Hierárquico

- CCC (Cubic Clustering Criterion)
 - compara o R^2 calculado com o seu esperado, $E[R^2]$, supondo que os clusters são gerados por uma distribuição uniforme. Se CCC > 3 (o que é bom), $R^2 > E[R^2]$, isto é, a estrutura de cluster é diferente da partição uniforme.



$$CCC = \ln \left[\frac{1 - E(R^2)}{1 - R^2} \right] \frac{\sqrt{\frac{np^*}{2}}}{(0.001 + E(R^2))^{1.2}}$$

Análise de Cluster

Técnicas para escolha do número de clusters – Método Elbow

Assim como no princípio de cluster, a ideia do método é minimizar a variabilidade dentro do cluster, ou seja:

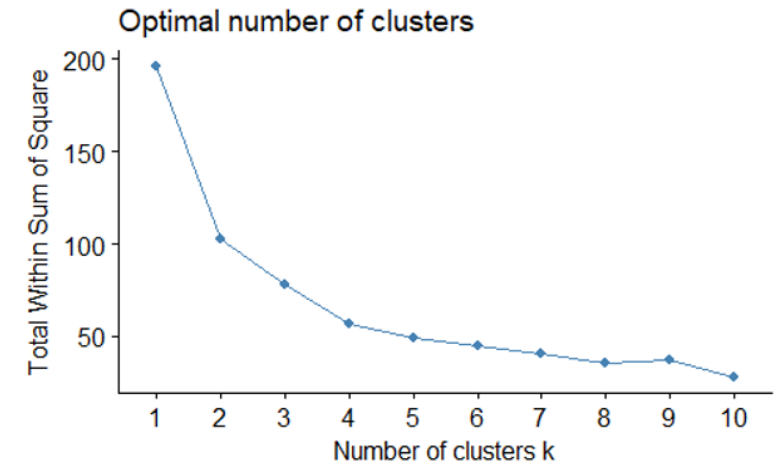
$$\text{minimize}(\sum_{i=1}^n W(C_k))$$

onde C_k é o k cluster e $W(C_k)$ é a variação dentro do cluster. Então, o total da soma dos quadrados dentro do cluster (wss) mede a homegenidade do cluster e queremos que seja tão pequeno quanto possível. Assim, podemos usar o seguinte algoritmo para definir os clusters como sendo ótimos:

Análise de Cluster

Técnicas para escolha do número de clusters – Método Elbow

1. Rodar o algoritmo de agrupamento para diferentes valores de k . Por exemplo, variando de 1 a 15 clusters.
2. Para cada k , calcular a soma dos quadrados total dentro do cluster (wss).
3. Traçar a curva de wss de acordo com o número de clusters.
4. A localização de uma curva (joelho) na trama é geralmente considerada como um indicador do número apropriado de clusters.



Análise de Cluster

Técnicas para escolha do número de clusters – Método Silhoutte

- O coeficiente de silhueta é a medida da relação entre um ponto e os membros do grupo dele.
- Se a medida “s” de um ponto é grande, a distância média do ponto dentro do cluster é menor que a distância média até os pontos no cluster vizinho, ou seja, o ponto está bem classificado.
- Se essa medida for pequena, a distância média do ponto dentro do grupo é maior que a distância média aos objetos no cluster vizinho, por isso o ponto foi classificado de forma indevida. ->
- O coeficiente de silhueta de todo o set é definido pela média dos coeficientes calculados para cada ponto.

$$sil(C) = \overline{sil(k)} = \frac{1}{k} \sum_{i=1}^k sil(C_i)$$

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

Sendo que:

"a" é a distância média entre o ponto e todos os demais pontos do cluster.

"b" é a distância média entre o ponto e todos os pontos do cluster vizinho mais próximo.

Análise de Cluster

Padronização

- Utilizando a Distância Euclidiana, o que impacta mais em um cálculo:

Uma diferença de 100 reais na renda anual????

ou

Uma diferença de 80 anos na idade????

- Quando temos variáveis com escalas diferentes, precisamos nos preocupar com a padronização dos dados.

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Padronização de variável

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

X: variável aleatória com média μ e desvio padrão S

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

Análise de Cluster

Prática no R – Hierárquico

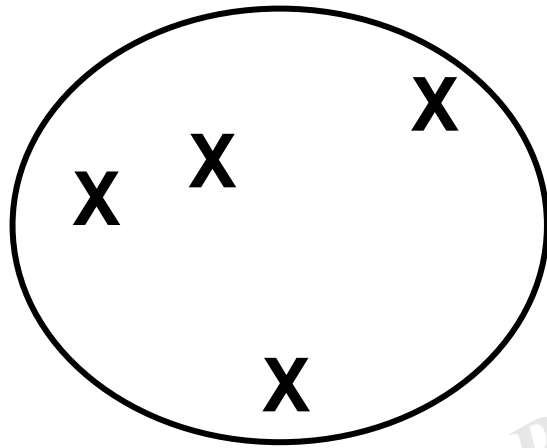
- Exemplo do Alunos no R
- Mcdonald

Paola Lucca Pizutti 006.246.810-33

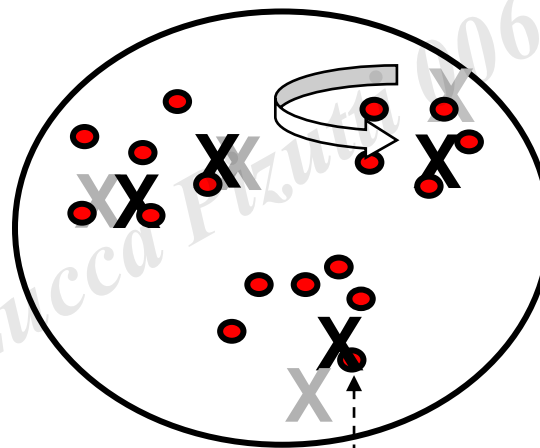
Análise de Cluster

Não Hierárquico

Sementes Iniciais



Centróides Finais



observações

Análise de Cluster

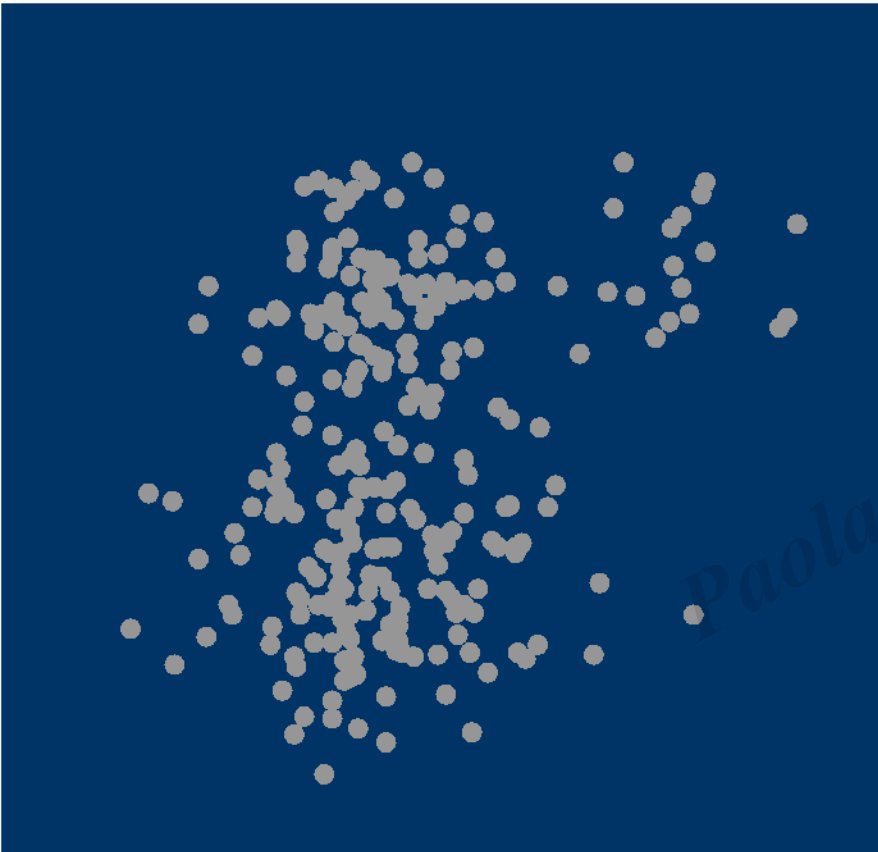
Não Hierárquico – *k-means*

A metodologia *k-means* segue 3 passos:

1. Seleciona (ou especifica) os centróides iniciais (sementes).
2. Lê as observações e atualiza os centróides, esse processo é repetido até a convergência.
3. Uma leitura final dos dados assinala cada observação ao centróide mais próximo.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

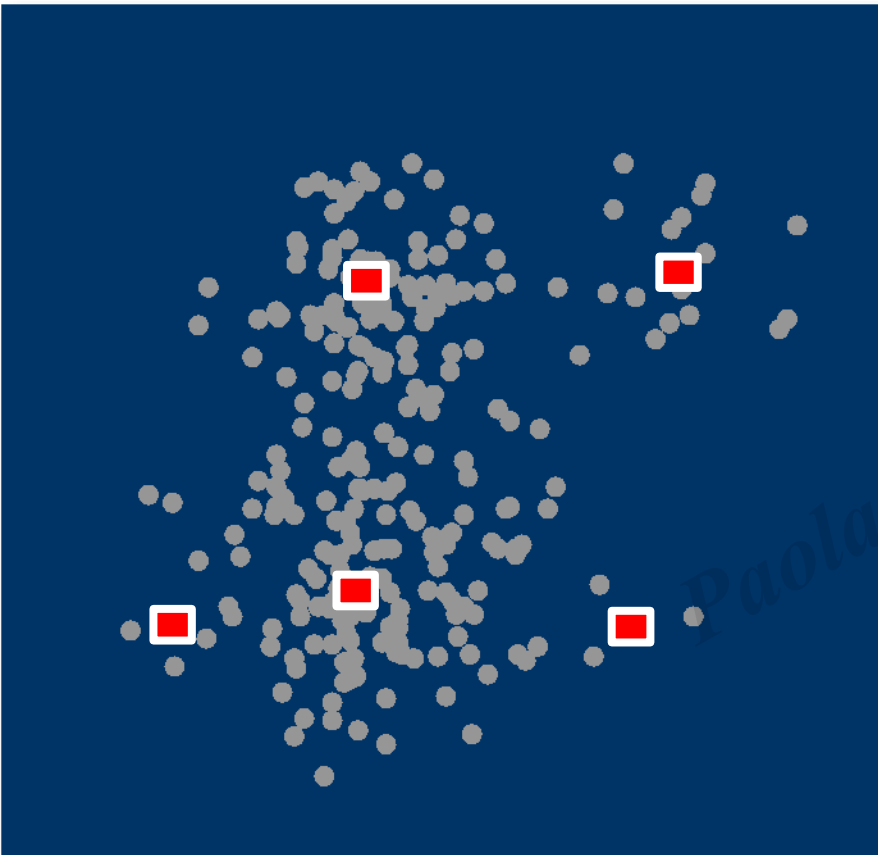
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

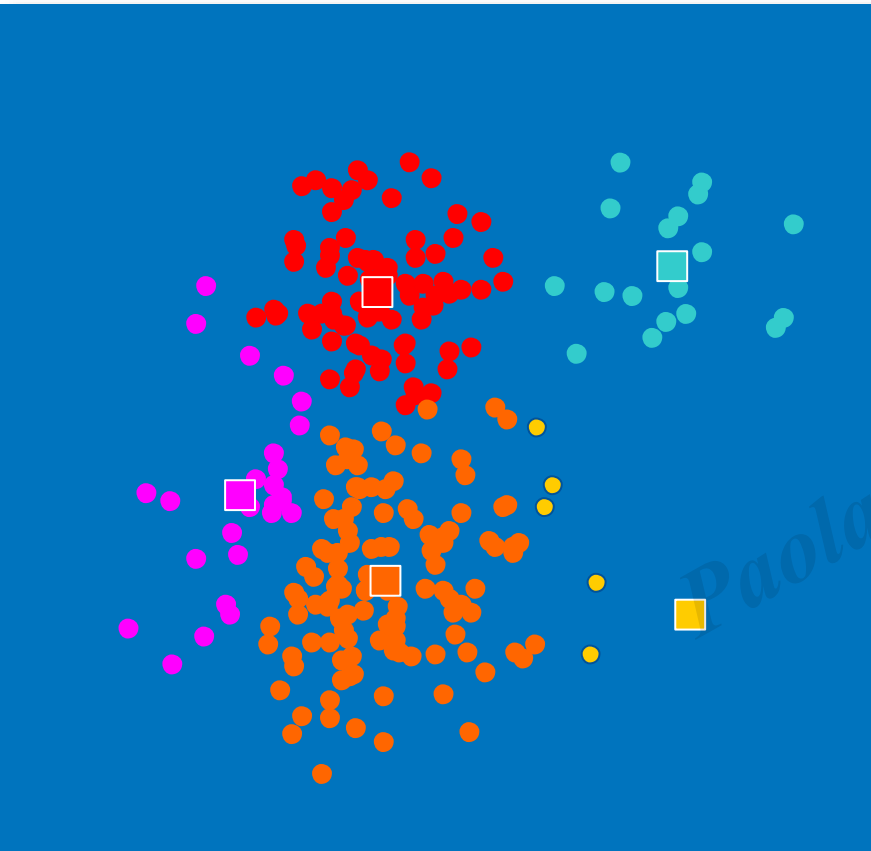
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

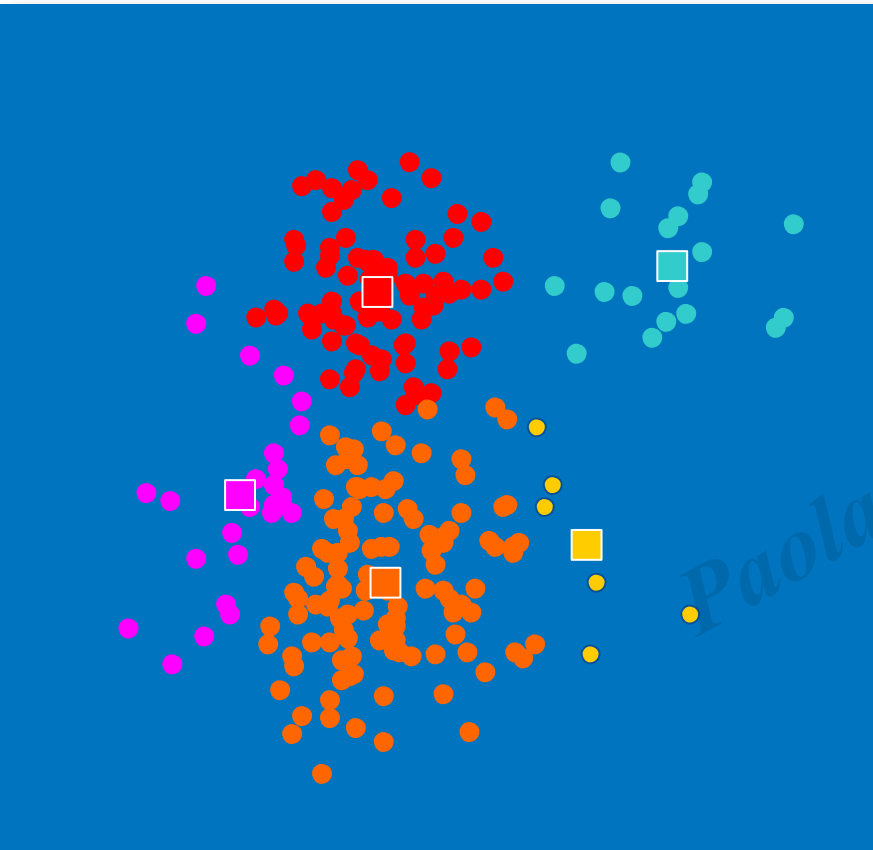
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

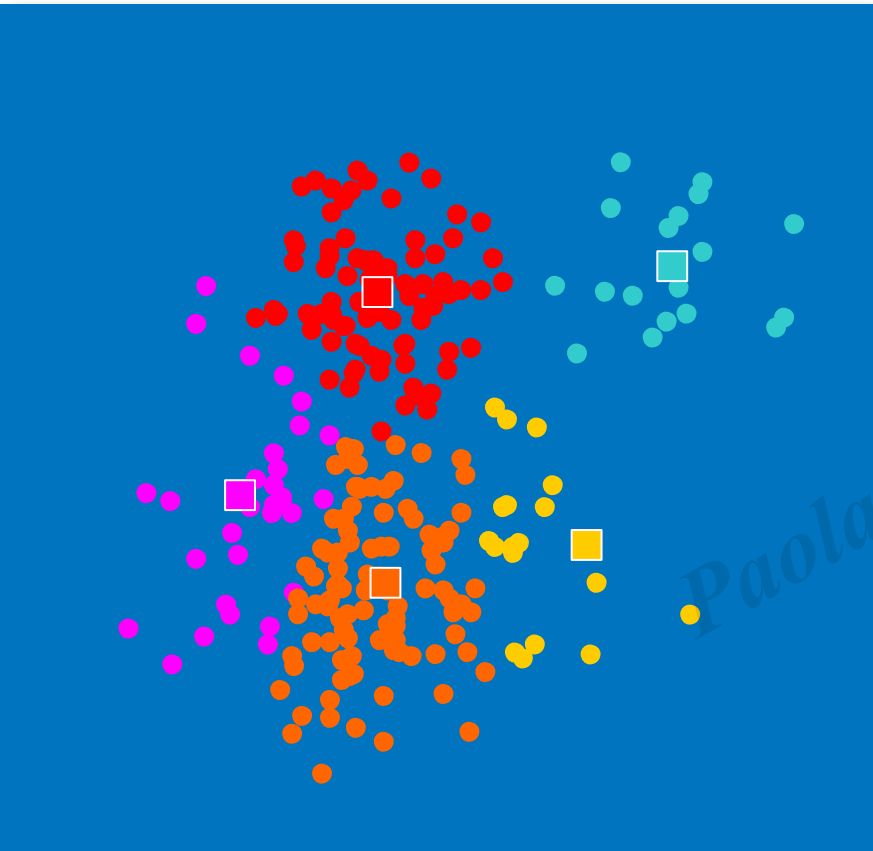
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

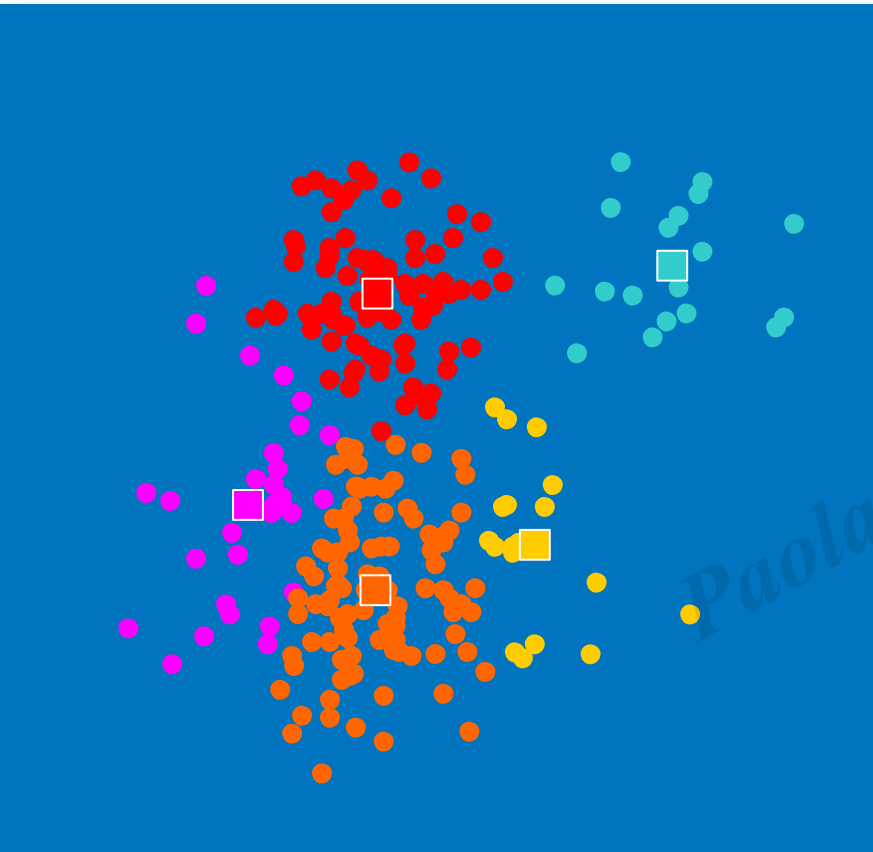
3. Atualiza os centróides.

4. **Re-assinala todos os casos.**

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

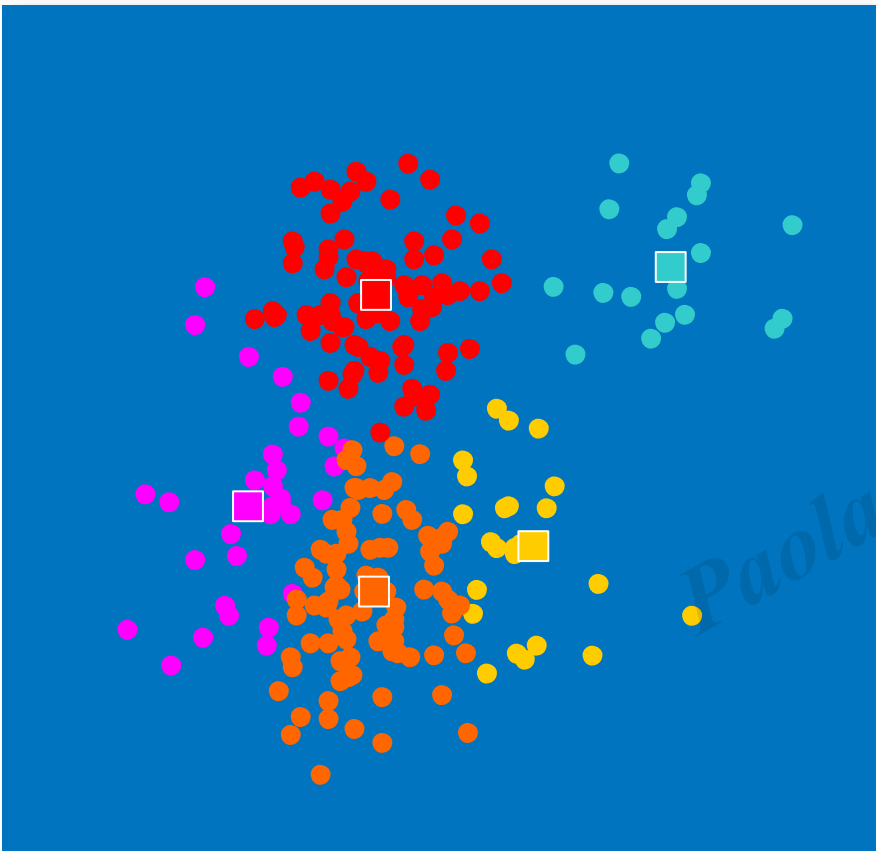
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

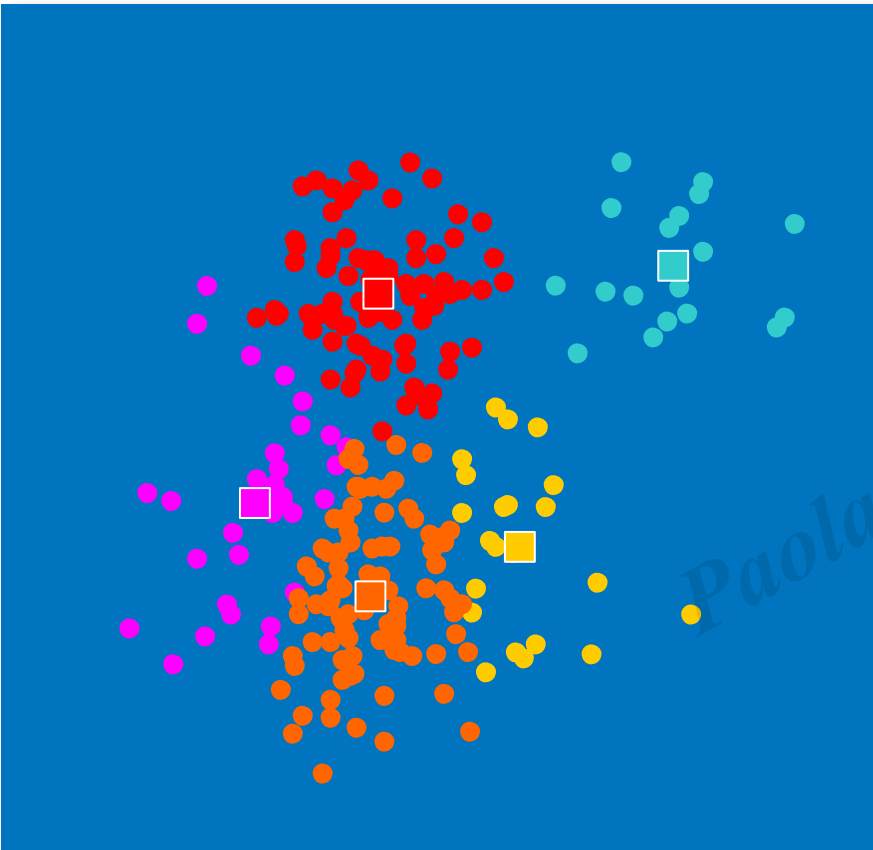
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

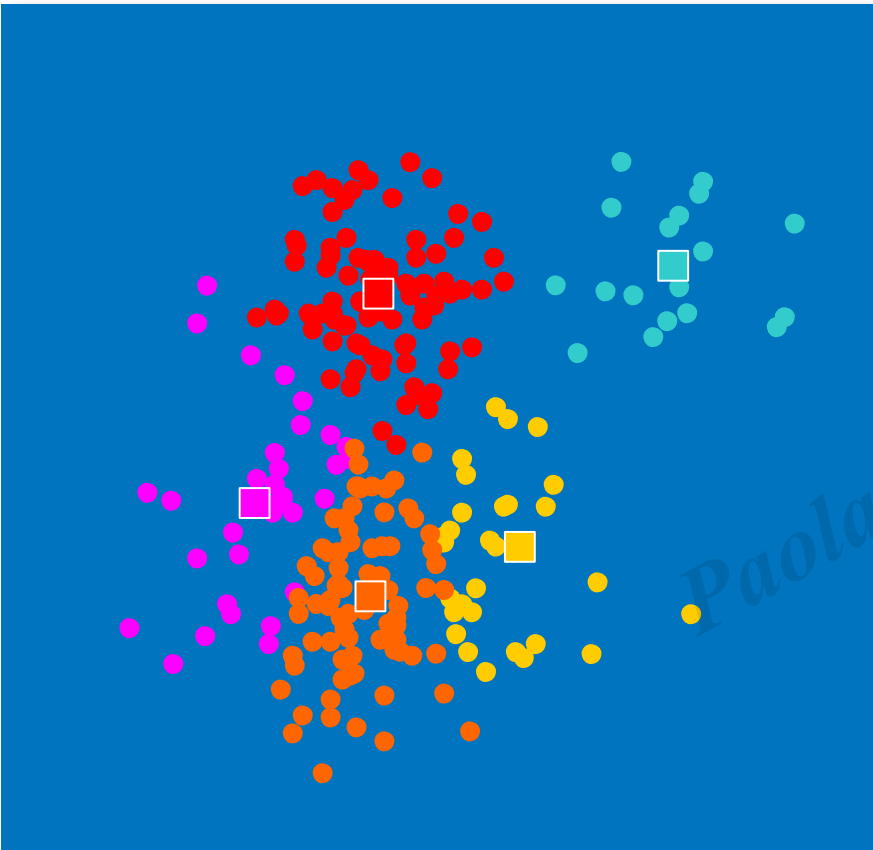
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

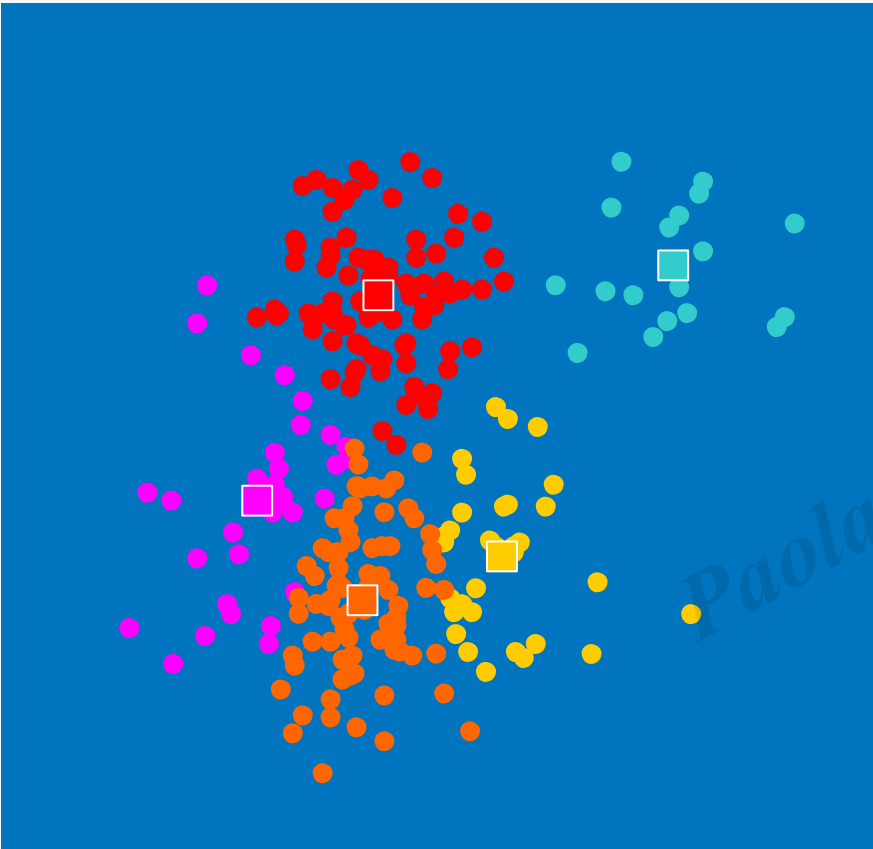
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

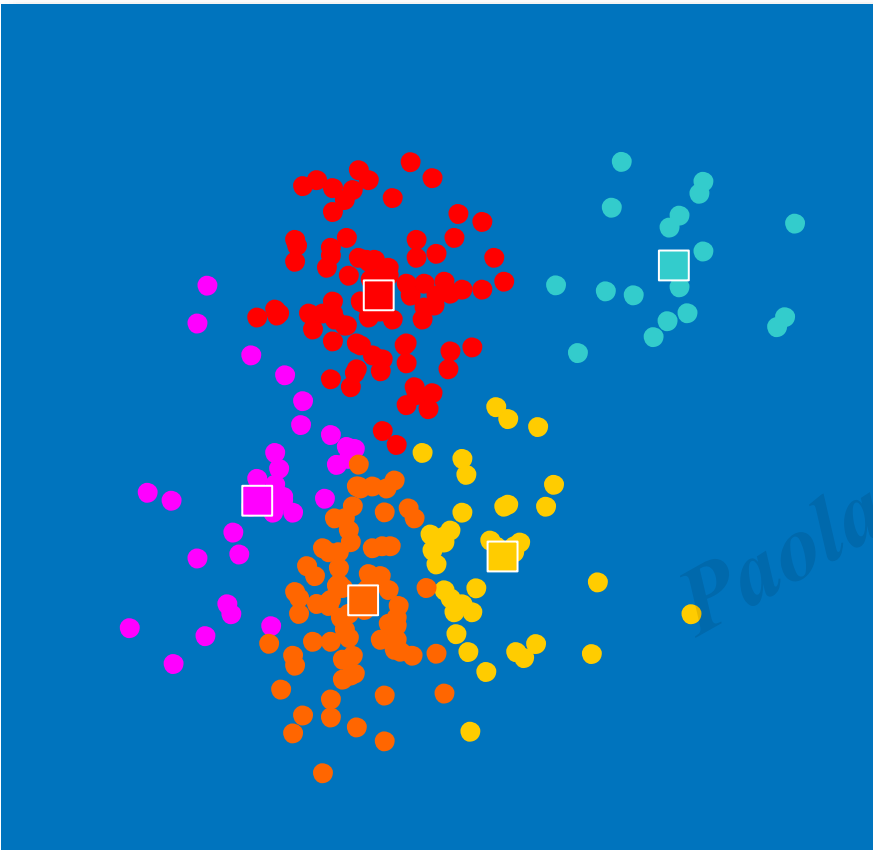
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

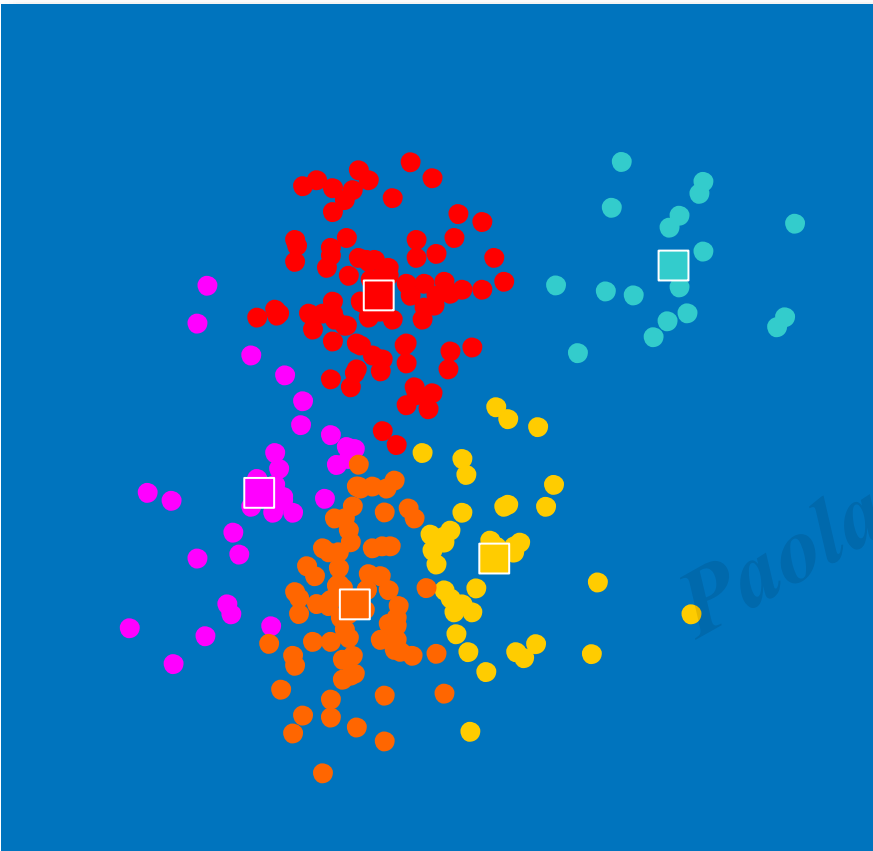
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

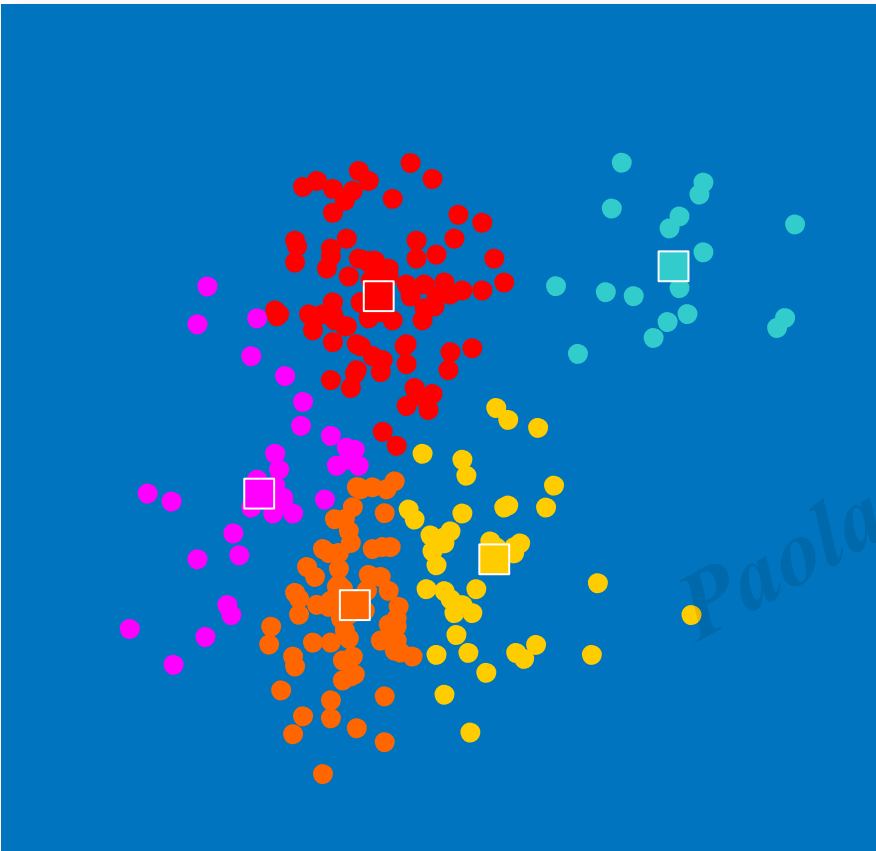
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

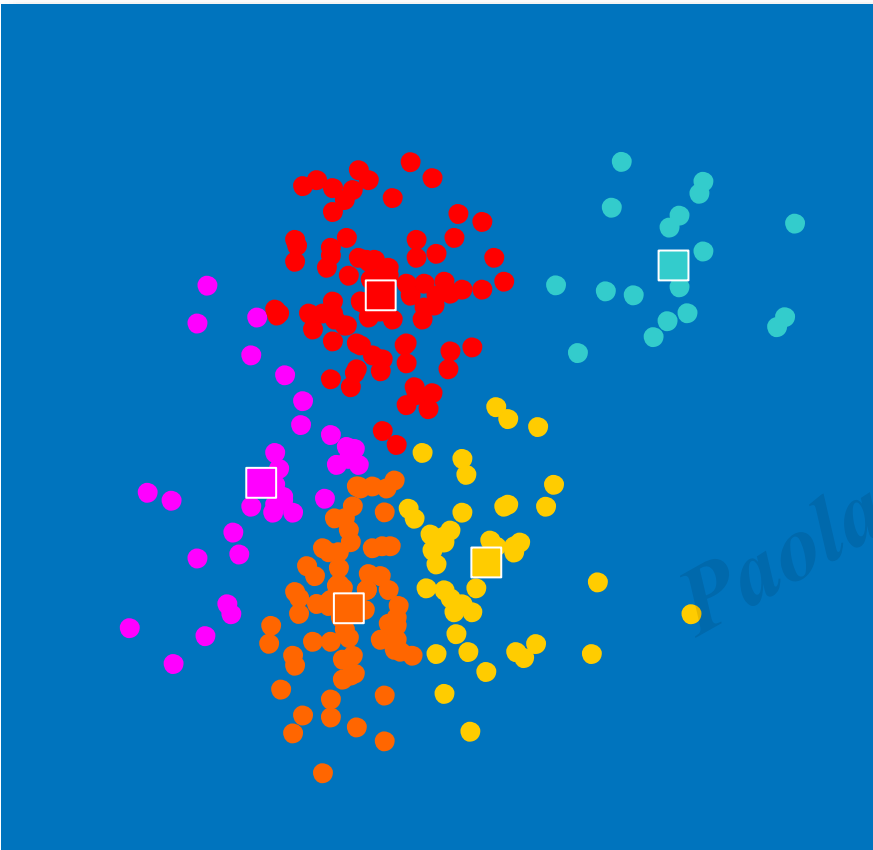
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

2. Assinala casos para o centróide mais perto.

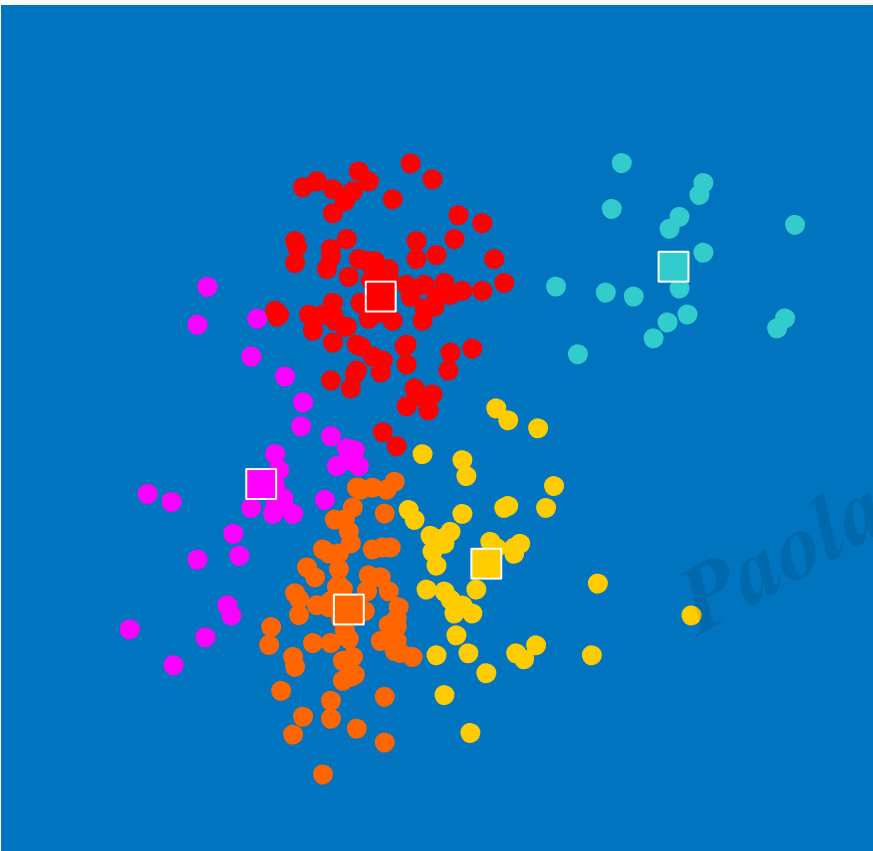
3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster

Não Hierárquico – *k-means*



1. Seleciona k centróides .

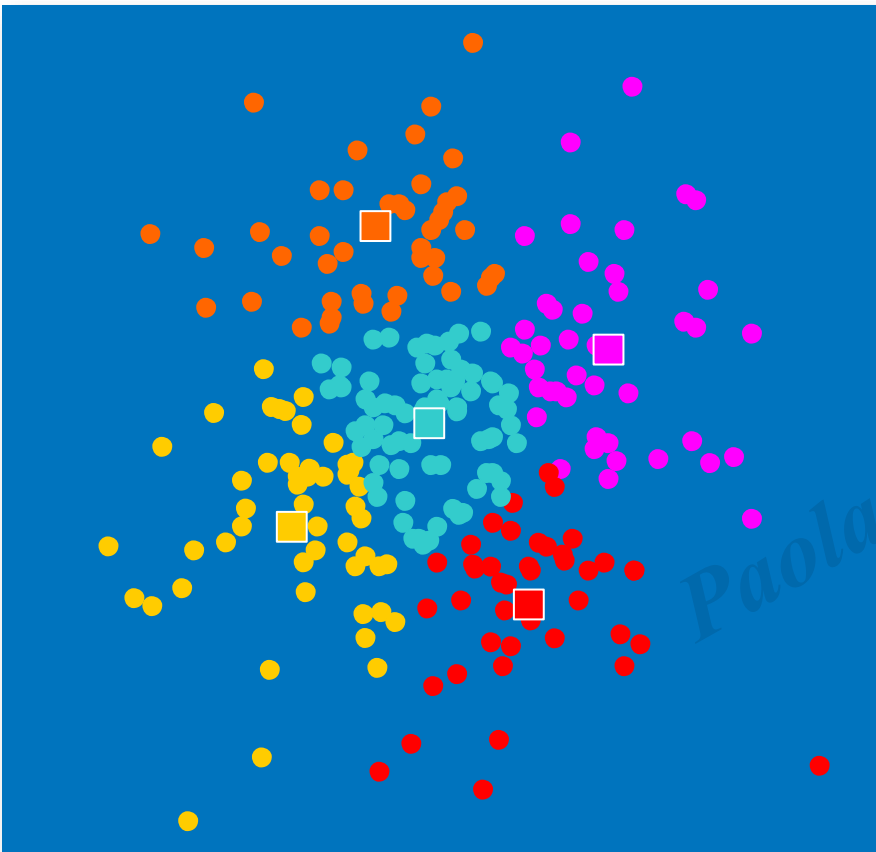
2. Assinala casos para o centróide mais perto.

3. Atualiza os centróides.

4. Re-assinala todos os casos.

5. Repita os passos 3 e 4 até a convergência.

Análise de Cluster



Quando os clusters não ocorrem naturalmente o algoritmo tende a dividir a tabela em partes iguais.

Análise de Cluster

Prática no R – Não Hierárquico

- Mcdonald
- Municípios

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

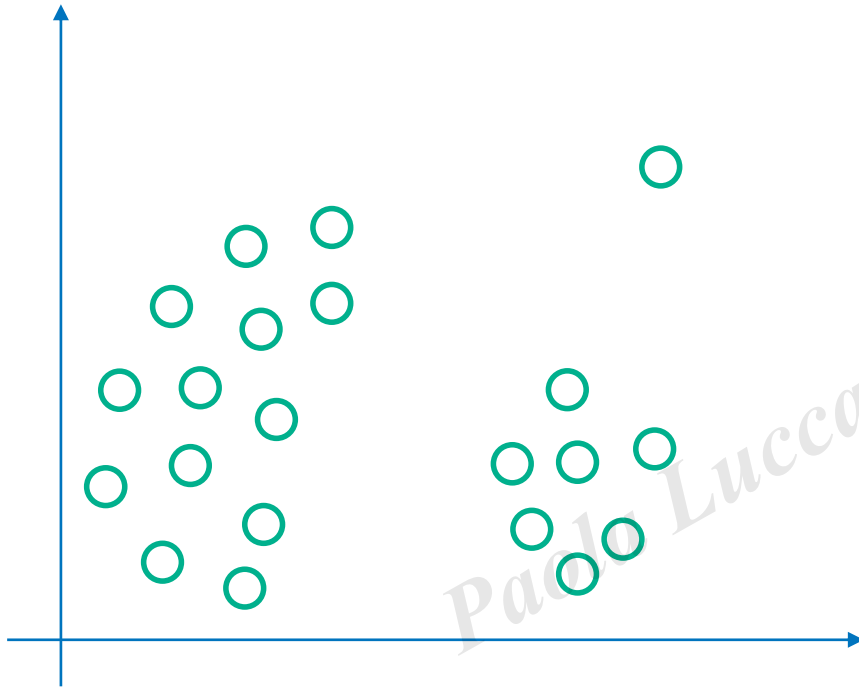
Dbscan

- “Density Based Spatial Clustering of Application with Noise”
- (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído)
- Dois parâmetros
 - Eps: raio para definição de ponto mais próximo
 - MinPts: número mínimo de pontos no Eps-vizinhos de um ponto

Paola Lucca Pizutti 006.246.810-33

Análise de Cluster

Dbscan



- Papel dos pontos
 - Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
 - Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
 - Outlier: não tem pontos no raio Eps.

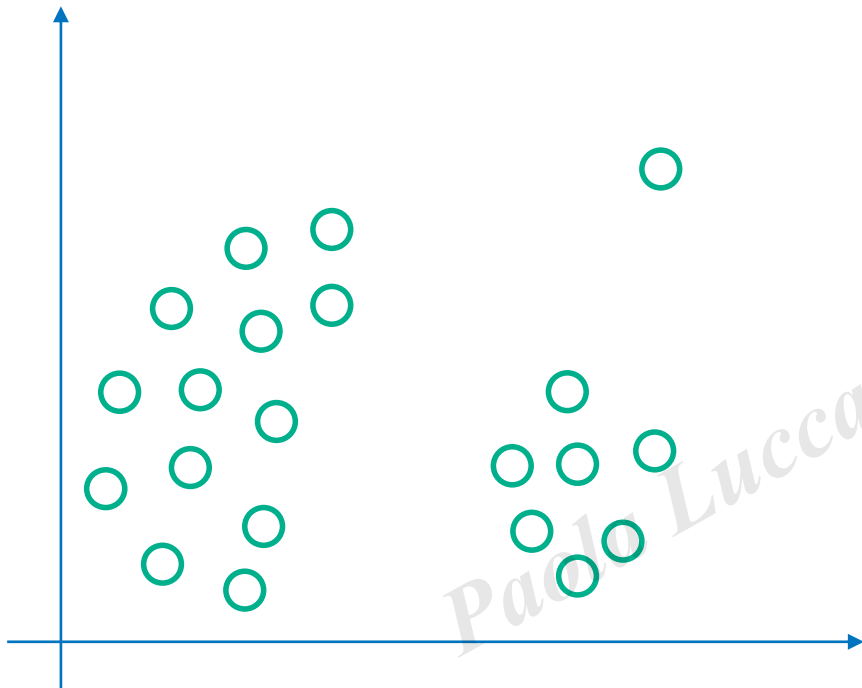
Análise de Cluster

Dbscan

- O método DBSCAN encontra clusters verificando a vizinhança **Eps** de cada ponto na base de dados, começando por um **objeto arbitrário** p . Se p é um ponto central, um novo cluster com p como um centro é criado. Se p é um ponto de fronteira, nenhum ponto é alcançável por densidade a partir de p e DBSCAN visita o próximo ponto na base. O método DBSCAN, então, iterativamente coleta objetos alcançáveis por densidade diretamente de pontos centrais, que pode envolver a união de alguns clusters alcançáveis por densidade. O processo termina quando nenhum novo ponto pode ser adicionado a qualquer cluster. Para o algoritmo DBSCAN assim definido, quaisquer dois pontos centrais com distância menor ou igual a **Eps** são colocados no mesmo cluster. Qualquer ponto de fronteira que está perto de um ponto central é colocado no mesmo cluster do ponto central. Pontos que não são diretamente atingíveis por algum ponto central são classificados como ruído.

Análise de Cluster

Dbscan

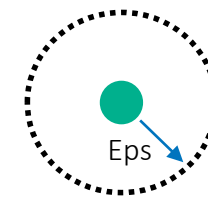


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

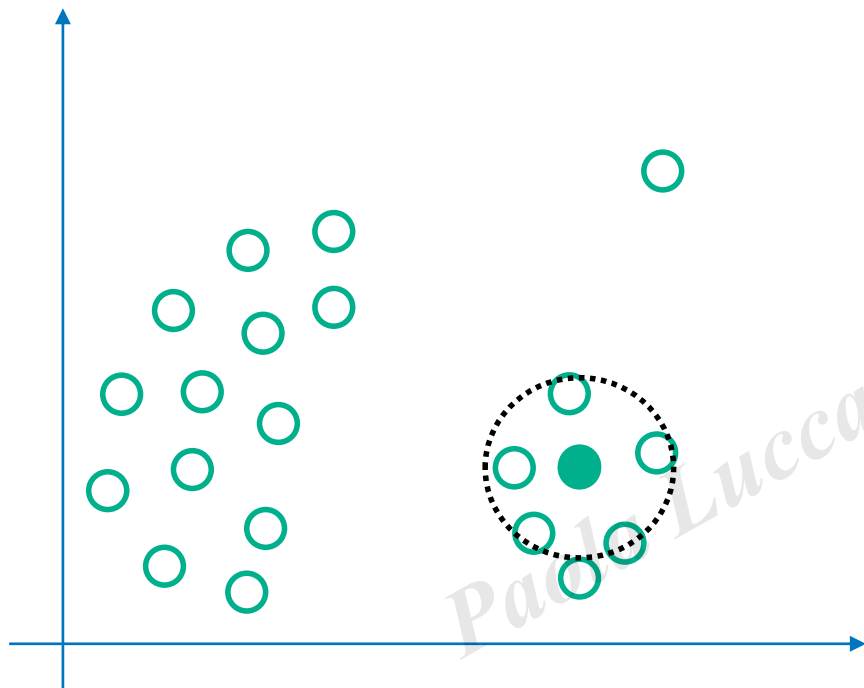
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

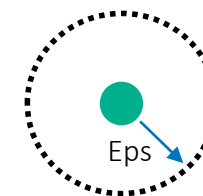


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

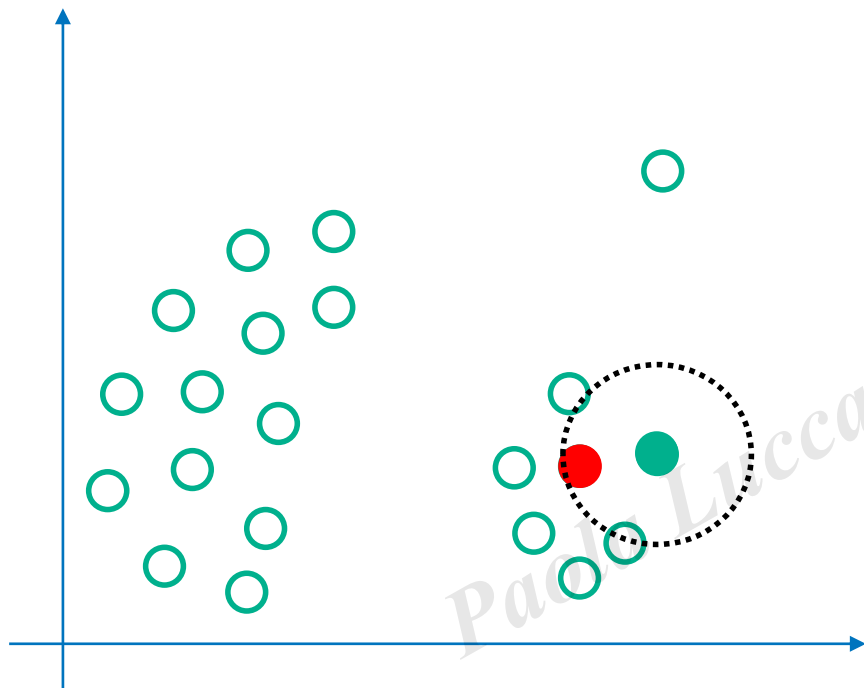
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

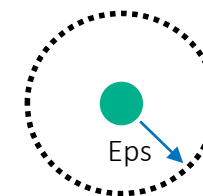


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

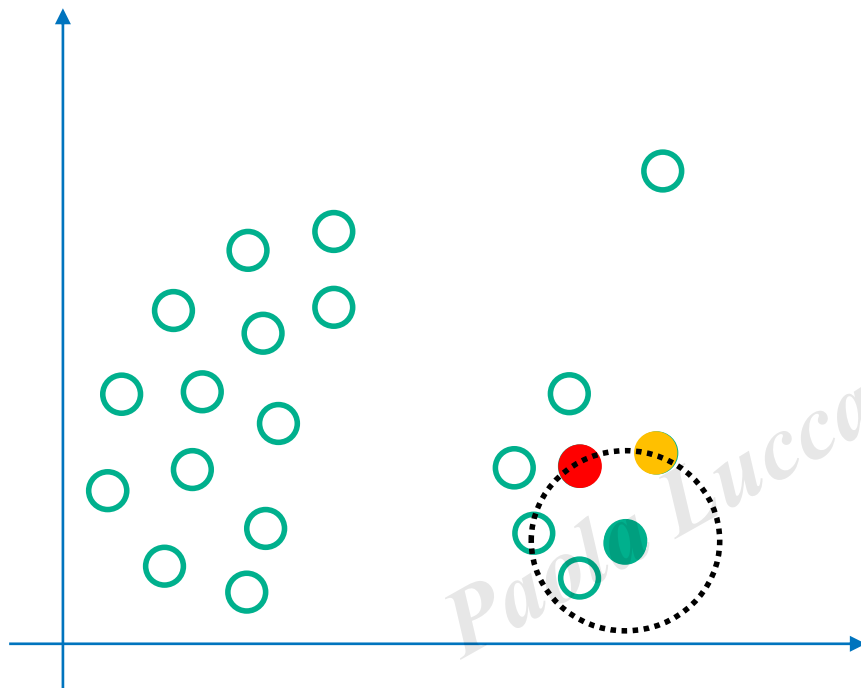
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

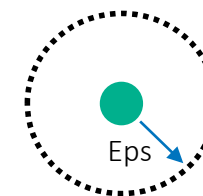


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

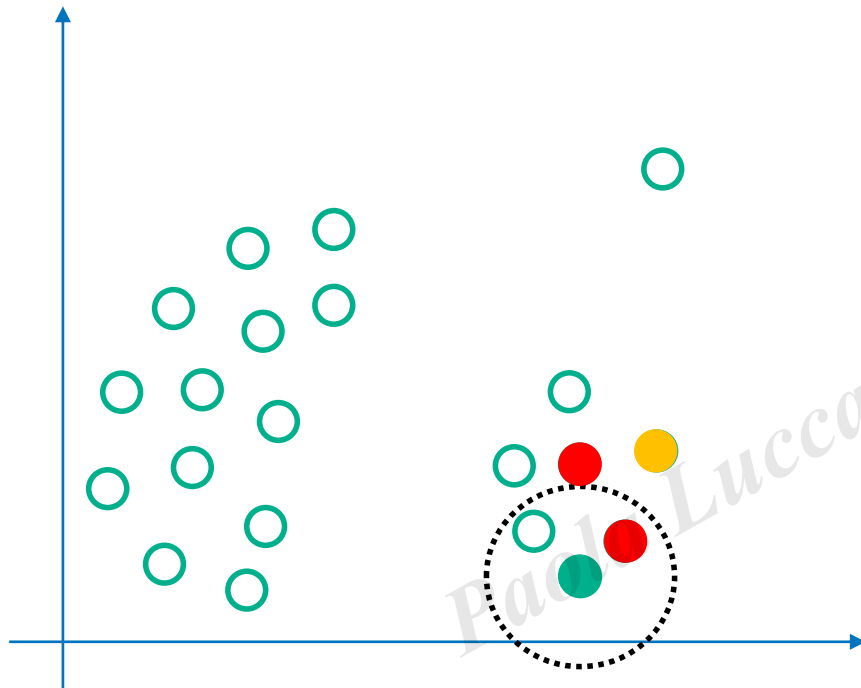
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

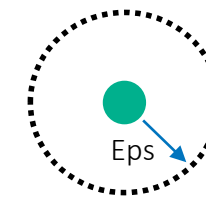


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

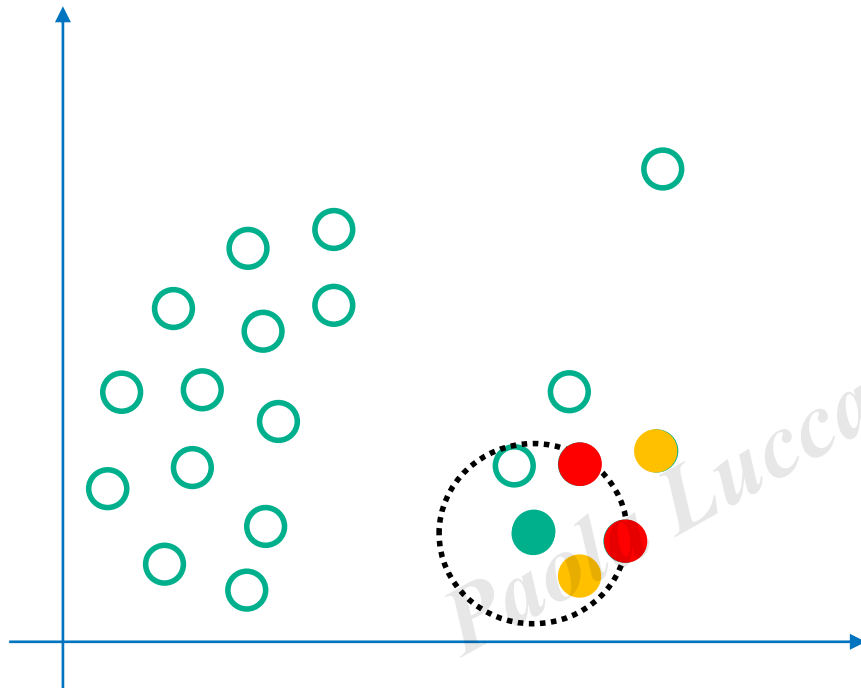
- $Eps = 1$

- $MinPts = 4$



Análise de Cluster

Dbscan

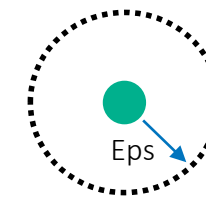


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

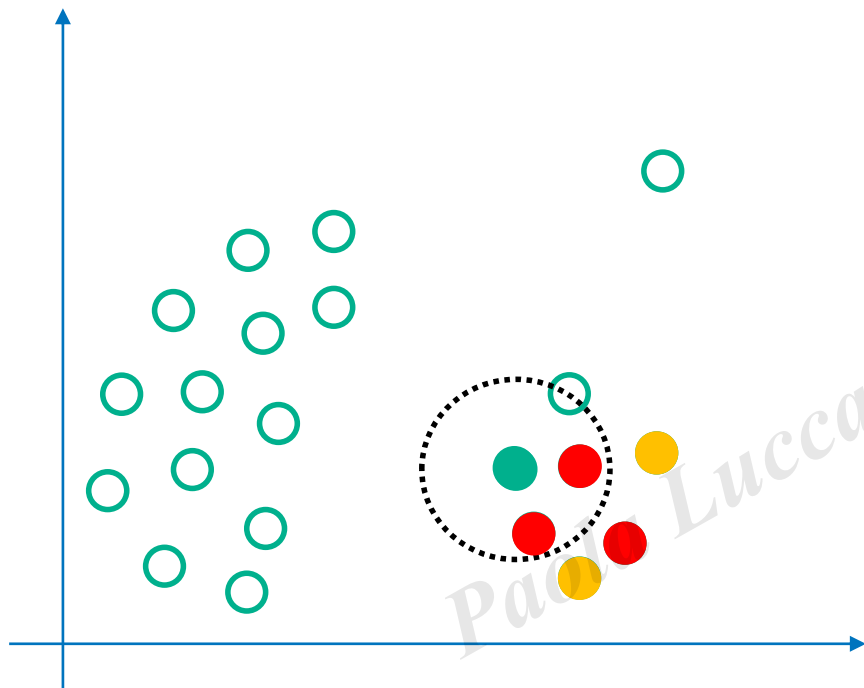
- $Eps = 1$

- $MinPts = 4$



Análise de Cluster

Dbscan

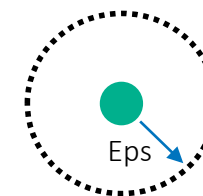


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

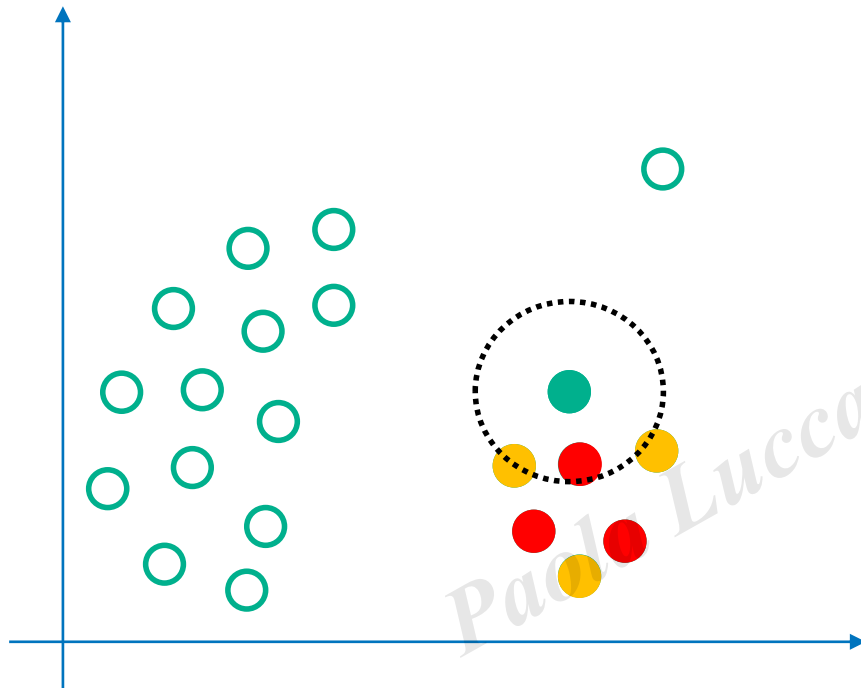
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

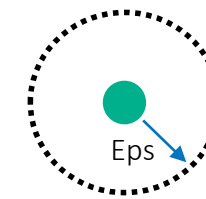


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

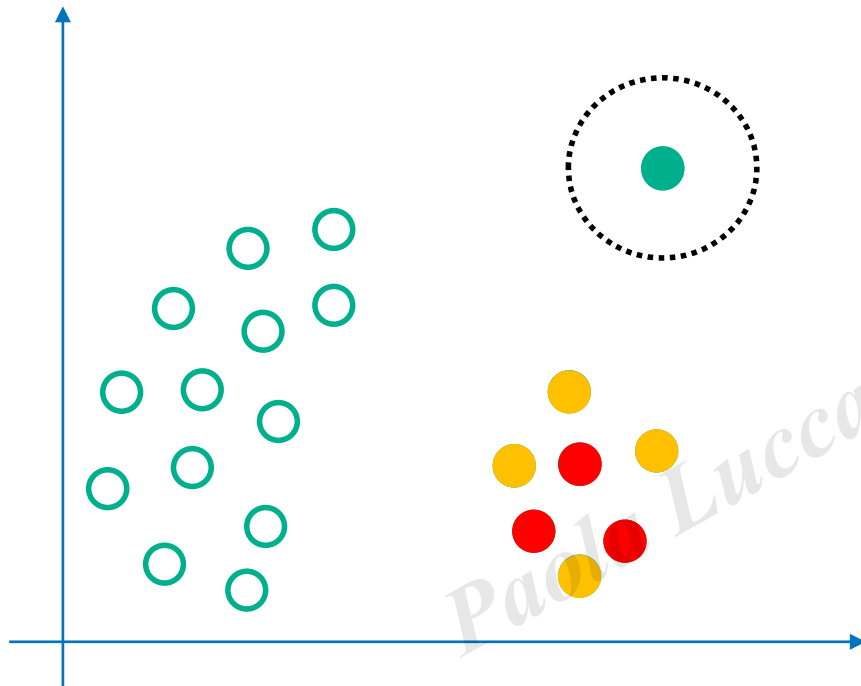
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

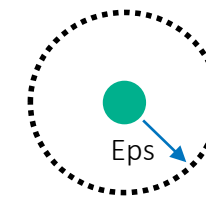


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

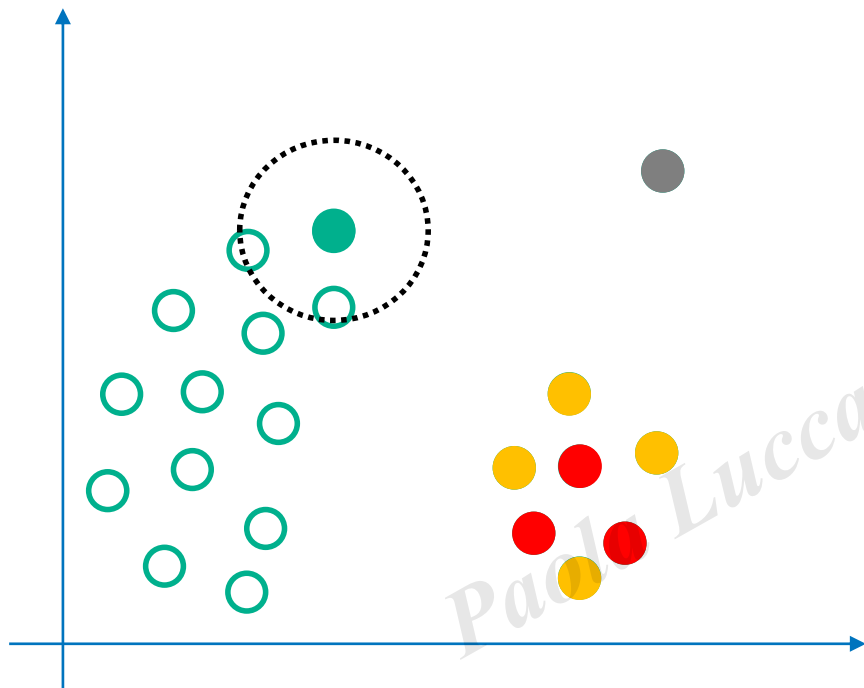
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

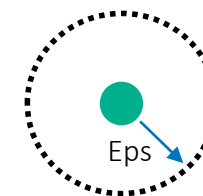


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

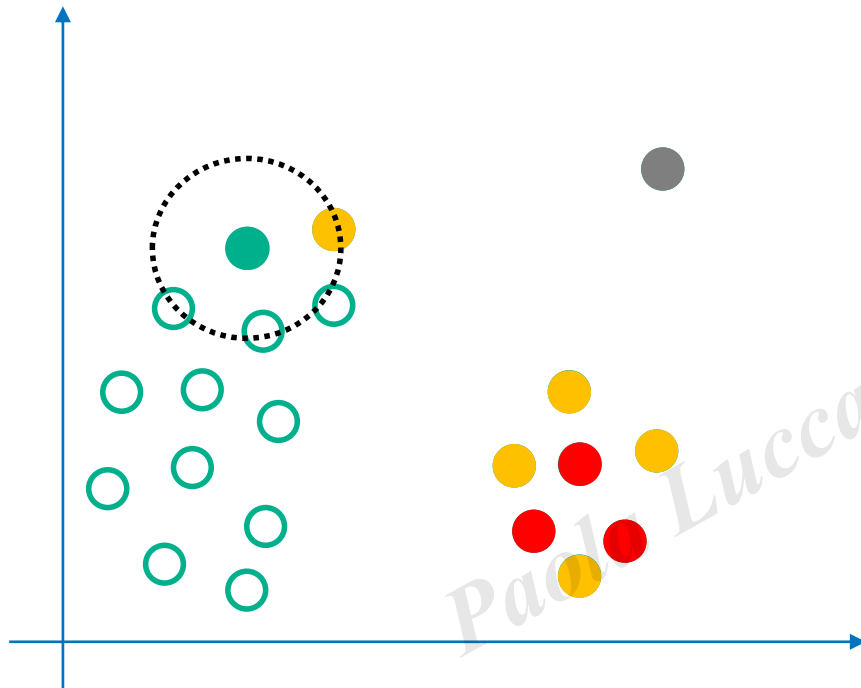
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

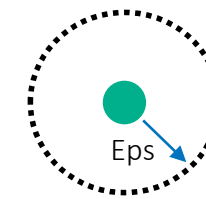


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

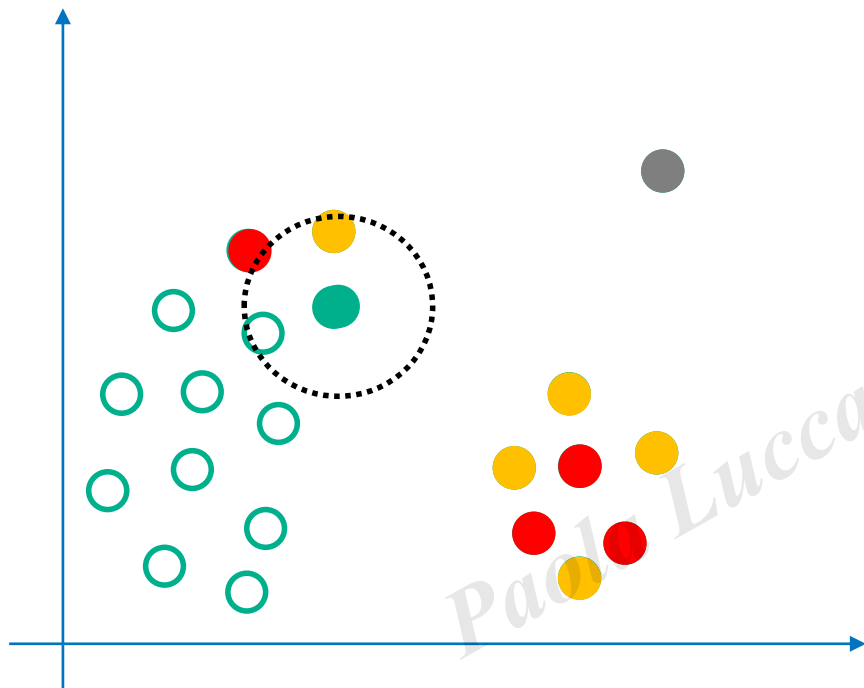
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

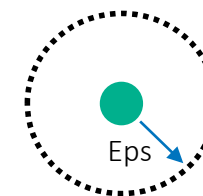


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

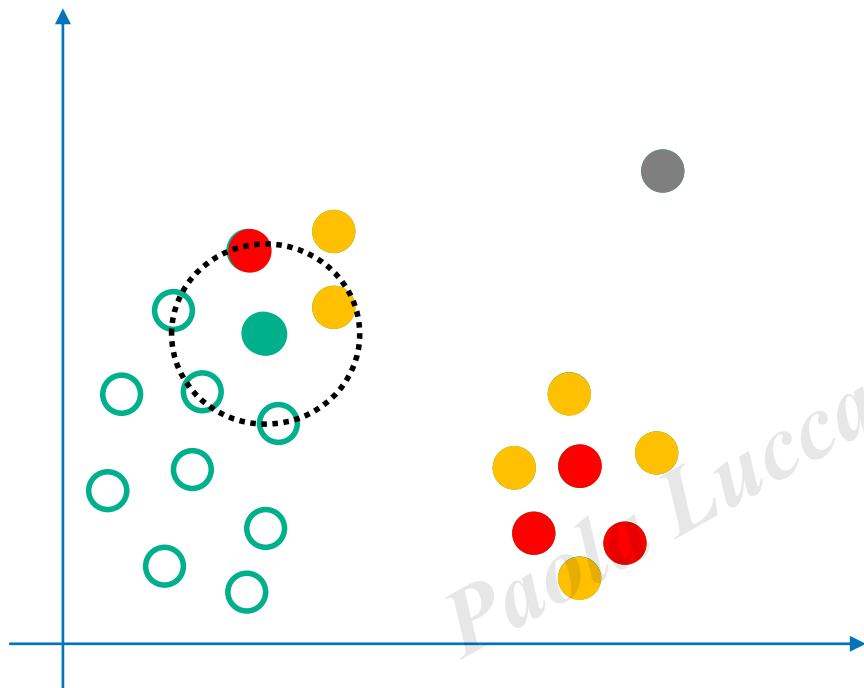
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

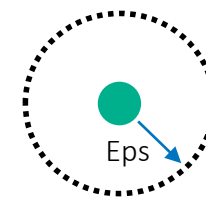


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

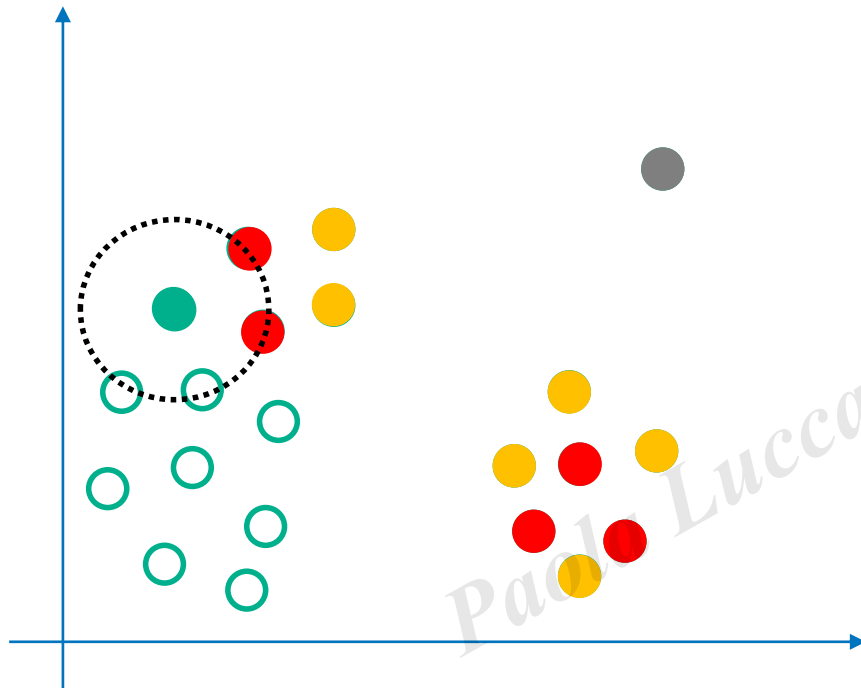
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

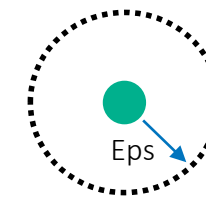


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

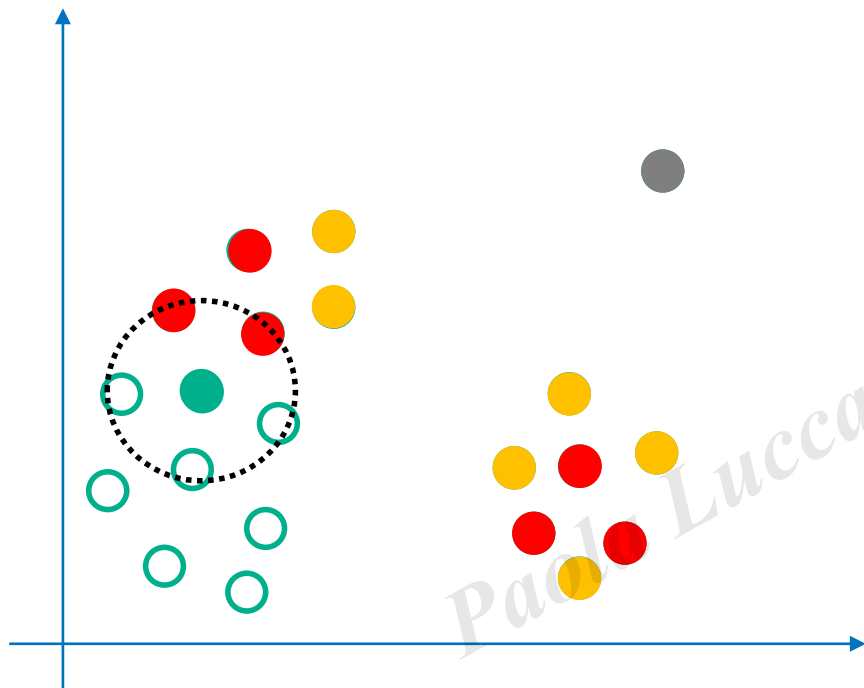
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

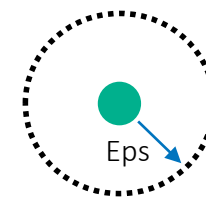


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

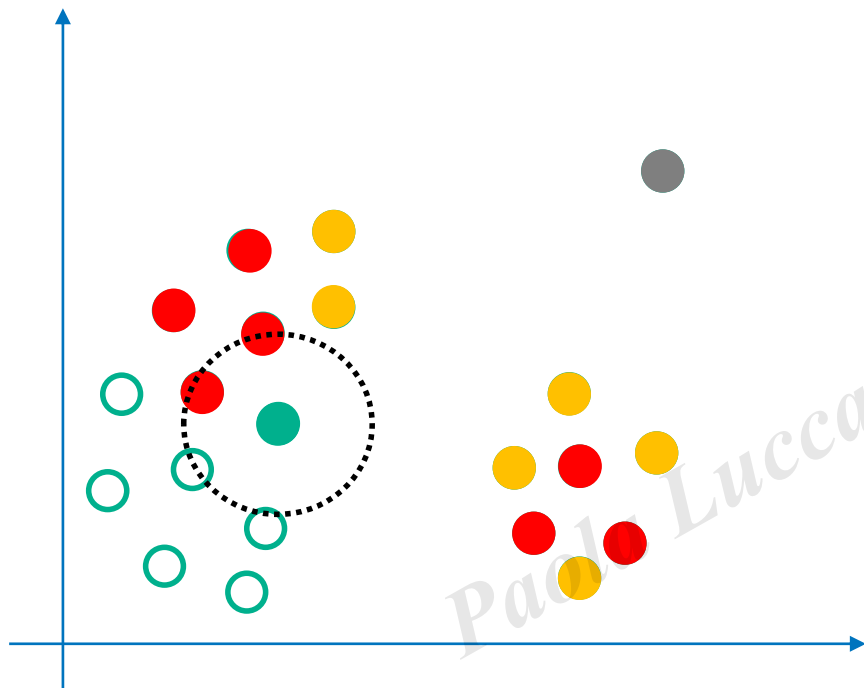
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

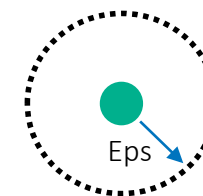


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

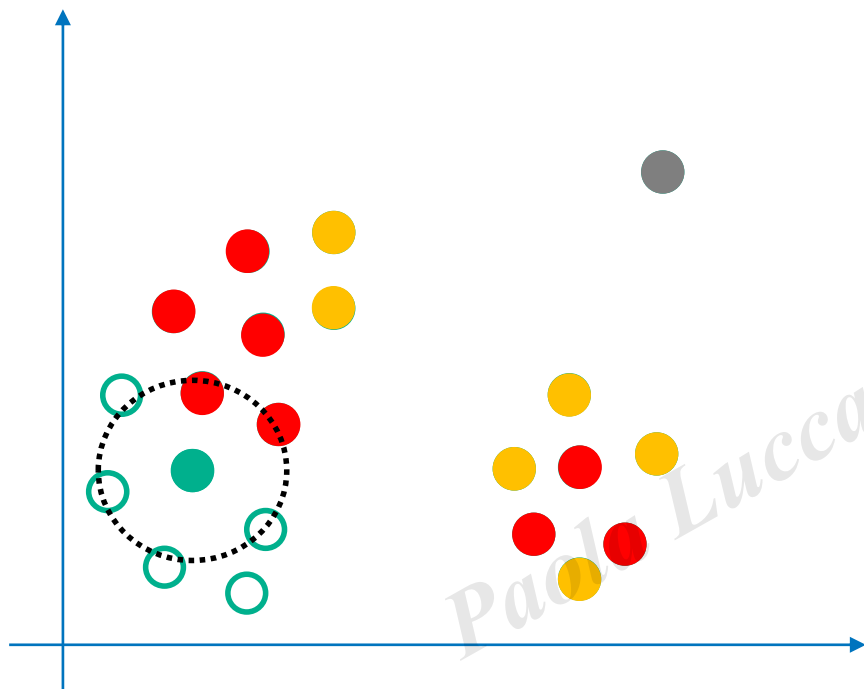
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

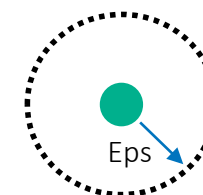


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

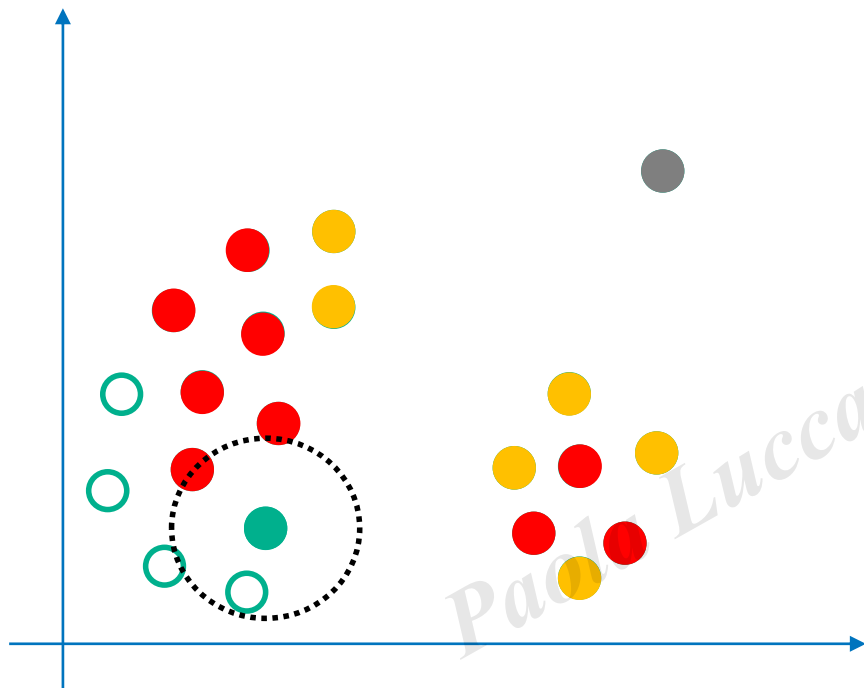
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

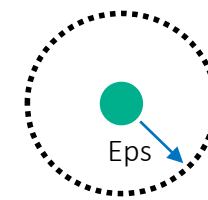


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

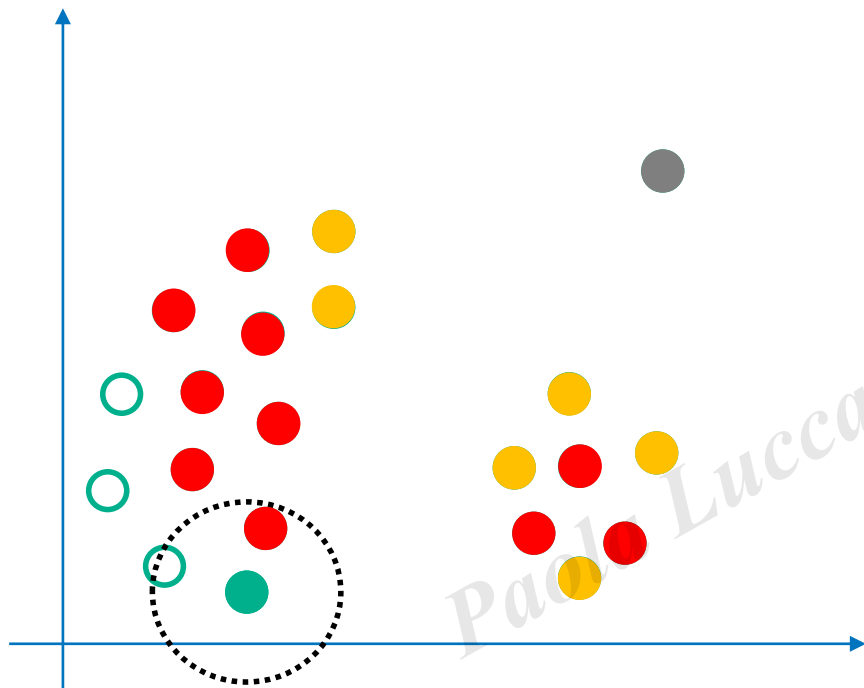
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

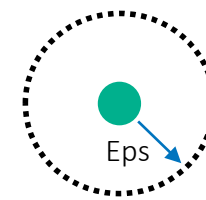


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

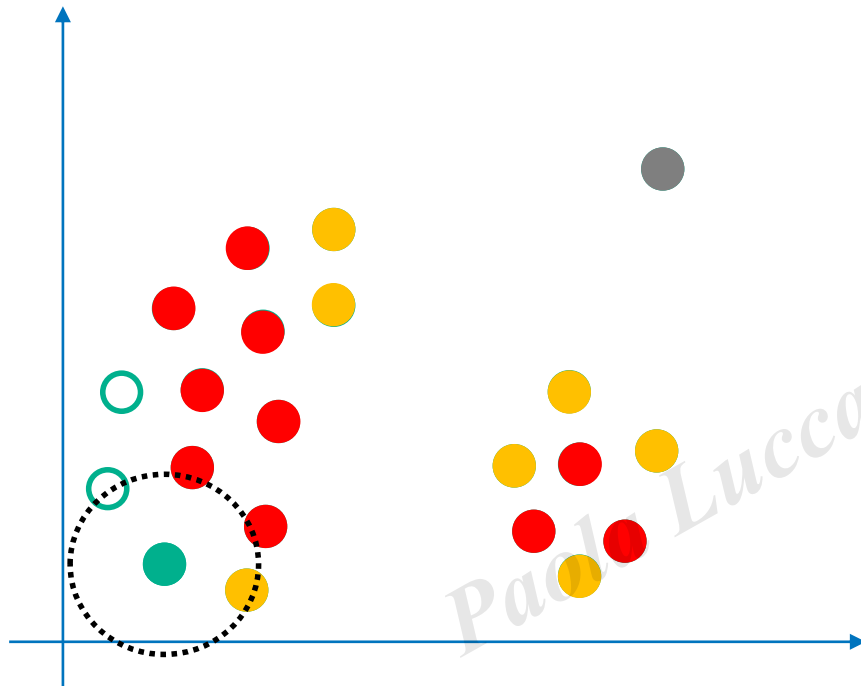
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

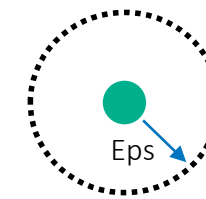


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

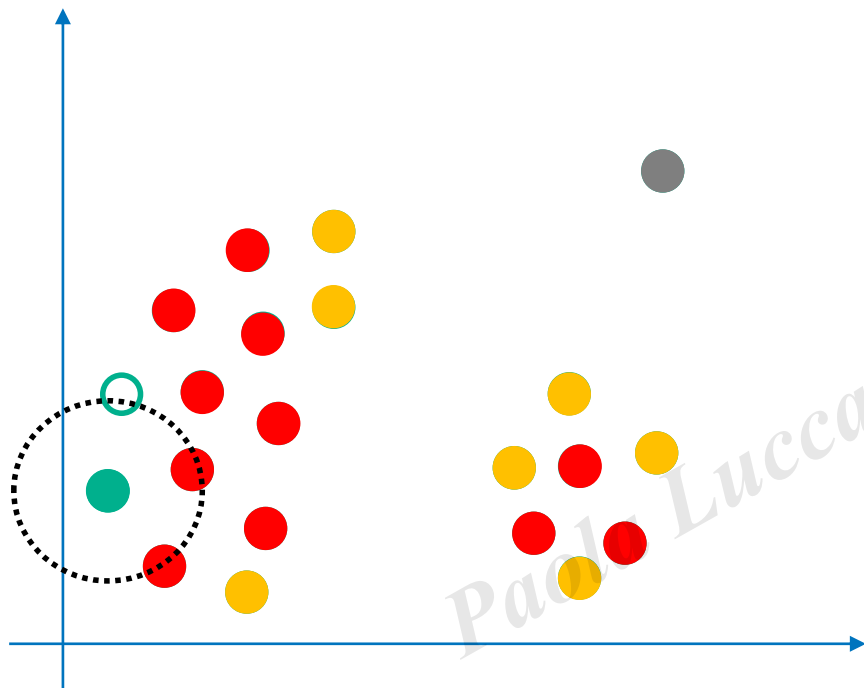
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

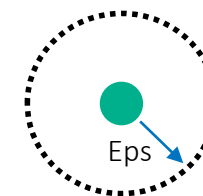


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

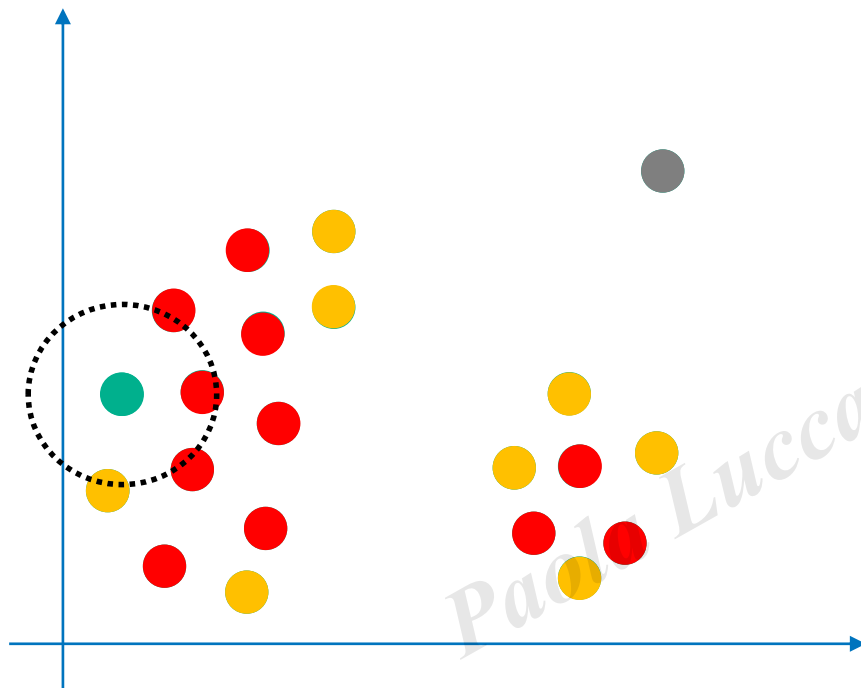
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

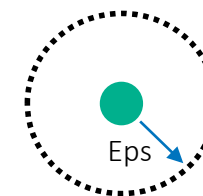


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

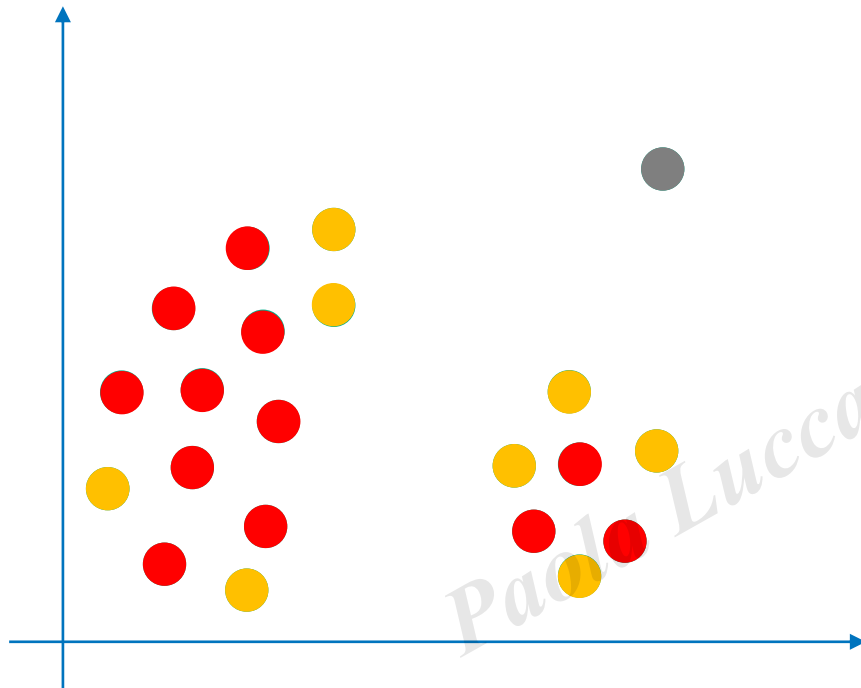
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

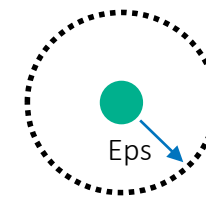


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

- Eps = 1

- MinPts = 4



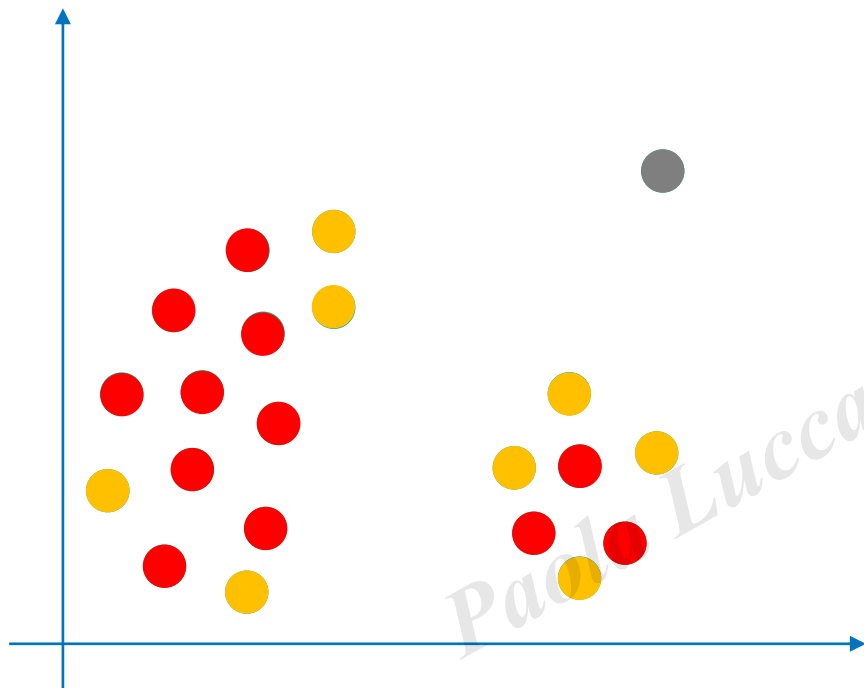
Análise de Cluster

Dbscan

- O método DBSCAN encontra clusters verificando a vizinhança **Eps** de cada ponto na base de dados, começando por um **objeto arbitrário** p . Se p é um ponto central, um novo cluster com p como um centro é criado. Se p é um ponto de fronteira, nenhum ponto é alcançável por densidade a partir de p e DBSCAN visita o próximo ponto na base. O método DBSCAN, então, iterativamente coleta objetos alcançáveis por densidade diretamente de pontos centrais, que pode envolver a união de alguns clusters alcançáveis por densidade. O processo termina quando nenhum novo ponto pode ser adicionado a qualquer cluster. Para o algoritmo DBSCAN assim definido, quaisquer dois pontos centrais com distância menor ou igual a **Eps** são colocados no mesmo cluster. Qualquer ponto de fronteira que está perto de um ponto central é colocado no mesmo cluster do ponto central. Pontos que não são diretamente atingíveis por algum ponto central são classificados como ruído.

Análise de Cluster

Dbscan

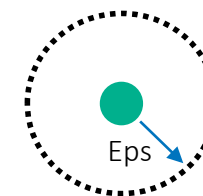


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

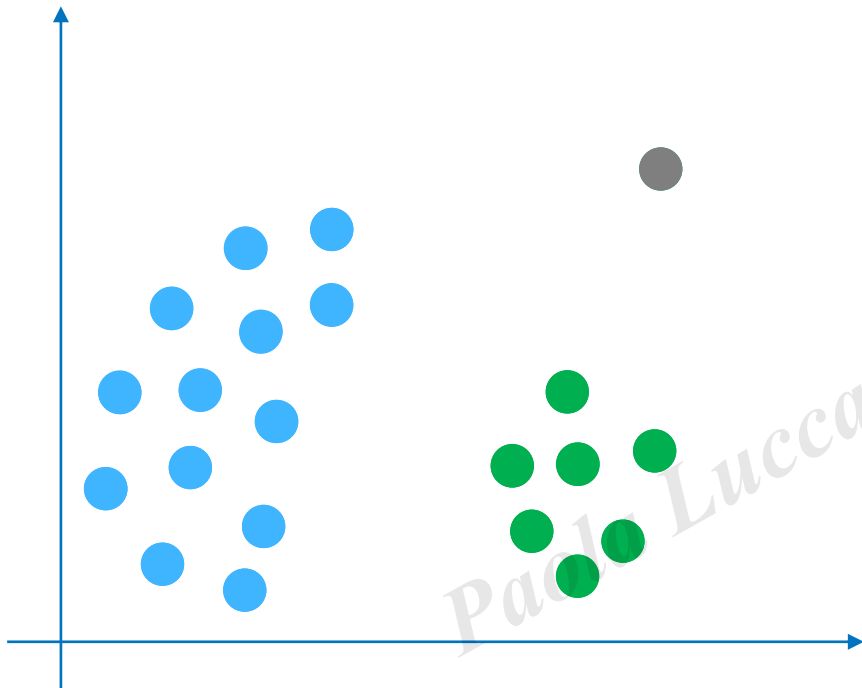
- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

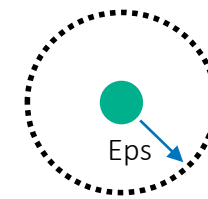


- Papel dos pontos

- Core (ponto central): tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (denso)
- Border (ponto de fronteira): não tem MinPts vizinhos dentro do Eps-vizinhos mais próximos (pouco denso), mas tem pontos no raio Eps.
- Outlier: não tem pontos no raio Eps.

- Eps = 1

- MinPts = 4



Análise de Cluster

Dbscan

Dados originais



Clusterização por
k-means



Clusterização por
DBSCAN



Análise de Cluster

Dbscan

- Fazer cluster pela técnica dbscan

Paola Lucca Pizutti 006.246.810-33



SEU
Novo
Mundo



Paola Lucca Pizutti 006.246.810-33

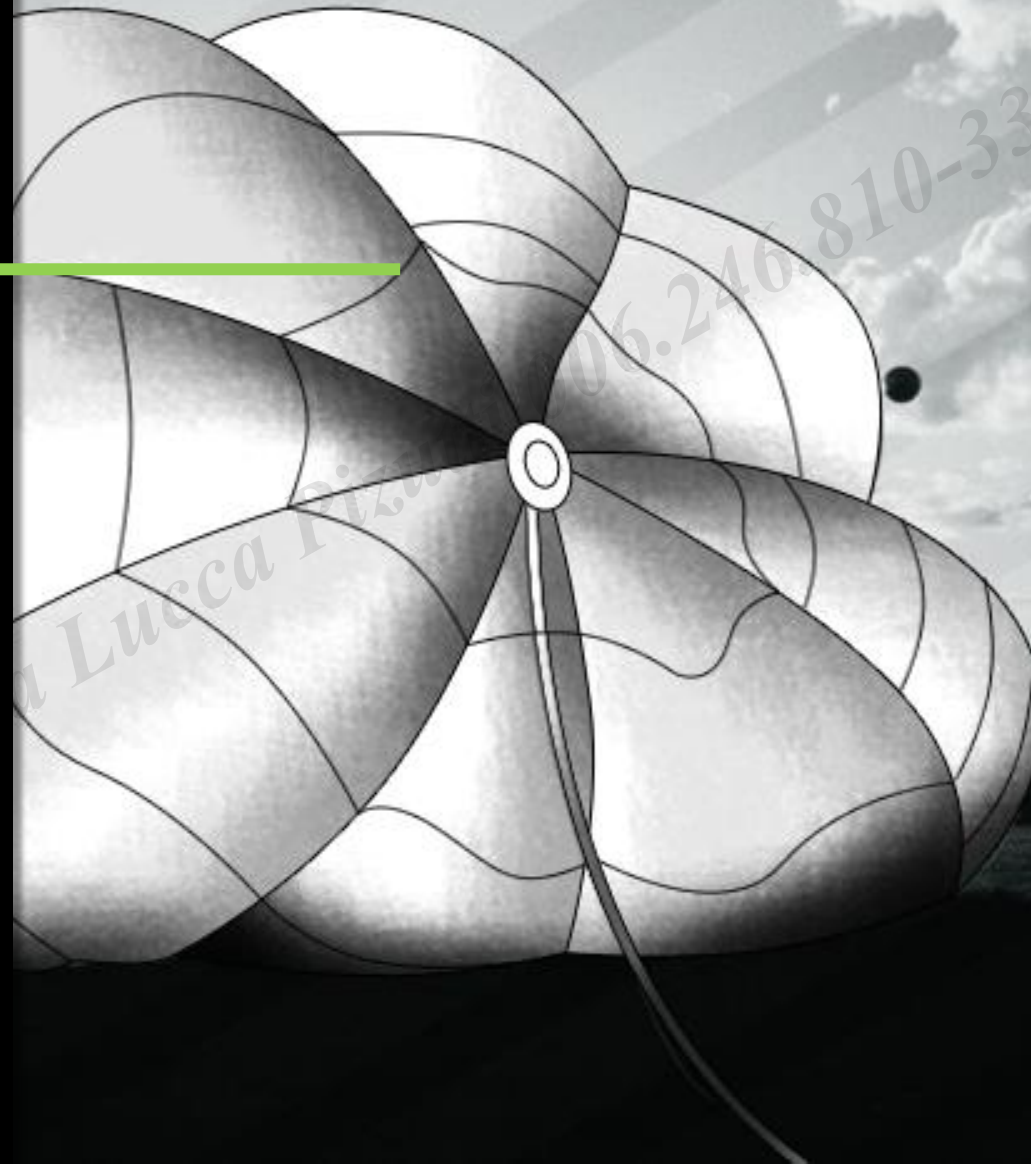
COMO É HOJE

- Grande esforço de manipulação de dados
- Análises univariadas
- Ferramentas não performáticas
- Dúvidas sobre a qualidade da execução



PASSO A PASSO E DESCOBERTAS

- SE O SHARE ESTÁ ABAIXO OU ACIMA DA MÉDIA DA UNIDADE
- SE O SHARE ESTÁ ABAIXO OU ACIMA DA MÉDIA DO SETOR
- QUAL A MARGEM MÉDIA DO CLIENTE
- SE A MARGEM VEZES O VOLUME POSSIVEL DE VENDA, É POSITIVO





ANÁLISE MULTIVARIADA

- Cálculo de distância multivariadamente
- Diminuição da variabilidade no grupo
- Descrição de cada grupo segundo a estratégia desenhada

NOSSOS diferenciais

Análise de Cluster é uma técnicas estatística que visa criar grupos com pouca variabilidade interna e que os grupos sejam diferentes entre si. Em uma visão multivariada!



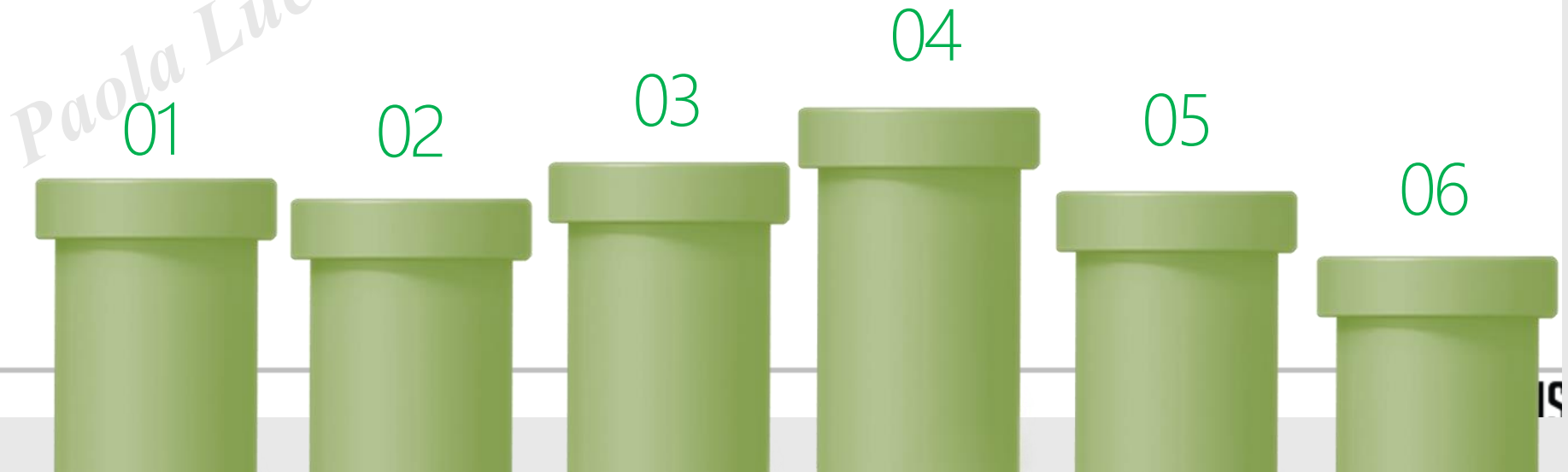
Lets play a GAME



Análise de Cluster é uma técnica estatística que visa criar grupos com pouca variabilidade interna e que os grupos sejam diferentes entre si. Em uma visão multivariada!



CLUSTER



Paola Lucca Pizutti 006.20639033



CLUSTER

Paola Lucca Pizutti 006.246.810-33

01

02

03

04

05

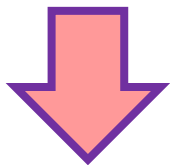
06



SHARE

P_A

0%



P_B

72%



P_C

87%



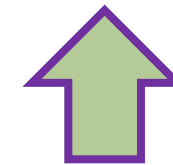
P_D

98%



P_E

97%



CLUSTER



02

03

04

05

06



Resumo

Produto	Num_treinamentos
P_A	63
P_B	17
P_C	35
P_D	47
P_E	17

Clus	P_A	P_B	P_C	P_D	P_E	Freq.	Estratégia	Comentários
01	0% ↓	72% ↑	87% ↑	98% ↑	97% ↑	16	Foco em P_A	Oportunidade
02	87% ↑	86% ↑	75% ↑	65% ↑	88% ↑	37	Manda bem	Parabéns
03	0% ↓	93% ↑	91% ↑	12% ↓	94% ↑	12	Foco em P_A e P_D	Oportunidade
04	79% ↑	46% ▢	43% ▢	41% ▢	45% ▢	7	Foco geral	Oportunidade
05	0% ↓	0% ↓	0% ↓	0% ↓	0% ↓	10	Foco geral	Oportunidade
06	0% ↓	98% ↑	13% ↓	8% ↓	86% ↑	18	Foco P_A, P_C e P_D	Oportunidade



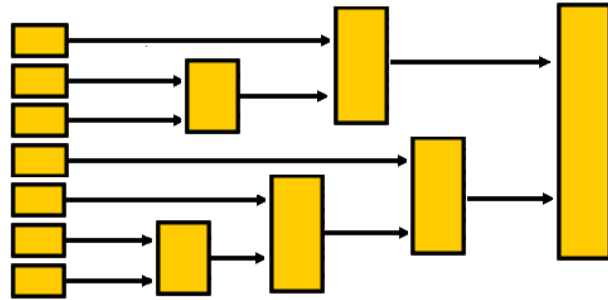
Aprecie



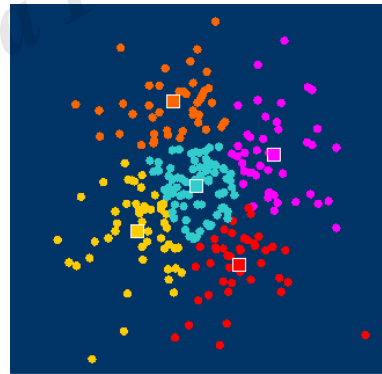
Análise de Cluster

Métodos de Agrupamento

- Hierárquico



- Cluster Não Hierárquico



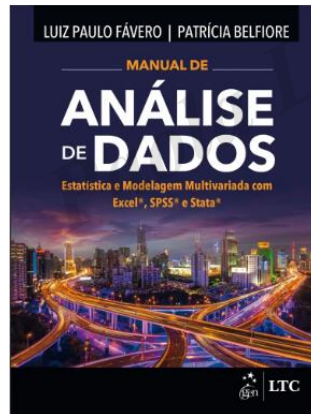
Análise de Cluster

Racional

- Definição do Problema
- Análise Exploratória da Base de Dados
- Padronização das Variáveis
- **Análise de Cluster**
- Caracterização dos grupos
- Aplicação de negócio

Referências

- Johnson, R. A. e Wichern, D. W. Applied Multivariate Statistical Analysis. Prentice-Hall Inc., 6th ed. 2007
- Timm, N.H. Applied Multivariate Analysis. Springer-Verlang, 2002
- Ferreira, D. F. Estatística Multivariada. 1.ed. —Lavras: Editora Ufla, 2008.



It's kind of fun to do the
IMPOSSIBLE

/in/adrianamms



Paola Lucca Pizutti 006.246.810-33

