Aula 17 - Algoritmos de Clusterização: K-means.

Profa. Gabrielly Queiroz

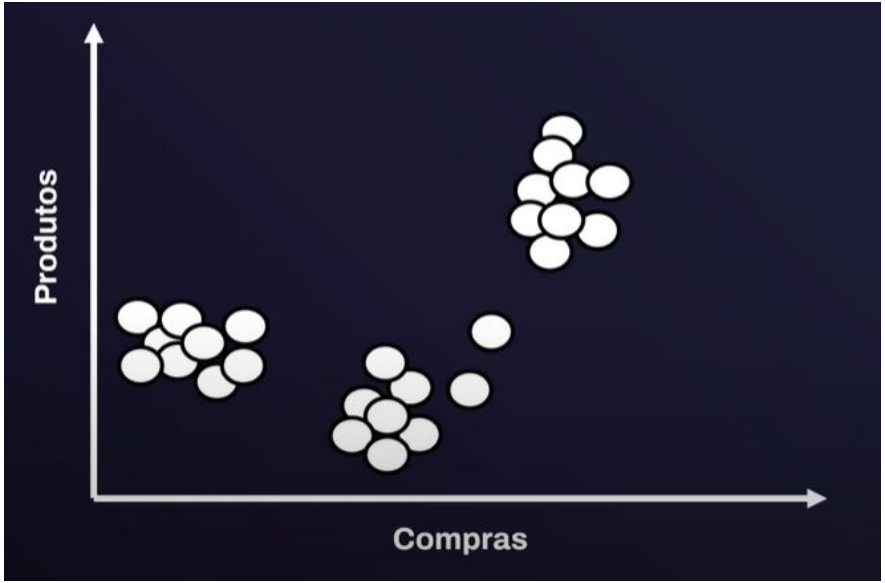
K-Means (clusterização).

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).

   k = 3



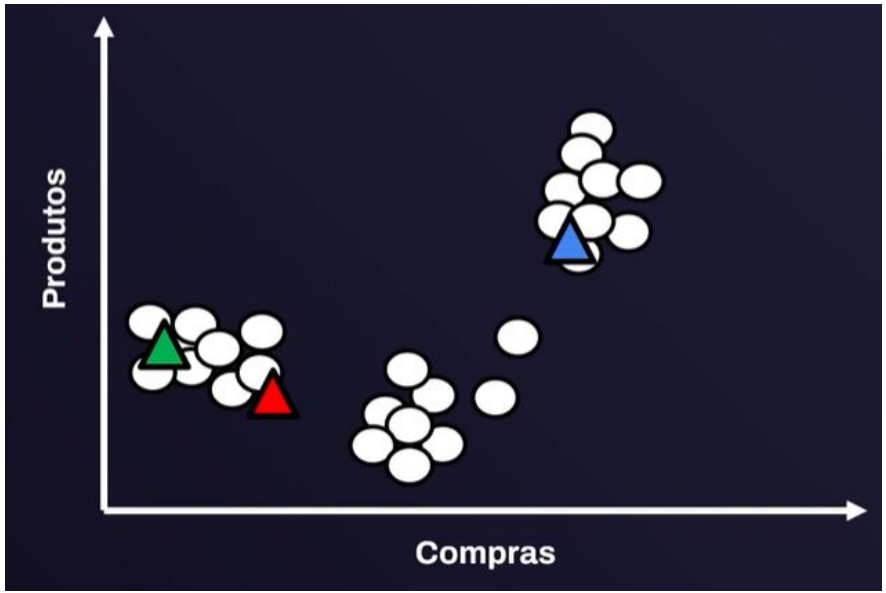
K-Means (clusterização).

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).
2. Inicializar os centróides aleatoriamente.
3. Calcular a Distância para todos os pontos para cada centróide.

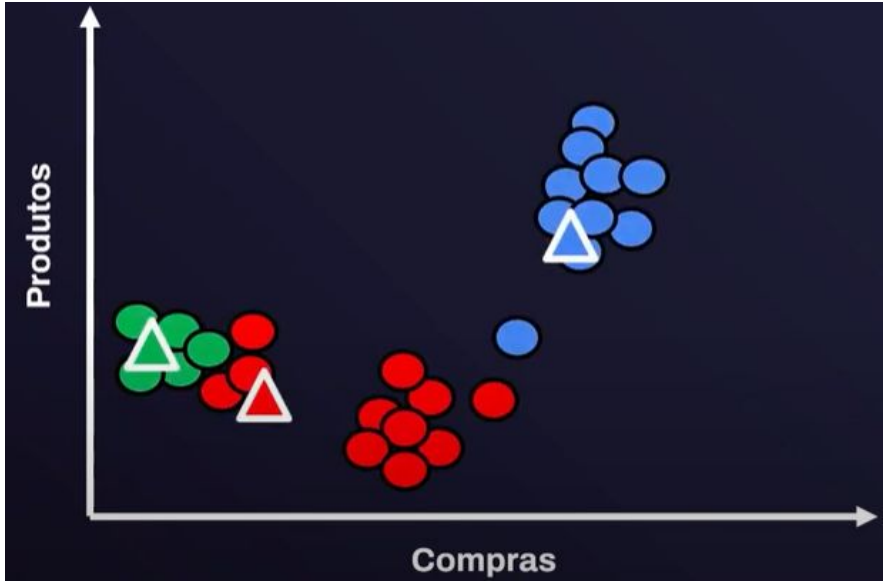


K-Means (clusterização).

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).
2. Inicializar os centróides aleatoriamente.
3. Calcular a Distância para todos os pontos para cada centróide.
4. Usar as distâncias mínimas para atribuir uma classe.

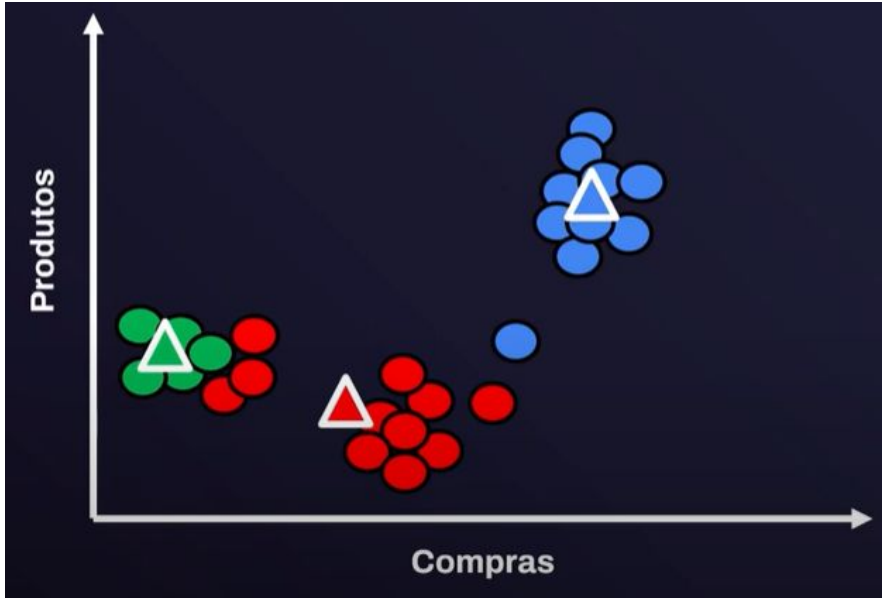


K-Means (clusterização).

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).
2. Inicializar os centróides aleatoriamente.
3. Calcular a Distância para todos os pontos para cada centróide.
4. Usar as distâncias mínimas para atribuir uma classe.
5. Recalcula o Centróide. Média das distâncias de cada grupo e coloca o centróide neste local.

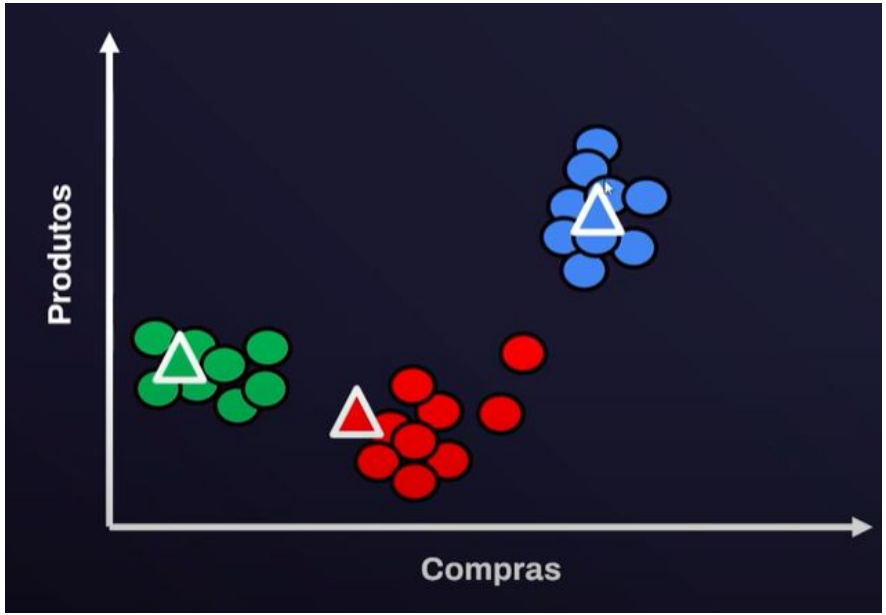


K-Means (clusterização).

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).
2. Inicializar os centróides aleatoriamente.
3. Calcular a Distância para todos os pontos para cada centróide.
4. Usar as distâncias mínimas para atribuir uma classe.
5. Recalcula o Centróide. Média das distâncias de cada grupo e coloca o centróide neste local.

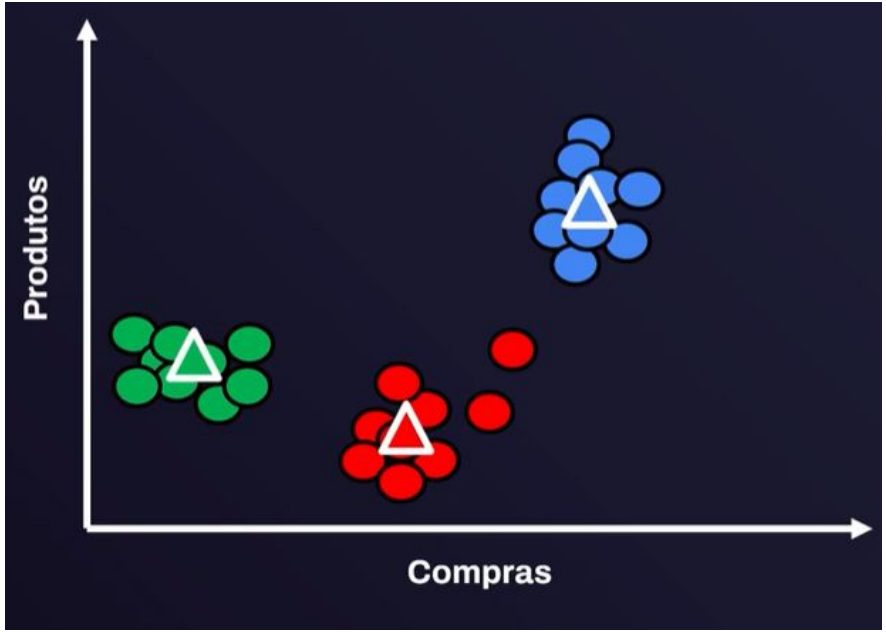


K-Means (clusterização).

- K-means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em k clusters. Ele funciona calculando os centróides dos grupos e associando cada ponto ao centróide mais próximo. O processo é repetido até que os centróides se estabilizem.
- **k**: é o número de clusters que você deseja formar. Esse valor é definido antes de executar o algoritmo.
- **Centróides**: são os "centros" de cada cluster, representando o ponto médio de todos os pontos do grupo.

Etapas principais:

1. Escolher o número de clusters (k).
2. Inicializar os centróides aleatoriamente.
3. Calcular a Distância para todos os pontos para cada centróide.
4. Usar as distâncias mínimas para atribuir uma classe.
5. Recalcula o Centróide. Média das distâncias de cada grupo e coloca o centróide neste local.
6. Recalcula novamente até não ter troca de cluster.



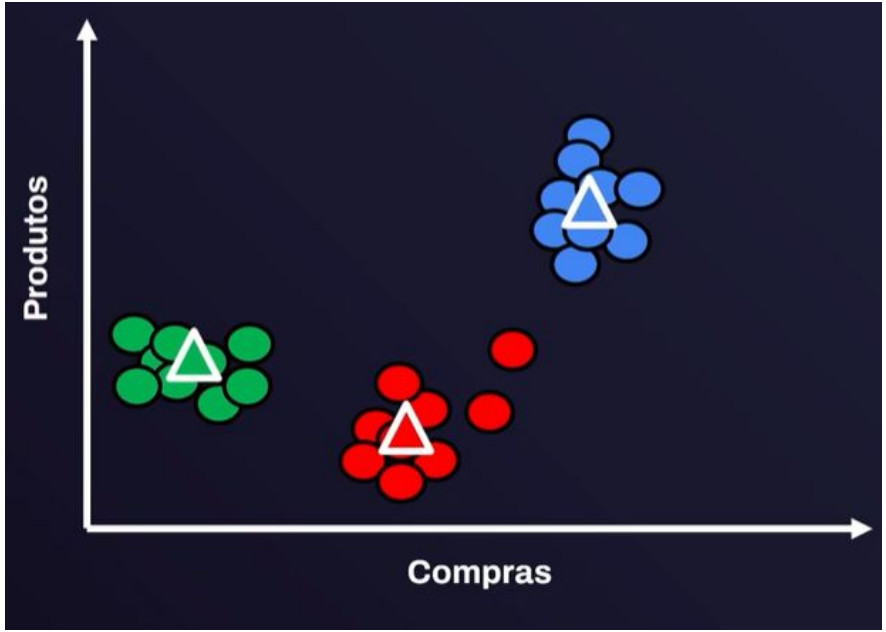
K-Means (clusterização).

Segmentação de cliente

Verde: clientes que compram poucos produtos e com pouca frequência.

Vermelho: clientes que compram poucos produtos e com alta frequência.

Azul: clientes que compram muitos produtos e com alta frequência.



K-Means - Resumo

- Declarar número de clusters K.
- Inicia aleatoriamente os centróides no espaço.
- Calcula a distância do centróide para todos os pontos.
- Atribui aos pontos a classe do centróide mais próximo.
- Reposiciona o centróide no centro médio dos clusters e refaz os cálculos.

- Simples de entender e aplicar.
- Computacionalmente eficiente, eficiente para quantidade de dados.

Exemplo

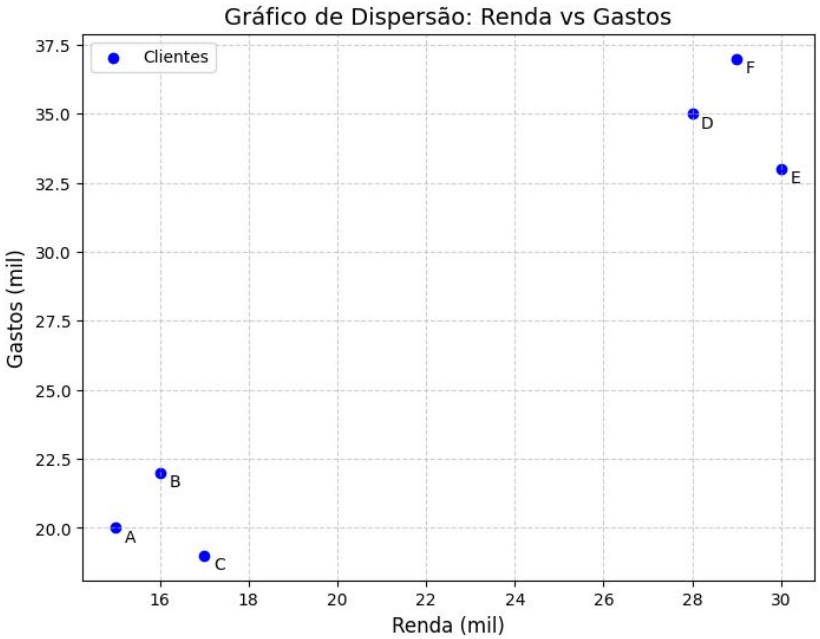
Agrupar clientes com base em renda e gastos.

Cliente	Renda (mil)	Gastos (mil)
A	15	20
B	16	22
C	17	19
D	28	35
E	30	33
F	29	37

Exemplo

- No K-means, **Renda e Gastos** são as dimensões (ou variáveis) que definem as características dos dados.
- O k representa o número de **grupos de observações semelhantes** que queremos formar a partir dessas dimensões.
- Visualizar os Dados (gráficos como dispersão auxiliam).
- Técnicas em algoritmos.

k = 2 (renda baixa e gasto baixo; renda alta e gasto alto).



Exemplo

Escolhemos k=2, o que significa que teremos **2 centróides iniciais**. Eles podem ser selecionados aleatoriamente ou com base em pontos do dataset.

Digamos que inicialmente escolhemos os centróides como:

- C1=(15,20)
- C2=(30,33)

Calculamos a distância até cada centróide. Exemplo para o ponto A = (15,20):

Ponto	Distância para $C1$	Distância para $C2$
A	0	19.85
B	2.24	17.80
C	2.24	19.13
D	19.85	2.24
E	20.81	0
F	22.02	2.24

Distância até $C1 = (15, 20)$:

$$d_{A,C1} = \sqrt{(15-15)^2 + (20-20)^2} = 0$$

Distância até $C2 = (30, 33)$:

$$d_{A,C2} = \sqrt{(15-30)^2 + (20-33)^2} = \sqrt{225 + 169} = \sqrt{394} \approx 19.85$$

Exemplo

Cada ponto é atribuído ao cluster cujo centróide está mais próximo:

- A,B,C: mais próximos de C1 (Cluster 1).
- D,E,F: mais próximos de C2 (Cluster 2).

Para cada cluster, recalculamos o centróide como a **média dos pontos do cluster**.

Para $C1$ (média de A, B, C):

$$C1 = \left(\frac{15+16+17}{3}, \frac{20+22+19}{3}\right) = (16, 20.33)$$

Para $C2$ (média de D, E, F):

$$C2 = \left(\frac{28+30+29}{3}, \frac{35+33+37}{3}\right) = (29, 35)$$

Com os novos centróides, repetimos os passos. O processo continua até que os centróides não mudem mais (convergência).

Exemplo

Calculamos novamente as distâncias de cada ponto aos novos centróides.

Cliente	$d(C1)$	$d(C2)$	Cluster
A	1.05	18.44	1
B	1.67	17.11	1
C	1.36	18.72	1
D	17.44	1.00	2
E	19.44	0.57	2
F	21.56	1.41	2

Nenhuma mudança nos clusters!

- Como os centróides não mudaram na segunda iteração, o algoritmo **convergiu**. Isso significa que os clusters finais foram encontrados.

Atividade

Agrupar cidades com base em temperatura média e umidade relativa.

Cidade	Temperatura (°C)	Umidade (%)
A	22.0	70
B	25.5	65
C	18.0	80
D	27.0	60