



Aprenda técnicas de Análise de Dados, Machine Learning e muito mais através de aulas 100% práticas partindo do nível zero até o avançado.

## Quero aprender



# Data Science do ZERO

Capítulo 06 - Machine Learning  
**Agrupamento de Dados.**

# O que é Agrupamento de Dados

- O Agrupamento de dados é uma técnica de Machine Learning que consiste segmentar itens que possuem algum tipo de similaridade.
- Em nosso dia a dia fazemos diversos agrupamentos através da identificação de padrões levando em consideração vários atributos como cor, forma, tamanho, peso etc.
- Tudo isso de forma rápida e intuitiva.



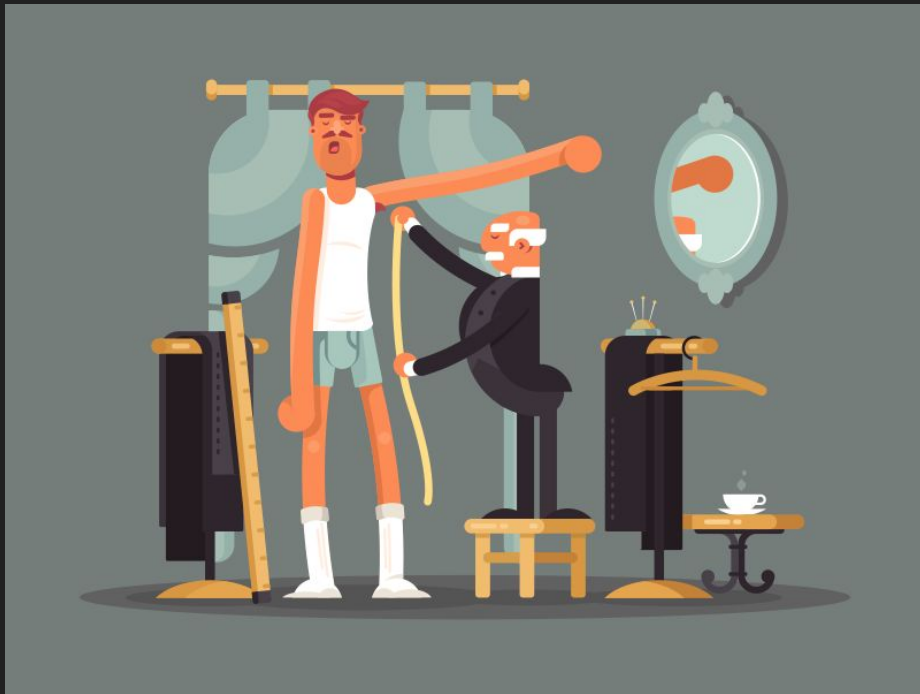
# Conceitos sobre Agrupamento de Dados

- Conseguimos agrupar coisas e objetos através de características em comum, ou seja, propriedades entre os objetos que sejam parecidas.
- A similaridade é o ponto chave para estabelecer os grupos.

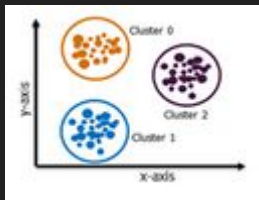


# Conceitos sobre Agrupamento de Dados

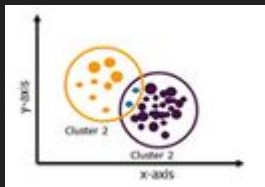
- No agrupamento de dados precisamos medir a similaridade entre as instâncias de dados.
- É através dessa distância entre cada ponto que iremos estabelecer os grupos.



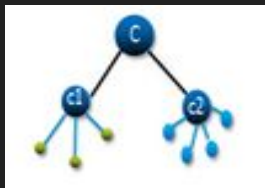
# Tipos de Grupos



**Exclusive Cluster** ou **Cluster Exclusivo** se refere a um tipo de agrupamento onde os registros são exclusivos ,ou seja, cada registro pertence a um único grupo.



**Overlapping Cluster** ou **Cluster Sobreposto** se refere a um tipo de agrupamento onde os registros podem pertencer a mais de um grupo ou cluster, diferente do Exclusive Cluster.

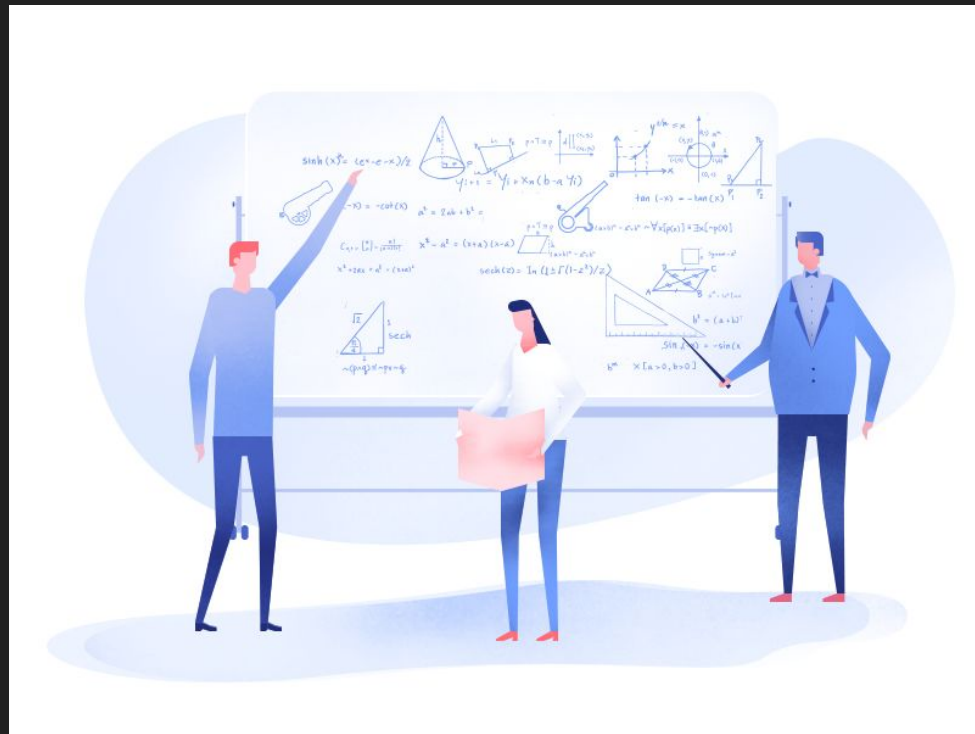


**Hierarchical Cluster** ou **Cluster hierárquico** se refere a um tipo de agrupamento onde possui uma hierarquia entre os grupos. Os registros podem ser agrupados em grupos que podem conter subgrupos contendo outros registros.

# Algoritmo K-Means

**Algoritmo do tipo não supervisionado** que tem como objetivo encontrar similaridades entre os dados e agrupá-los conforme o número de cluster passado pelo argumento k.

A similaridade entre cada ponto é calculada através de uma função de distância.



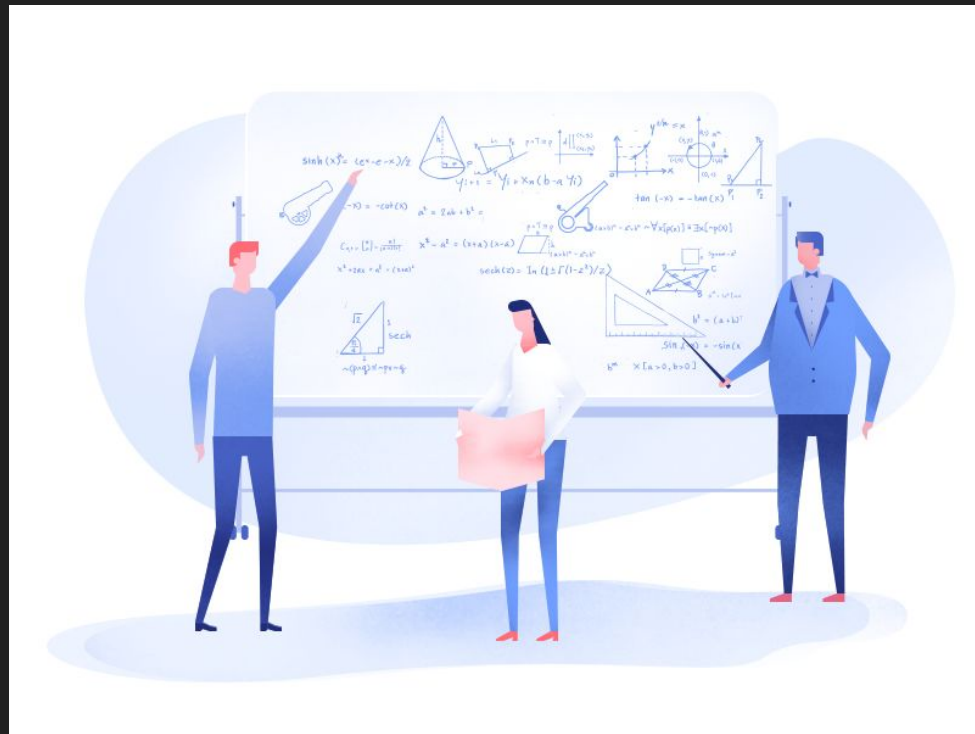
# Algoritmo K-Means

O funcionamento do algoritmo é dividido em quatro fases que são.

**Inicialização:** geração de forma aleatória ou (através de outro método) de  $k$  centroids, onde o número de centroids é representado ao parâmetro  $k$ .

Pontos de dados que serão utilizados, como o nome sugere, de pontos centrais dos clusters.

Referências que serão utilizadas para calcular a distância entre os dados e gerar os clusters.



# Algoritmo K-Means

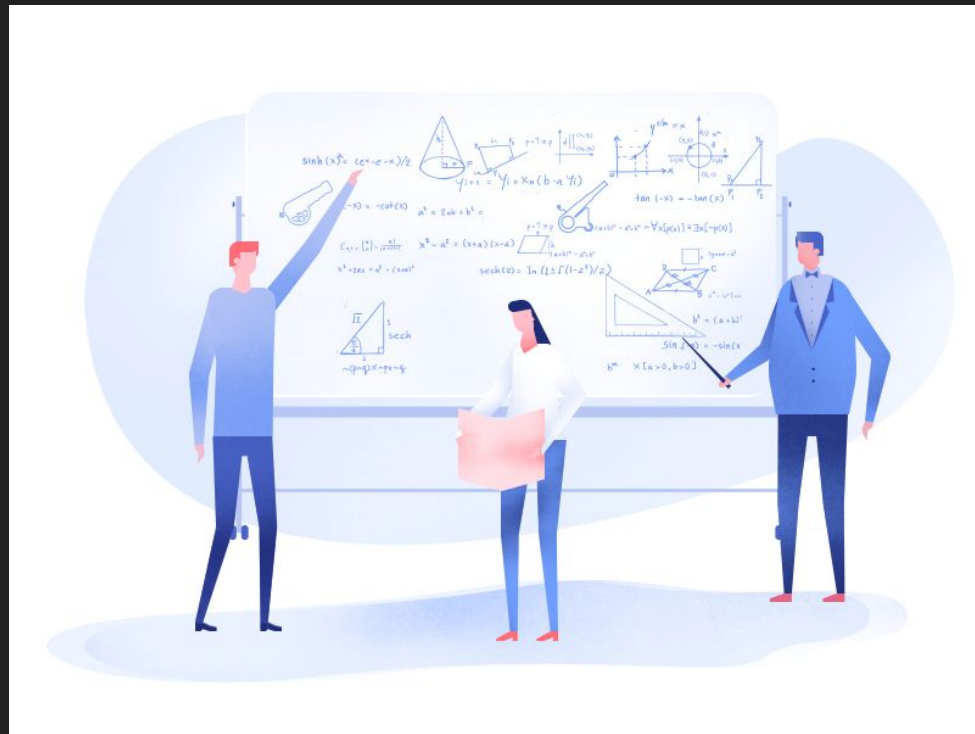
## Atribuição ao Cluster

Calculado a distância entre todos os pontos de dados e cada um dos centroids.

Atribuído ao centroid ou cluster que tem a **menor distância**.

Função de distância como **Euclidiana**.

Etapa finalizada com os dados entre cada cluster/centroid.



# Algoritmo K-Means

## Movimentação de Centroids

Recalcula o valor dos centróides através da média dos valores dos pontos de dados.

Novo valor de centroíde.



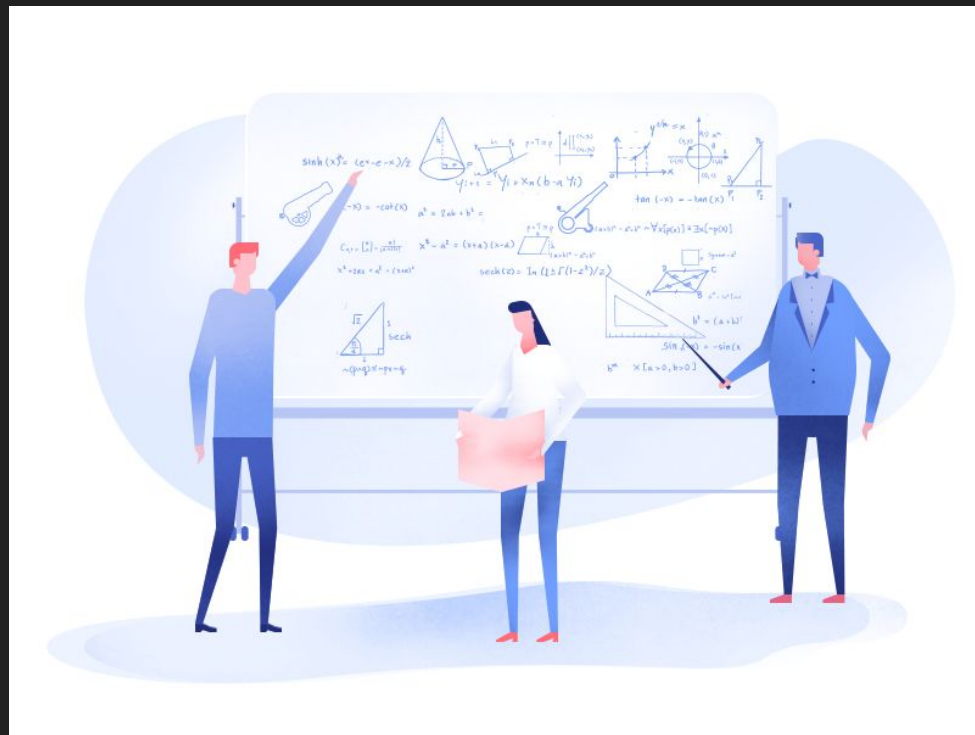
# Algoritmo K-Means

## Otimização do K-médias

Na fase final da execução do K-means as fases **Atribuição ao Cluster** e **Movimentação de Centroids** são repetidas até o cluster se tornar estático ou algum critério de parada tenha sido atingido.

O cluster se torna estático quando nenhum dos pontos de dados alteram de cluster.

Critério de parada como número de iterações.



Hands on!