Resumo Jonathan Alves

Kmeans -> Algoritmo Teoria

O **K-means** é um **algoritmo de agrupamento** que separa dados em **K grupos** (ou clusters). Ele busca colocar itens parecidos no mesmo grupo, com base em características que você definir.

Quando Usar?

- Quando você quer **encontrar padrões ocultos** ou organizar dados em grupos.
- Quando você **não tem rótulos** definidos (como A, B, C) e quer que o algoritmo ***descubra os grupos automaticamente.***
- ***Exemplo: Agrupar clientes por comportamento ou funcionários por desempenho.***

Por que Ele é Bom?

- **Simples e rápido** para conjuntos de dados grandes.
- **Identifica quando um grupo começa e o Outro termina**
- **Útil para explorar dados** e entender padrões.
- Ajuda a **organizar e segmentar** informações sem precisar rotular manualmente.

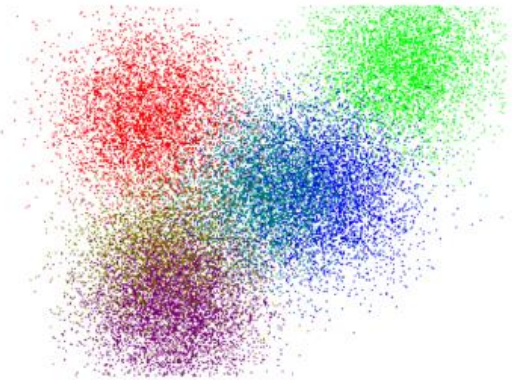
Por fim, ele é ótimo quando você **não sabe exatamente como dividir seus dados** e precisa de uma forma automática e inteligente para encontrar essas divisões.

Algoritmo de Loyd

➔ O algoritmo de Lloyd é usado no K-means para formar grupos.

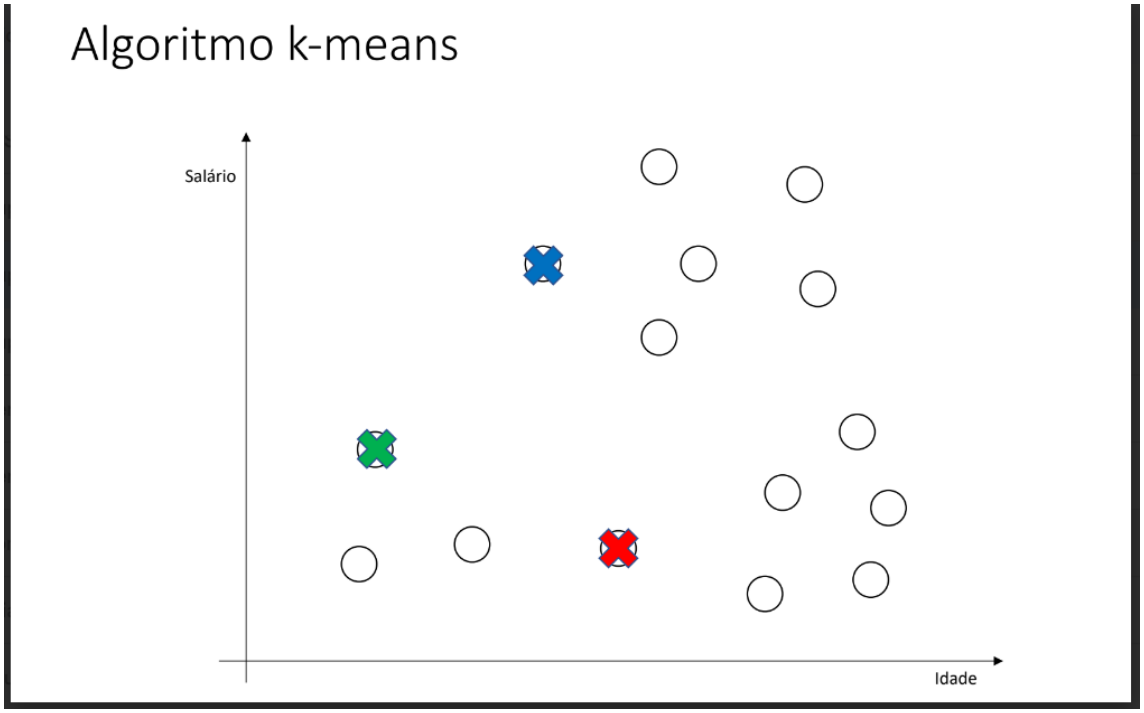
➔ Primeiro, ele escolhe centros aleatórios. Depois, cada dado é colocado no grupo com o centro mais próximo.

➔ Em seguida, o centro de cada grupo é atualizado com a média dos pontos. Esse processo se repete até os centros pararem de mudar. No final, os dados ficam organizados nos grupos mais próximos possíveis.

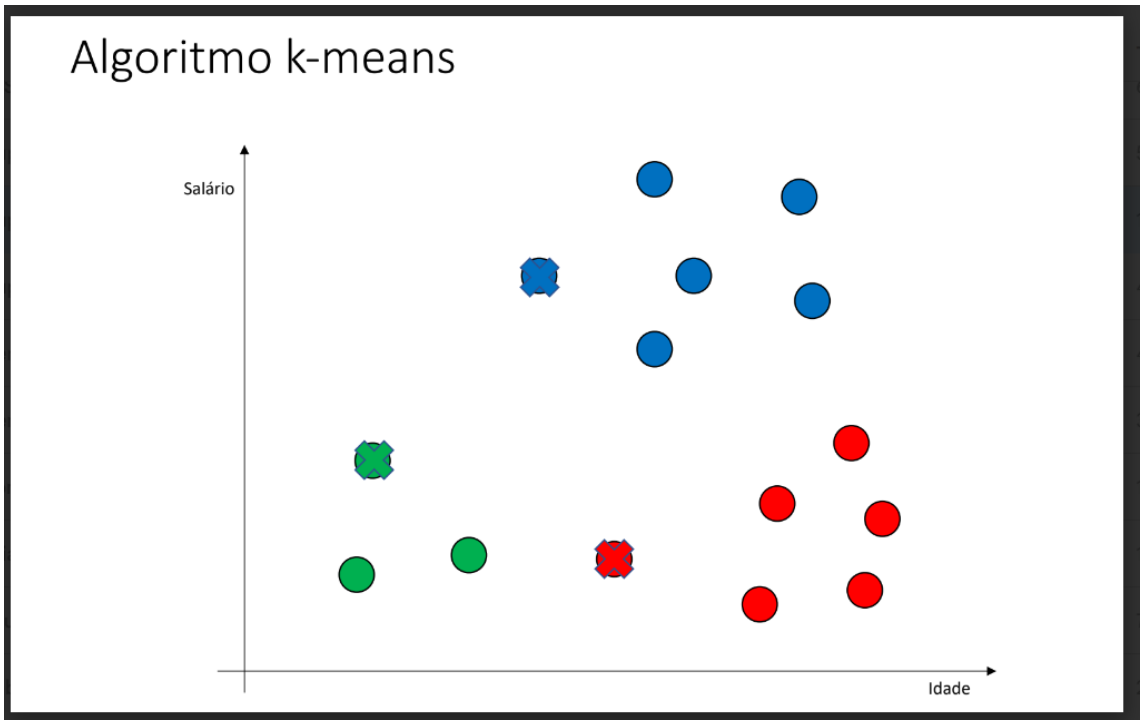


O Algoritmo nesse primeiro momento encontra os grupos, e ao encontrar os grupo, ele irá buscar os mais próximos para fazer a classificação exemplo:

***Exemplo Antes*:**



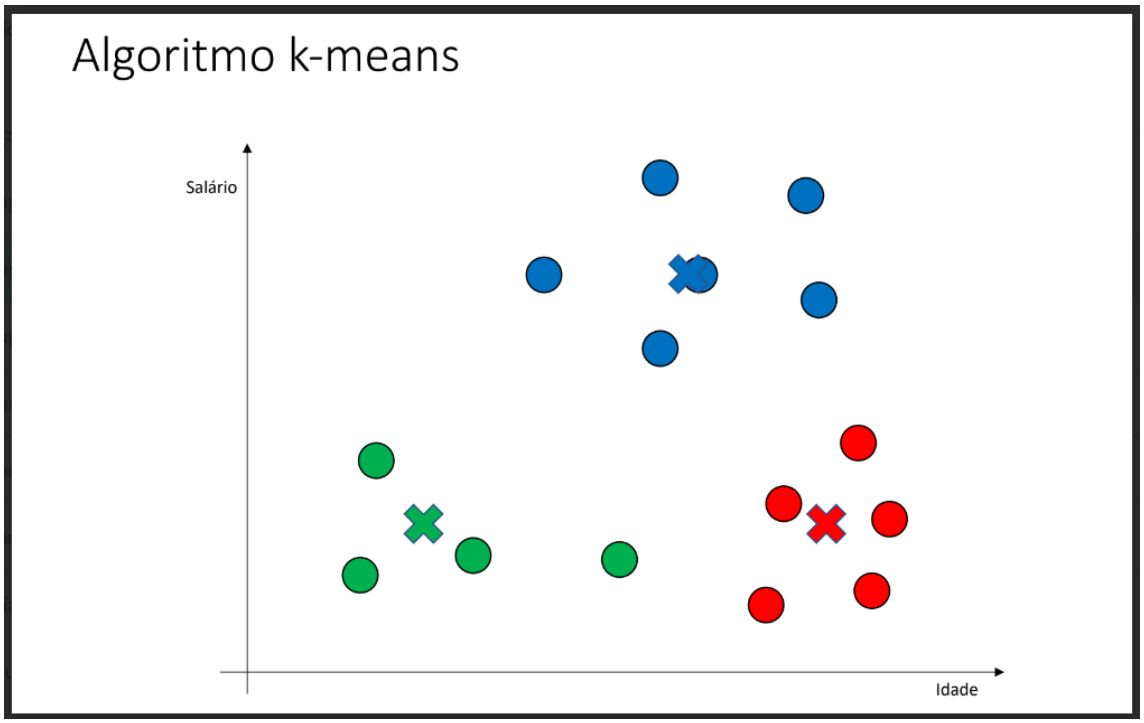
Depois:



Após fazer as classificações dos grupos, a ideia é achar o centro da classificação, e classifica novamente os registros pelos mais próximos.

Obs: Os pontos marcados com" X" na imagem abaixo, em tese seriam o centro buscado da classificação

Resultado:



Cálculo do K-means

O K-means usa a distância Euclidiana para calcular a proximidade entre os pontos e os centros dos grupos. A fórmula é:

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2}$$

Exemplo de como calcular

Base de dados X: [5, 7, 9]

Base de dados C ou Y: [5, 5, 5]

Primeiro Passo: Subtraia cada posição do vetor:

5 – 5 = 0

7- 5 = 2

9- 5 = 4

Segundo Passo: Eleva os valores ao Quadrado

$0^2 = 0$

$2^2 = 4$

$4^2 = 16$

Terceiro Passo: Somatório

0 + 4 + 16 = 20

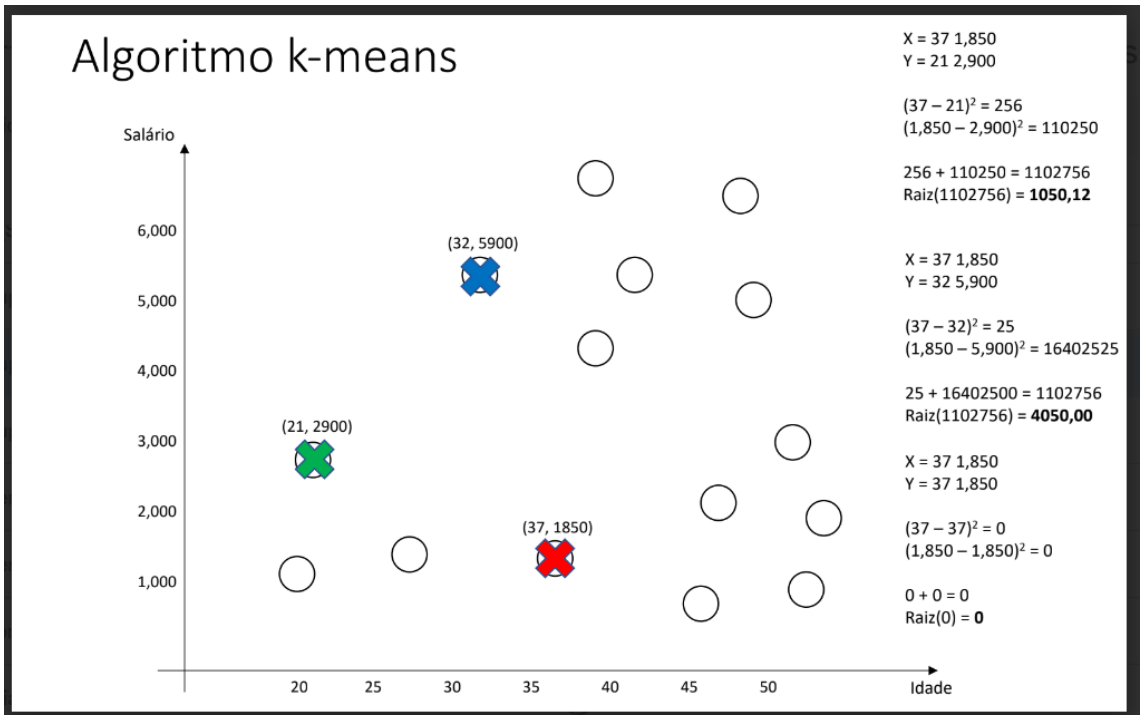
Último Passo: Raiz Quadrada

Raiz quadrada de 20 = 4,47

Cálculo Reais

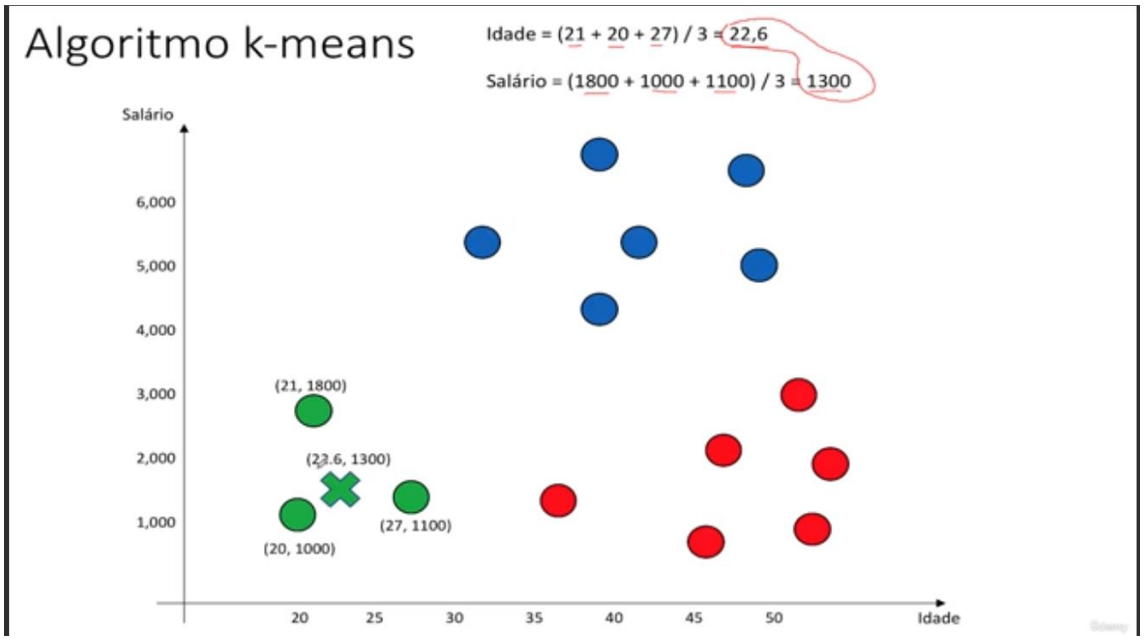
Aplicando os cálculos sobre uma base de dados real.

Tentando encontrar o grupo mais próximo do registro (37 anos, 1850 Salário), o mais próximo acaba sendo o grupo vermelho, ou seja, ele mesmo nesse exemplo:

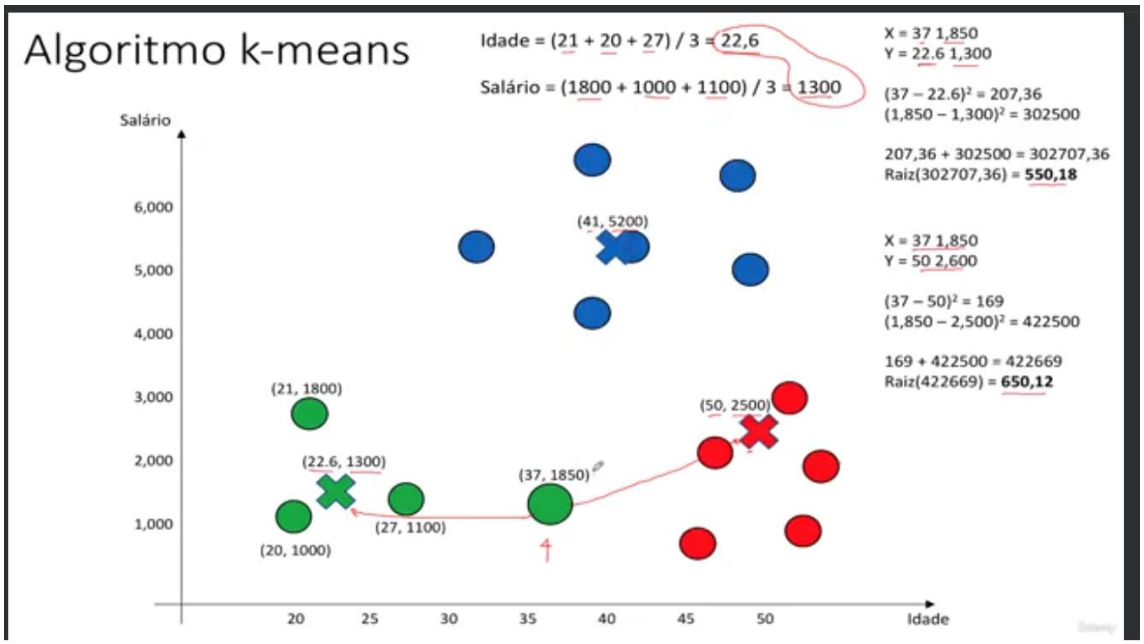


Atualizando os Centroides (centro da classificação):

Para calcular o ponto onde o centroide ficara é simples. Tiramos as médias dos valores, e depois definimos no eixo onde está



Agora recalculando os outros pontos para validar, notamos que esse ponto destacado, anteriormente classificado como vermelho, tem a menor distancia se comparado com os verdes do que com o vermelho, logo ele será classificado como verde agora



Obs: Sempre lembre que a classificação se dá pela menor distancia com o centroide.

K-MEANS ++

O K-means++ é uma melhoria do K-means que resolve um problema comum: a escolha dos centros iniciais. No K-means tradicional, os centros são escolhidos aleatoriamente, o que pode levar a resultados ruins se os pontos iniciais forem mal distribuídos. O K-means++ evita isso escolhendo centros de forma mais inteligente.

Como funciona o K-means++?

1. Escolhe o primeiro centro aleatoriamente entre os dados.

2. Para o próximo centro, dá preferência aos pontos mais distantes do centro já escolhido.
3. Repete o processo até escolher todos os K centros, garantindo que eles estejam bem espalhados.
4. Depois, o K-means segue normalmente: agrupa dados, recalcula centros e repete até estabilizar.

Por que é melhor?

- Os centros iniciais são mais bem distribuídos.
- O algoritmo converge mais rápido e encontra grupos melhores.
- Reduz a chance de resultados ruins causados por uma escolha aleatória ruim.

Quando usamos o K-means, precisamos definir sempre o número de Cluster (grupos) que queremos ter como saída. Exemplo em uma base de Clientes Bancarios. Precisamos definir os Bons e Mals Pagadores.

Logo, a saída é igual a 2 grupos. Mas caso não tenhamos o número de grupos que queremos ter em mente, podemos calcular as possíveis saídas com essa fórmula aqui: