Ciência de dados = transformar um dado em conhecimento

Inteligência artificial = Rede neural que tenta simular a inteligência humana

Subcampos da IA:

- Machine Learning: Aprende com dados, sem explicitamente programar cada tarefa

- Deep Learning: Machine Learning para padrões complexos

- Processamento de Linguagem Natural: chat gpt

- Visão computacional: Interpreta imagens e vídeos

Vamos usar para nosso sistema de recomendação:

- Mineração de dados

- IA

- Big data: Tratamento de dados em grandes volumes em grande velocidade e variedade

Os 5Vs do Big data:

* Volume: quantidade de dados gerados
* Velocidade: Rapidez na geração e processamento
* Variedade: Diferentes tipos e formatos de dados
* Veracidade: Confiabilidade e qualidade dos dados
* Valor: Utilidade dos dados para gerar benefícios

Clusterização:

- Método de aprendizado não supervisionado

- Se baseia em centroide: joga informação para o grupo mais próximo (define pela distância euclidiana) e define o centro de forma que fique o menos longe possível de todos, depois fico refazendo as contas até o centroide se estabilize (pare de alterar tanto o valor)

\*Distância euclidiana = ponto 1 – ponto 2

- Quanto mais cluster, mais longe os valores ficam do centroide

Para escolher a quantidade de clusters:

1 - Método do Cotovelo (Elbow Method):

* Plota a inércia (soma das distâncias dos pontos aos centroides) para diferentes valores de k.
* Procura o ponto onde a redução da inércia começa a diminuir de forma menos acentuada (o “cotovelo”).

2 - Coeficiente de Silhueta:

* Mede coesão e separação dos clusters.
* Valores próximos de 1 indicam clusters bem separados e compactos, próximo do 0 ótimo e se negativo significa que não separou direito

- Inercia = soma da distância do ponto até o centroide