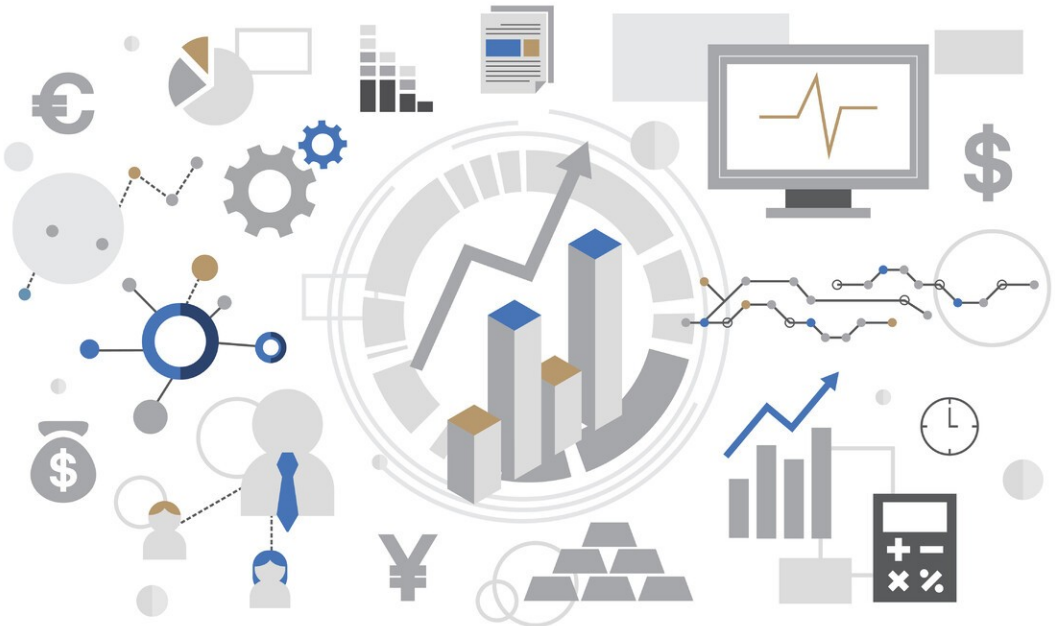
Ciência de Dados



Prof. Me. Clênio Silva
e-mail: clenio.silva@uniube.br

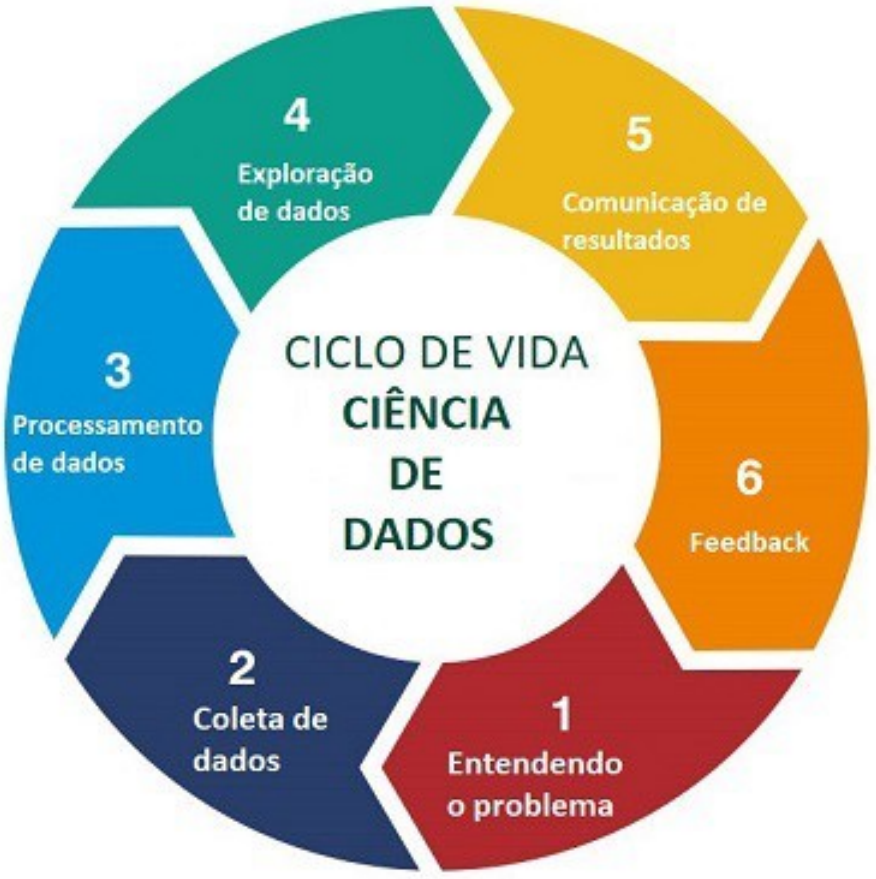
O que é ciência de dados

- Definição:
  - Ciência de Dados é um campo interdisciplinar que utiliza técnicas estatísticas, algoritmos de aprendizado de máquina e análise para extrair conhecimento e insights a partir de dados.
- Objetivo:
  - Transformar dados brutos em informações úteis para tomada de decisões

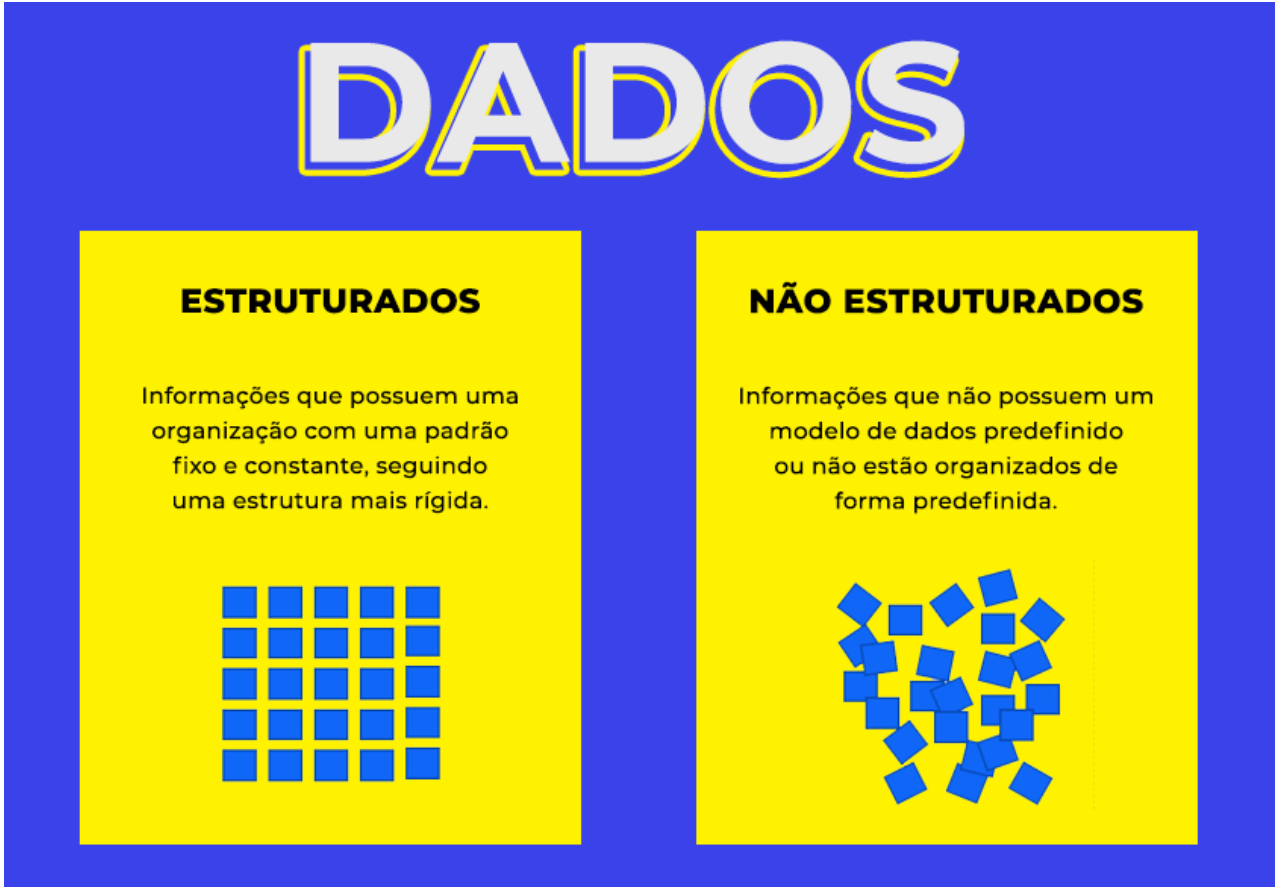
Componentes da Ciência de Dados

- Coleta de Dados:
  - Obtenção de dados de diversas fontes, como bancos de dados, APIs, e arquivos.
- Limpeza e Preparação:
  - Processamento dos dados para remover inconsistências e preparar para análise.
- Análise e Modelagem:
  - Aplicação de técnicas estatísticas e algoritmos para identificar padrões e construir modelos.
- Visualização e Comunicação:
  - Apresentação dos resultados de forma compreensível para stakeholders.

Componentes da Ciência de Dados



Coleta de dados



Coleta de dados

- Identificar e coletar dados de diferentes fontes, como bancos de dados, arquivos CSV, APIs, etc.
- Verificar se os dados estão em um formato utilizável.

Limpeza de dados

- Remoção de Duplicatas: Identificar e remover registros duplicados.
- Tratamento de Valores Faltantes: Lidar com valores ausentes usando técnicas como imputação (substituição por média, mediana, valor mais frequente) ou exclusão de registros.
- Correção de Erros: Corrigir erros de digitação, valores fora do esperado, e inconsistências.
- Formatação: Ajustar a formatação dos dados para que estejam consistentes (por exemplo, datas no mesmo formato, números com a mesma precisão).

Análise de dados

- Análise Exploratória de Dados (EDA)
  - Visualizações:
    - Utilizar gráficos como histogramas
  - Estatísticas Descritivas:
    - Calcular métricas como média, mediana, desvio padrão, e percentis para entender a distribuição e características dos dados.
  - Correlação e Relações:
    - Examinar relações entre variáveis utilizando correlações e testes de dependência.

Análise de dados

- Análise Inferencial
  - Testes Estatísticos:
    - Realizar testes como t-tests, ANOVA, e qui-quadrado para fazer inferências sobre populações com base em amostras.
  - Intervalos de Confiança:
    - Calcular intervalos de confiança para estimar a precisão de parâmetros populacionais.

Análise de dados

- Segmentação de Dados
  - Clustering:
    - Usar algoritmos de agrupamento como K-means, DBSCAN ou Hierarchical Clustering para identificar grupos semelhantes dentro dos dados.
  - Segmentação:
    - Dividir os dados em segmentos com base em características comuns para análises mais detalhadas.

Modelagem de dados

- Modelos Descritivos:
  - Modelos que ajudam a entender os dados e explicar padrões. Exemplos incluem regressão linear e análise de componentes principais (PCA).
- Modelos Preditivos:
  - Modelos que fazem previsões sobre dados futuros com base em dados históricos. Exemplos incluem regressão logística, árvores de decisão, redes neurais e máquinas de vetores de suporte (SVM).

Preparação de dados para modelagem

- Divisão de Dados:
  - Separar os dados em conjuntos de treinamento, validação e teste para avaliar a performance do modelo de forma justa.
- Feature Engineering:
  - Criar novas variáveis (features) que possam melhorar a capacidade preditiva do modelo. Exemplos incluem criar variáveis a partir de data/hora, combinar features existentes ou extrair características de texto.

Exemplo prático

```python
# Importando bibliotecas
import pandas as pd
import matplotlib.pyplot as plt

# Criando um DataFrame simples
data = {
    'Ano': [2020, 2021, 2022],
    'Vendas': [1500, 1800, 2100]
}
df = pd.DataFrame(data)

# Exibindo o DataFrame
print(df)

# Criando um gráfico simples
plt.plot(df['Ano'], df['Vendas'], marker='o')
plt.title('Vendas Anuais')
plt.xlabel('Ano')
plt.ylabel('Vendas')
plt.grid(True)
plt.show()
```