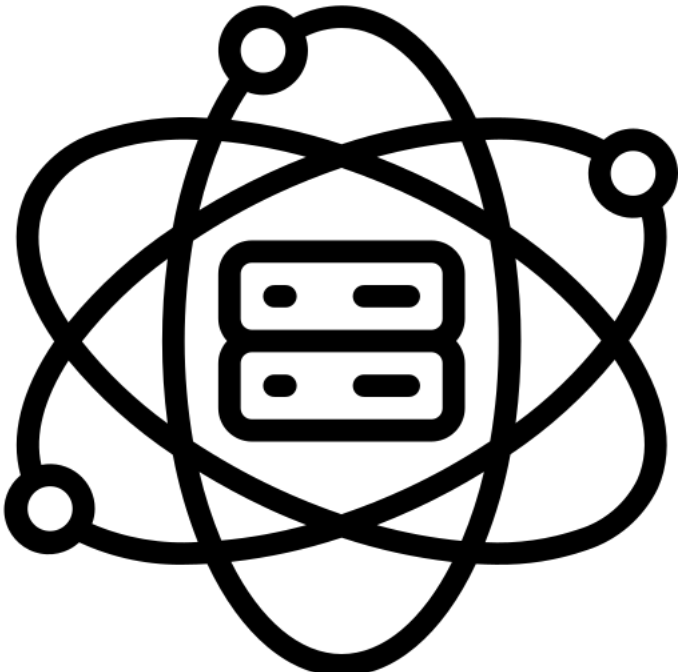
Paws on Data: Um Guia de Ciência de Dados para iniciantes

Christian Matheus

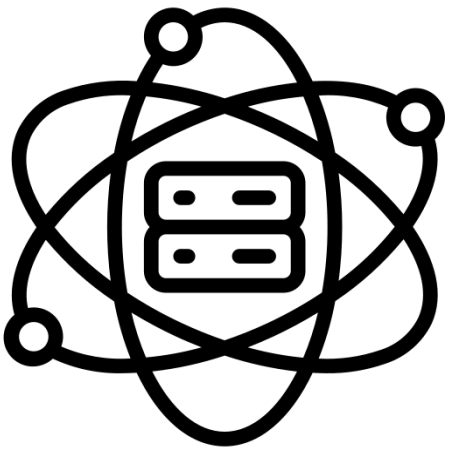
O que é Ciência de dados

A ciência de dados é a prática de usar dados para extrair informações valiosas e tomar decisões informadas. Com o crescente volume de dados gerados diariamente, a capacidade de analisar e interpretar esses dados se tornou essencial em diversas áreas.



Princípios básicos da Ciência de dados

1. Coleta de Dados: Reunir dados de várias fontes.
2. Limpeza de Dados: Remover inconsistências e dados faltantes.
3. Análise Exploratória de Dados (EDA): Compreender os padrões nos dados.
4. Modelagem de Dados: Criar modelos estatísticos e de aprendizado de máquina.
5. Interpretação e Comunicação de Resultados: Transformar resultados técnicos em insights acionáveis.



I

Coleta de Dados

A coleta de dados é a primeira etapa no processo de ciência de dados. Envolve reunir dados de várias fontes, que podem incluir bancos de dados, arquivos CSV, APIs e web scraping.

Coleta de Dados

- Bancos de Dados Relacionais: MySQL, PostgreSQL.
- Arquivos CSV: Planilhas e outros arquivos de texto.
- APIs: Interfaces de programação de aplicativos que fornecem dados.
- Web Scraping: Extração de dados de sites.
- Veja um exemplo genérico, de dados coletados de uma API, no código abaixo:

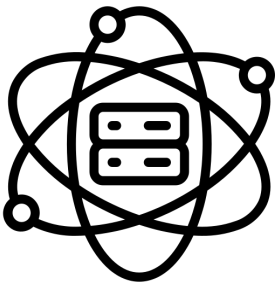
Coleta_dado.py

```python
import requests
import pandas as pd

# Coletando dados de uma API
url = "https://api.exemplo.com/dados"
response = requests.get(url)
data = response.json()

# Convertendo os dados para um DataFrame do Pandas
df = pd.DataFrame(data)

# Exibindo as primeiras linhas do DataFrame
print(df.head())
```



II

Limpeza e Preparação de Dados

A limpeza e preparação de dados são etapas cruciais no processo de ciência de dados. Dados brutos geralmente contêm erros, valores ausentes e inconsistências que podem afetar a análise.

Limpeza e Preparação de Dados

- Tratamento de Valores Ausentes: Substituir ou remover valores nulos.
- Remoção de Duplicatas: Eliminar registros duplicados.
- Correção de Tipos de Dados: Assegurar que os dados estejam no formato correto.
- Veja um exemplo genérico, de remoção de dados duplicados, no código abaixo:

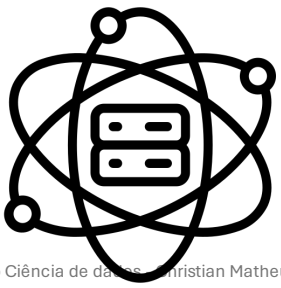
Limpeza_dados.py

```python
import pandas as pd

# Removendo valores ausentes
dados = dados.dropna()

# Removendo duplicatas
dados = dados.drop_duplicates()

# Convertendo coluna para o tipo correto
dados['data'] = pd.to_datetime(dados['data'])
```



III

Análise Exploratória de Dados (EDA)

A análise exploratória de dados (EDA) é o processo de examinar conjuntos de dados para resumir suas principais características, frequentemente com a ajuda de visualizações gráficas.

Análise Exploratória de Dados (EDA)

- Resumo Estatístico: Calcular médias, medianas, etc.
- Visualizações: Usar gráficos para entender padrões e tendências.
- Detecção de Outliers: Identificar valores atípicos que podem distorcer a análise.
- Veja um exemplo genérico, de criação de um histograma, no código abaixo:

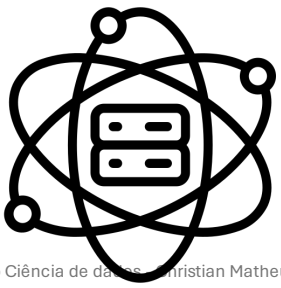
Analise_exploratoria_dados.py

```python
import matplotlib.pyplot as plt

# Resumo estatístico
print(dados.describe())

# Histograma de uma coluna
dados['coluna_interesse'].hist()
plt.show()

# Gráfico de dispersão
dados.plot.scatter(x='coluna_x', y='coluna_y')
plt.show()
```



IV

Modelagem de Dados

A modelagem de dados é uma etapa fundamental na ciência de dados. Envolve a criação de modelos matemáticos ou algoritmos de aprendizado de máquina que podem fazer previsões ou identificar padrões nos dados.

Modelagem de Dados

- Escolha do Modelo: Selecionar um modelo apropriado para o problema.
- Treinamento do Modelo: Ajustar o modelo aos dados de treinamento.
- Avaliação do Modelo: Avaliar a precisão do modelo usando dados de teste.
- Veja um exemplo genérico, de um modelo de Regressão linear, no código abaixo:

Regressao_linear.py

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# Carregando os dados
dados = pd.read_csv('dados.csv')

# Separando as variáveis dependentes e independentes
X = dados[['variavel_independente']]
y = dados['variavel_dependente']

# Dividindo os dados em conjuntos de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Criando o modelo de regressão linear
modelo = LinearRegression()

# Treinando o modelo
modelo.fit(X_train, y_train)

# Fazendo previsões
previsoes = modelo.predict(X_test)

# Avaliando o modelo
erro_medio_quadrado = mean_squared_error(y_test, previsoes)
print(f'Erro Médio Quadrado: {erro_medio_quadrado}')
```

V

Interpretação e Comunicação de Resultados

Interpretação e comunicação de resultados são etapas críticas na ciência de dados. Os insights derivados dos modelos precisam ser traduzidos em termos compreensíveis para a tomada de decisão.

Interpretação e Comunicação de Resultados

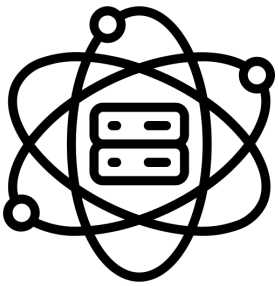
- Interpretação dos Resultados: Compreender o que os resultados significam no contexto do problema.
- Visualização dos Resultados: Usar gráficos e tabelas para ilustrar os resultados.
- Comunicação Eficaz: Apresentar os resultados de forma clara e concisa para as partes interessadas.
- Veja um exemplo genérico, de um modelo de gráfico, no código abaixo:

Grafico.py

```python
import matplotlib.pyplot as plt

# Visualizando as previsões vs. valores reais
plt.scatter(y_test, previsoes)
plt.xlabel('Valores Reais')
plt.ylabel('Previsões')
plt.title('Previsões vs. Valores Reais')
plt.show()

# Plotando os resíduos
residuos = y_test - previsoes
plt.hist(residuos)
plt.xlabel('Resíduos')
plt.ylabel('Frequência')
plt.title('Distribuição dos Resíduos')
plt.show()
```



AGRADECIMENTOS

OBRIGADO POR LER!

Esse ebook foi criado por IA, e diagramado por mim. O passo a passo se encontra no meu github.

Conteúdo criado para fins didáticos.

T

https://github.com/Christian-Matheus/Ebook_com_chatGPT