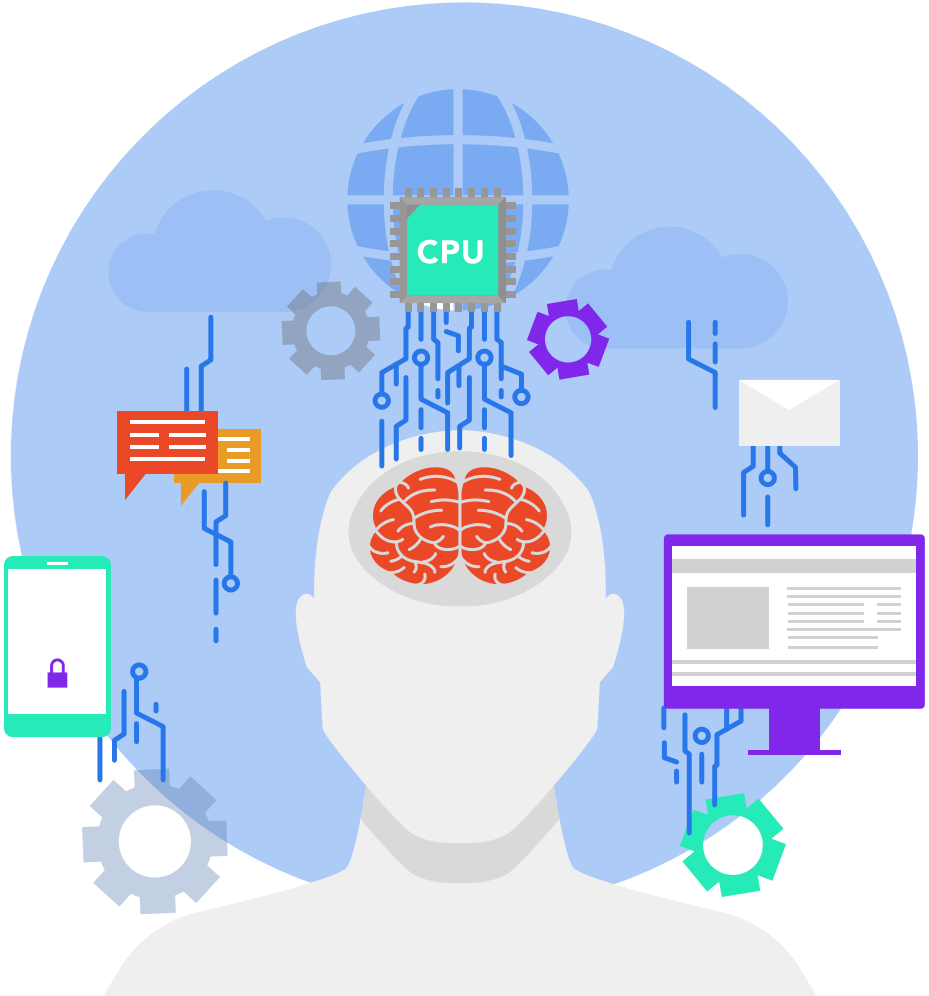


Pipeline de ML

Professor: Gabriel Oliveira Assunção

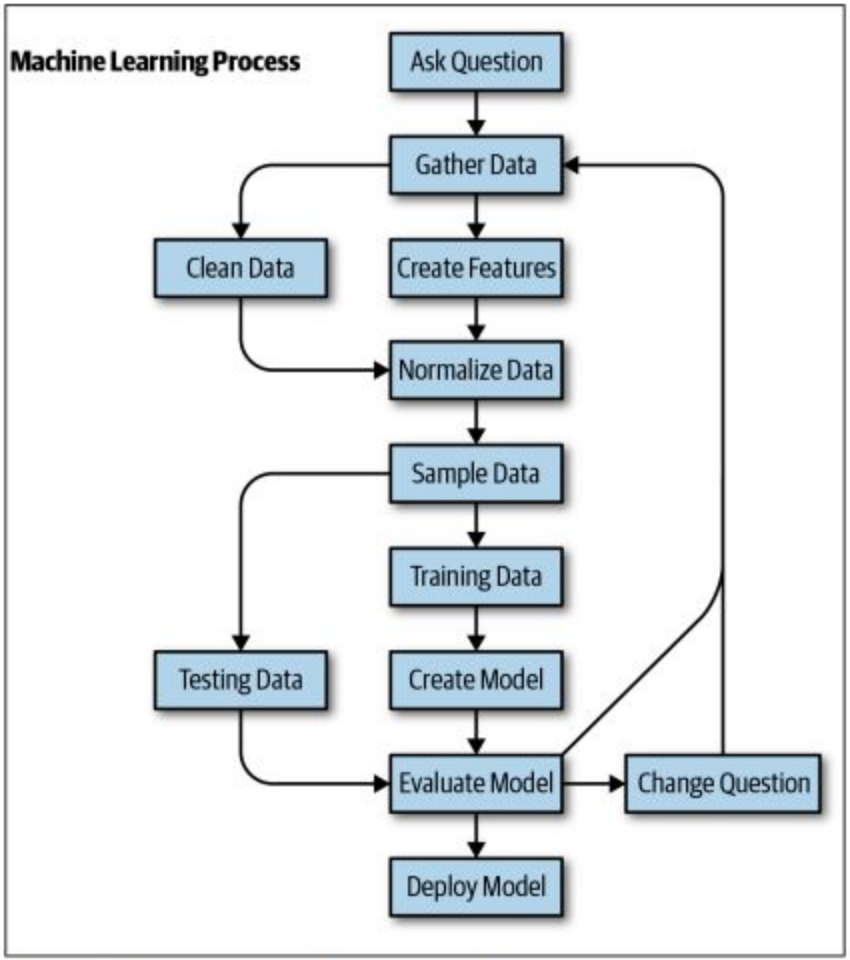


Figure 2-1. Common workflow for machine learning.

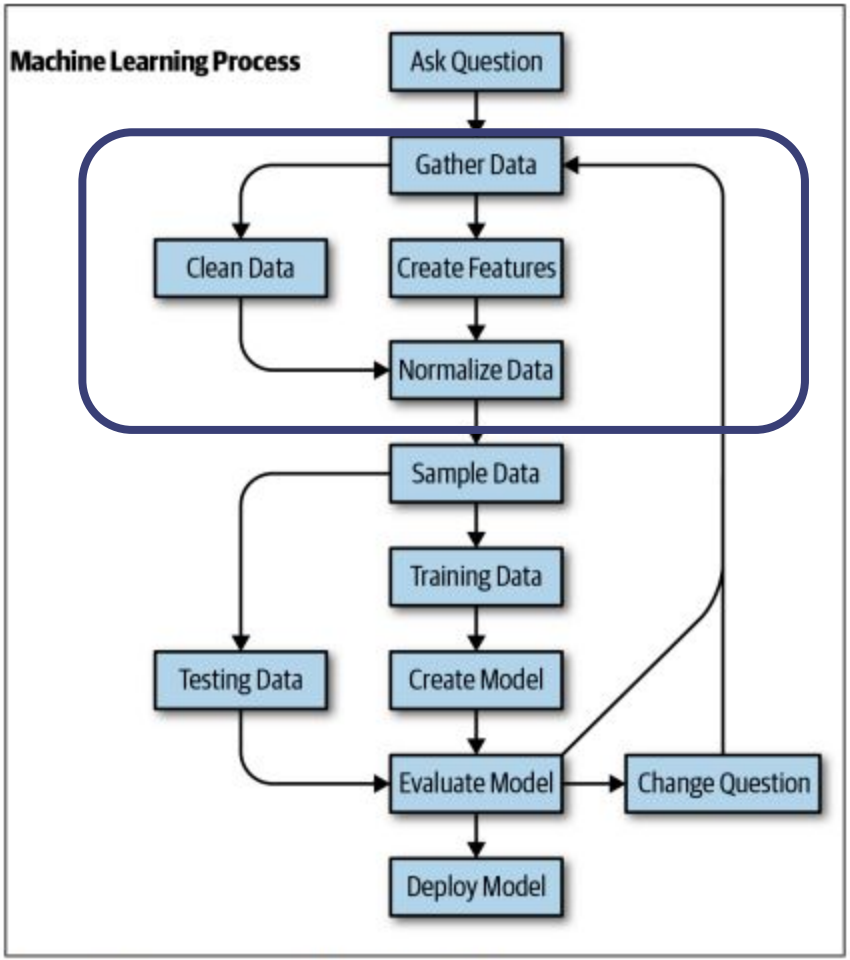


Figure 2-1. Common workflow for machine learning.

Table of contents
Regressão -Preparação de features
Import das principais funções e leitura dos dados
Preparação de dados numéricos
Preparação de dados categóricos
Limpeza das categorias
Variáveis one hot
+ Section

Preparação de dados categóricos

Para o processamento das variáveis categóricas faremos o seguinte:

1. Transformação de variáveis numéricas que representam a categoria: a. Dummy b. One hot encoding

Variáveis Dummy

Variáveis Dummy é uma representação de uma variável categórica com k categorias em k-1 variáveis, todas as novas variáveis são binárias.

Exemplo: Variável faixa de idade

Variável original	18-30	30-50	50+
0-18	0	0	0
18-30	1	0	0
30-50	0	1	0
50+	0	0	1

Para criar variáveis dummy iremos utilizar a função get_dummy do pacote Pandas

Exemplo:

```
import pandas as pd
df_dummy = pd.get_dummy(df, drop_first=True)
```

Exercícios:

Crie uma variável dummy para o conjunto de dados? Notou algum problema na criação das dummys?

```
[ ] df.dtypes
```

```
car_name              object
registration_year     object
insurance_validity    object
fuel_type             object
seats                  int64
kms_driven             int64
ownership             object
```

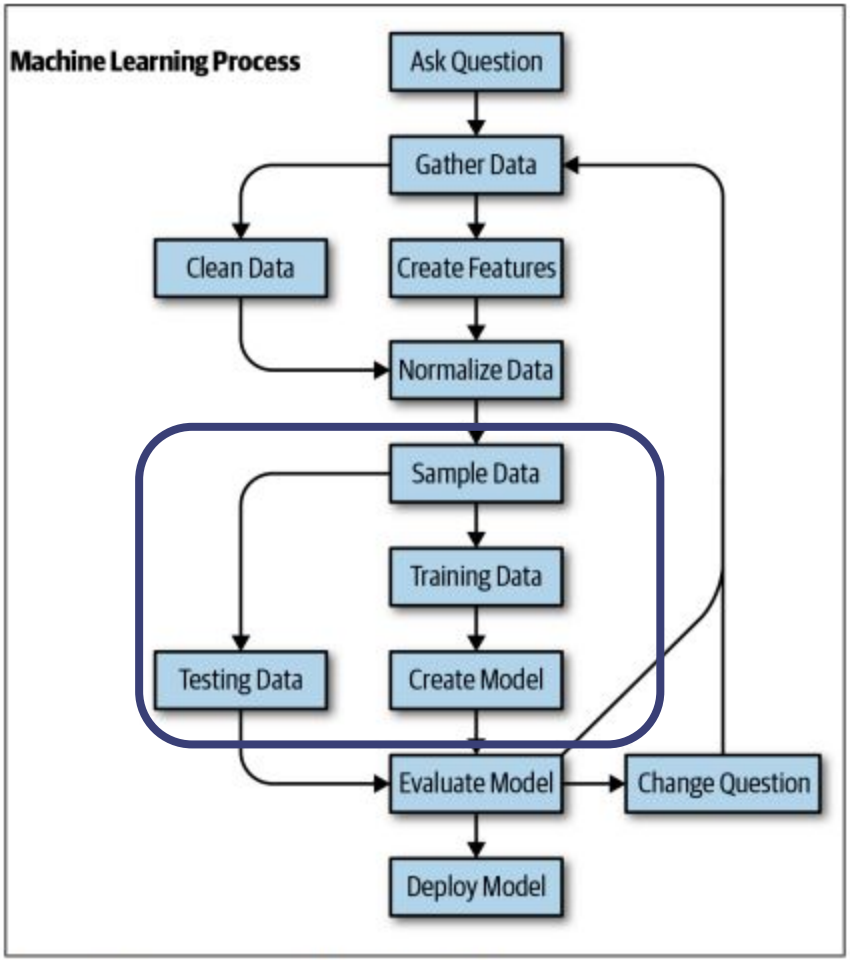
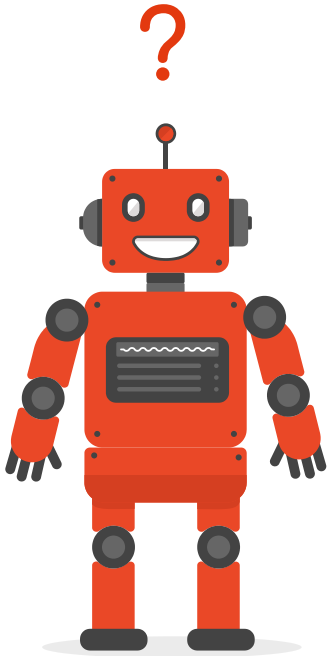
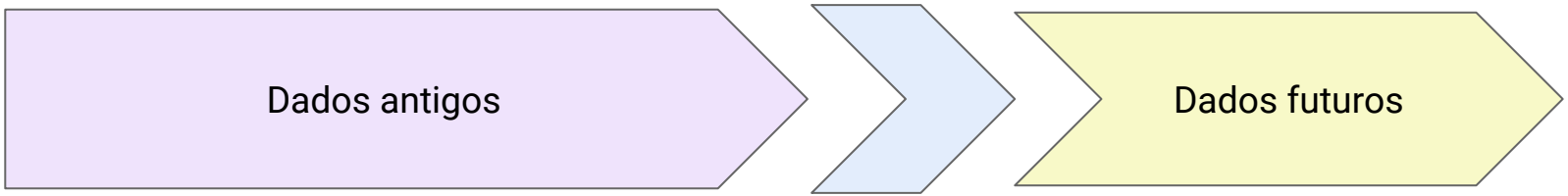
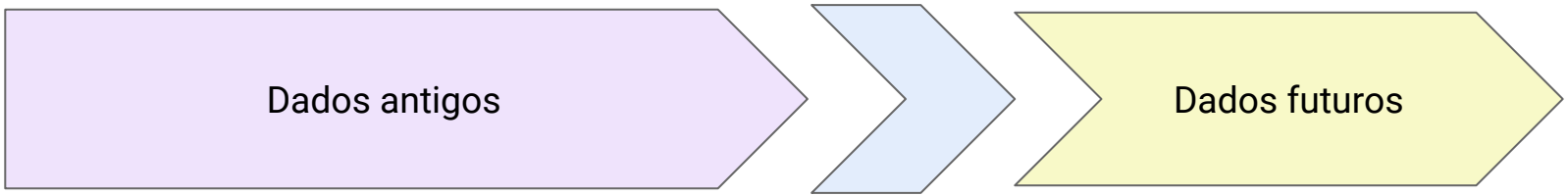


Figure 2-1. Common workflow for machine learning.

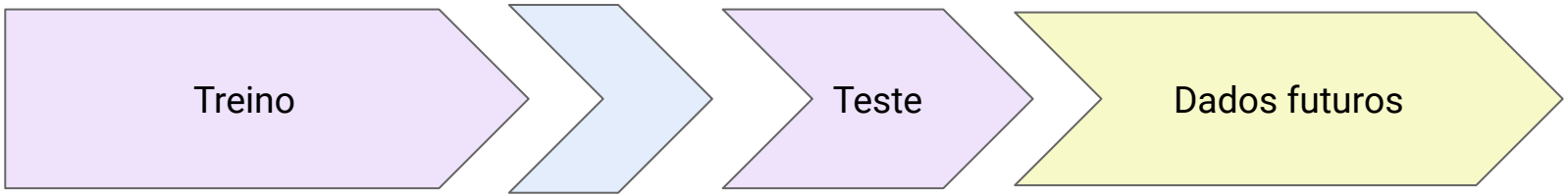
Porque separar o dado?




Dados antigos
Dados futuros



Como mensurar o comportamento do modelo no novo dado?
Qual é o melhor modelo?



Máximo de informação para aprender sobre padrões

Cenário desconhecido para ver a performance

Aplicação do modelo + monitoramento

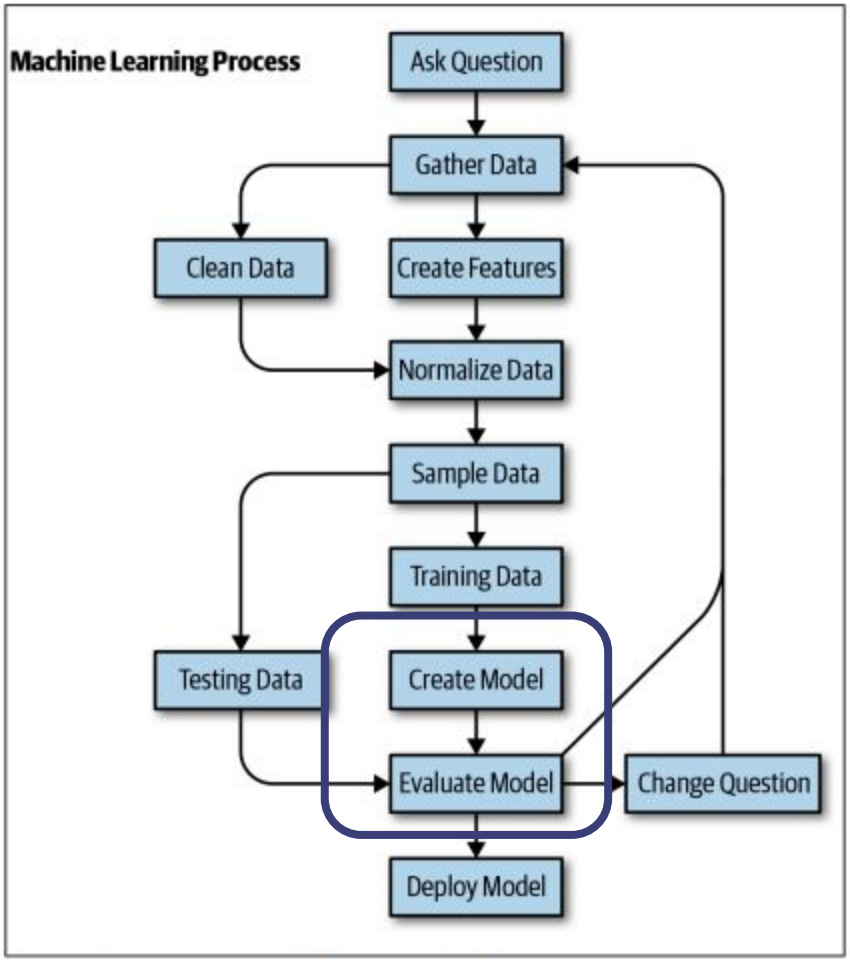


Figure 2-1. Common workflow for machine learning.

Seleção de modelos

01

Modelo de Regressão Logística

02

Árvore de decisão

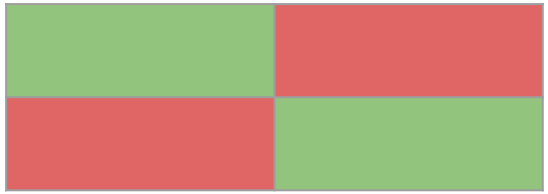
Regressão Logistica

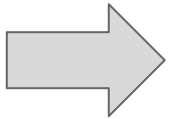
Treino

Random Forest

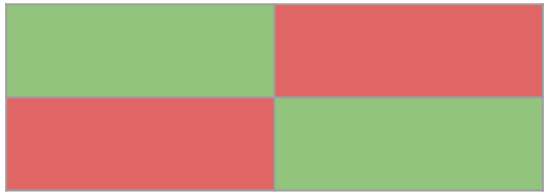
Teste

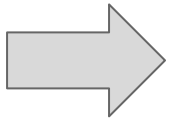
XGBoost



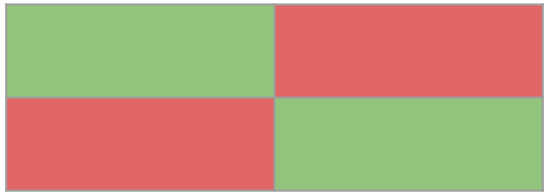


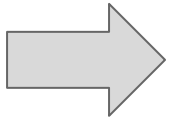
Acuracia: 90%



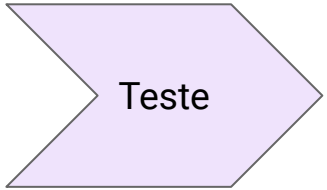


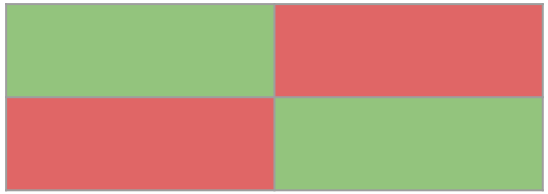
Acuracia: 85%

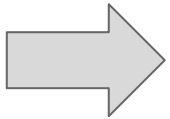




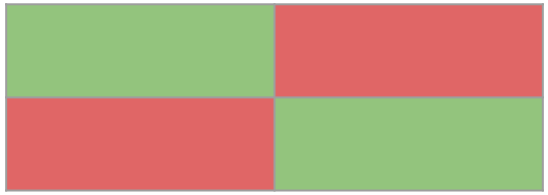
Acuracia: 93%

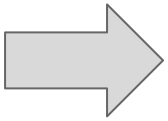




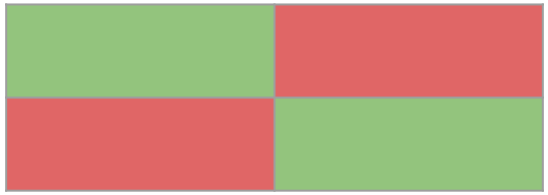


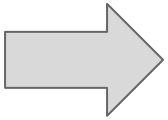
Acuracia: 90%





Acuracia: 85%





Acuracia: 93%

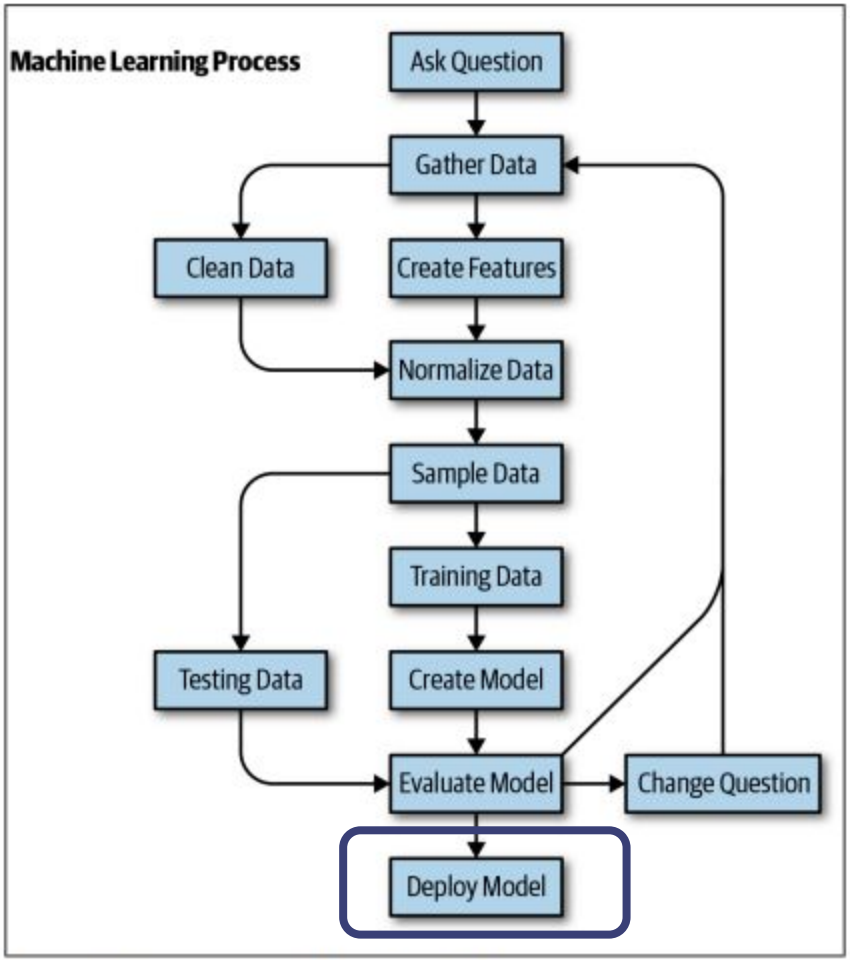
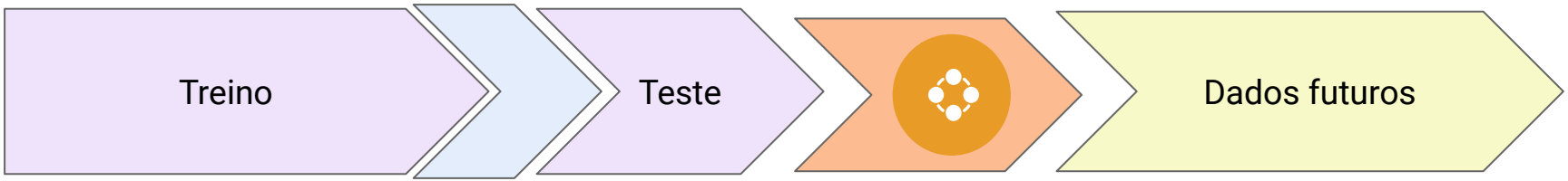


Figure 2-1. Common workflow for machine learning.

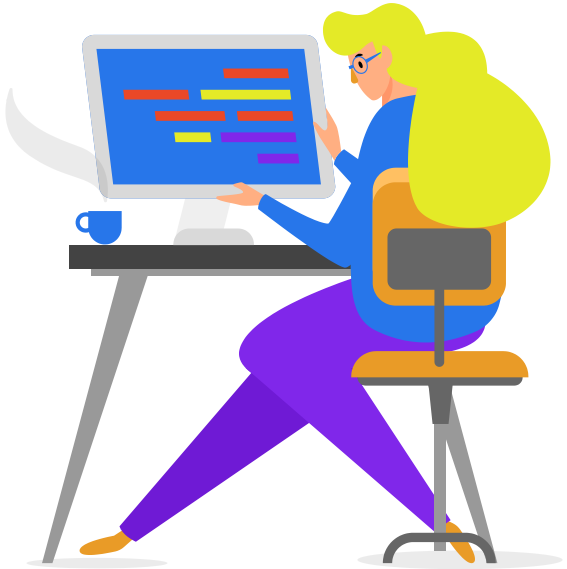

Treino
Teste
Dados futuros

Hora da prática!!!!

Separação do banco

Treino do modelo

Avaliação do modelo



Separação do banco

Treino do modelo

Treino do modelo

Treino do modelo

Avaliação do modelo