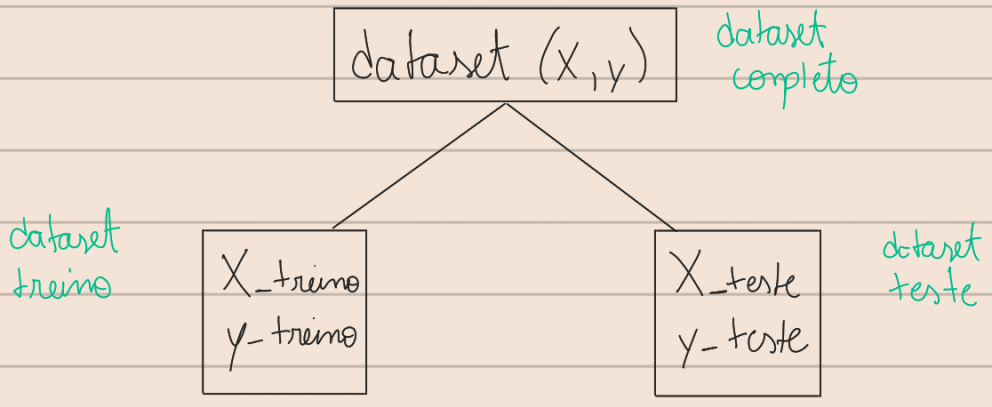
Etapas de um projeto de Machine Learning

Etapa 1: análise exploratória preliminar

- Foco é conhecer a natureza das features individualmente
  - estat descritiva
  - visualizações
- descobrir "anomalias" nos dados: outliers, erros, artefatos (saturação e etc) "buracos" e etc
- Não investigar interrelações

Etapa 2: separação treino e teste



Etapa 3: análise exploratória posterior

- Foco: analisar interrelações
  - correlações
  - visualizações cogentes

As Etapas 1, 2, 3 correspondem às atividades "Business Understanding" e "Data understanding". Ao término das etapas 1-3, temos decisões sobre: objetivos, limpeza, (filtragem) e transformações

Etapa 4: modelagem

- Filtragem: eliminar
  - não entra nas "Pipelines" do sklearn
- Transformação e modelagem
  - Pipeline do sklearn
- Escolha do melhor modelo

A Etapa 4 corresponde às atividades "Data preparation" e "modeling" do CRISP-DM

Etapa 5: certificação

- Medir desempenho do modelo escolhido
- Decidir se pode entrar em produção

≠ Corresponde ao "Evaluation" do CRISP-DM

Etapa 6: produção

- treino final e avançado de modo treinado
- para responsabilidade para equipes MLOPS/DEVOPS

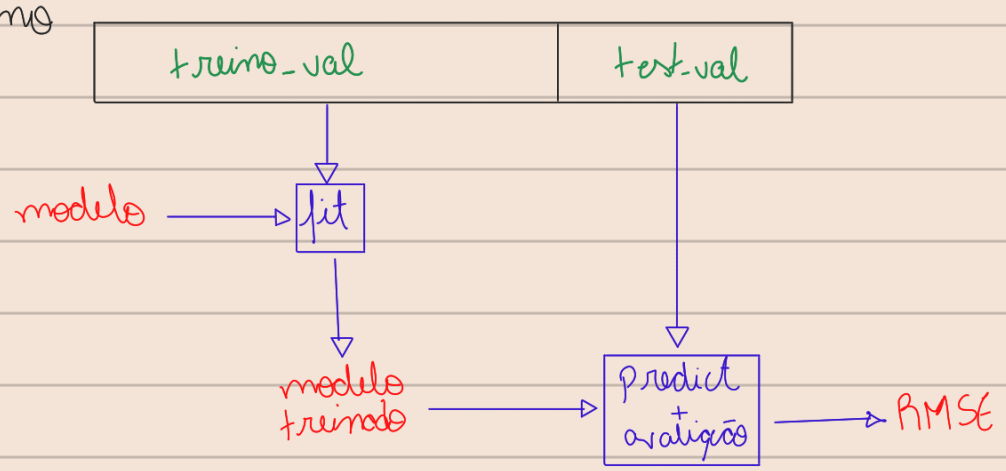
≠ Corresponde ao "Deployment" do CRISP-DM

"Evaluation"

CRISP-DM	Etapas	Nível
Business Understanding	Etapa 1	pré Nível 2
Data Understanding	Etapa 2	
	Etapa 3	
Data prep / Modeling	Etapa 4	nível 2: escolha mod
Evaluation	Etapa 5	nível 1: certificação
Deployment	Etapa 6	nível 0: deploy

Escolha do modelo

Estratégia: separação treino_val / teste_val



* em um pipeline no fit, os transformers aprendem e transformam caso não sejam o ultimo, CASO Pipe tenha um modelo os transformers vão aplicar fit_transform. Caso tenha um modelo podemos dar fit_predict, caso não apenas fit_transform

* Toda busca por modelos tem que ter Dummy, RAW Reg Lin } benchmark de "base"

Estratégia: validação cruzada

treino	1	2	3	4	5

cv = 5
cross validation

round	treino_val	teste_val
1	2 3 4 5	1
2	3 4 5 1	2
3	4 5 1 2	3
4	5 1 2 3	4
5	1 2 3 4	5

5 RMSE