MACHINE LEARNING APLICADO À ANÁLISE DE DADOS

AULA 6 – 23/03/2021

INTRODUÇÃO A MACHINE LEARNING COM SCIKIT-LEARN

SCIKIT LEARN

- É uma biblioteca do Python que nos permite construir modelos preditivos. Ela implementa uma ampla variedade de algoritmos e processos de Machine Learning para análises avançadas.

PROBLEMA DE APRENDIZAGEM

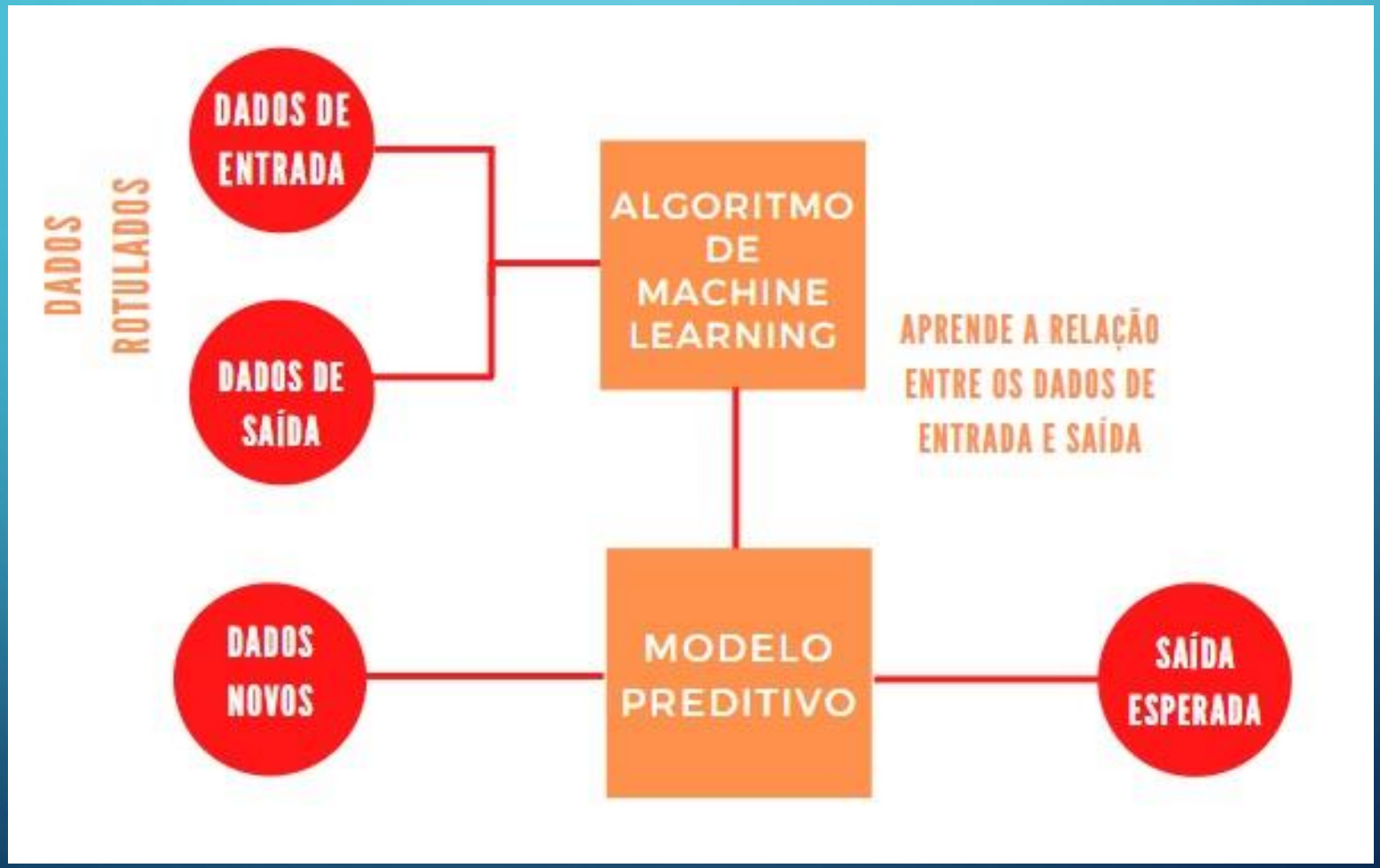
CONJUNTO DE AMOSTRA DE DADOS

PREVER PROPRIEDADES EM DADOS DESCONHECIDOS

APRENDIZADO SUPERVISIONADO

Quando um programa é treinado sobre um conjunto de dados pré-definidos, assim ele poderá tomar decisões precisas quando receber novos dados. Os dados pré-definidos (dados de treino) precisam conter valores de entrada e saída, para que o modelo aprenda, a partir de novos dados de entrada, a gerar a saída correta.

MODELO DE APRENDIZAGEM SUPERVISIONADA



CLASSIFICAÇÃO

Possui o objetivo de classificar dados com base em rótulos. É necessário ter uma base de dados histórica, que irá ser utilizada como treinamento do algoritmo, para classificar novos registros de dados. Sistemas de classificação são usados geralmente quando as previsões são de naturezas distintas, ou seja, ‘sim’ ou ‘não’.

EXEMPLO: CLASSIFICAÇÃO DE ANIMAIS

	type	name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone
1	mammal	aardvark	yes	no	no	yes	no	no	yes	yes	yes
2	mammal	antelope	yes	no	no	yes	no	no	no	yes	yes
3	fish	bass	no	no	yes	no	no	yes	yes	yes	yes
4	mammal	bear	yes	no	no	yes	no	no	yes	yes	yes
5	mammal	boar	yes	no	no	yes	no	no	yes	yes	yes
6	mammal	buffalo	yes	no	no	yes	no	no	no	yes	yes
7	mammal	calf	yes	no	no	yes	no	no	no	yes	yes
8	fish	carp	no	no	yes	no	no	yes	no	yes	yes
9	fish	catfish	no	no	yes	no	no	yes	yes	yes	yes
10	mammal	cavy	yes	no	no	yes	no	no	no	yes	yes
11	mammal	cheetah	yes	no	no	yes	no	no	yes	yes	yes
12	bird	chicken	no	yes	yes	no	yes	no	no	no	yes

(UCI Machine Learning Repository: Zoo Data Set)

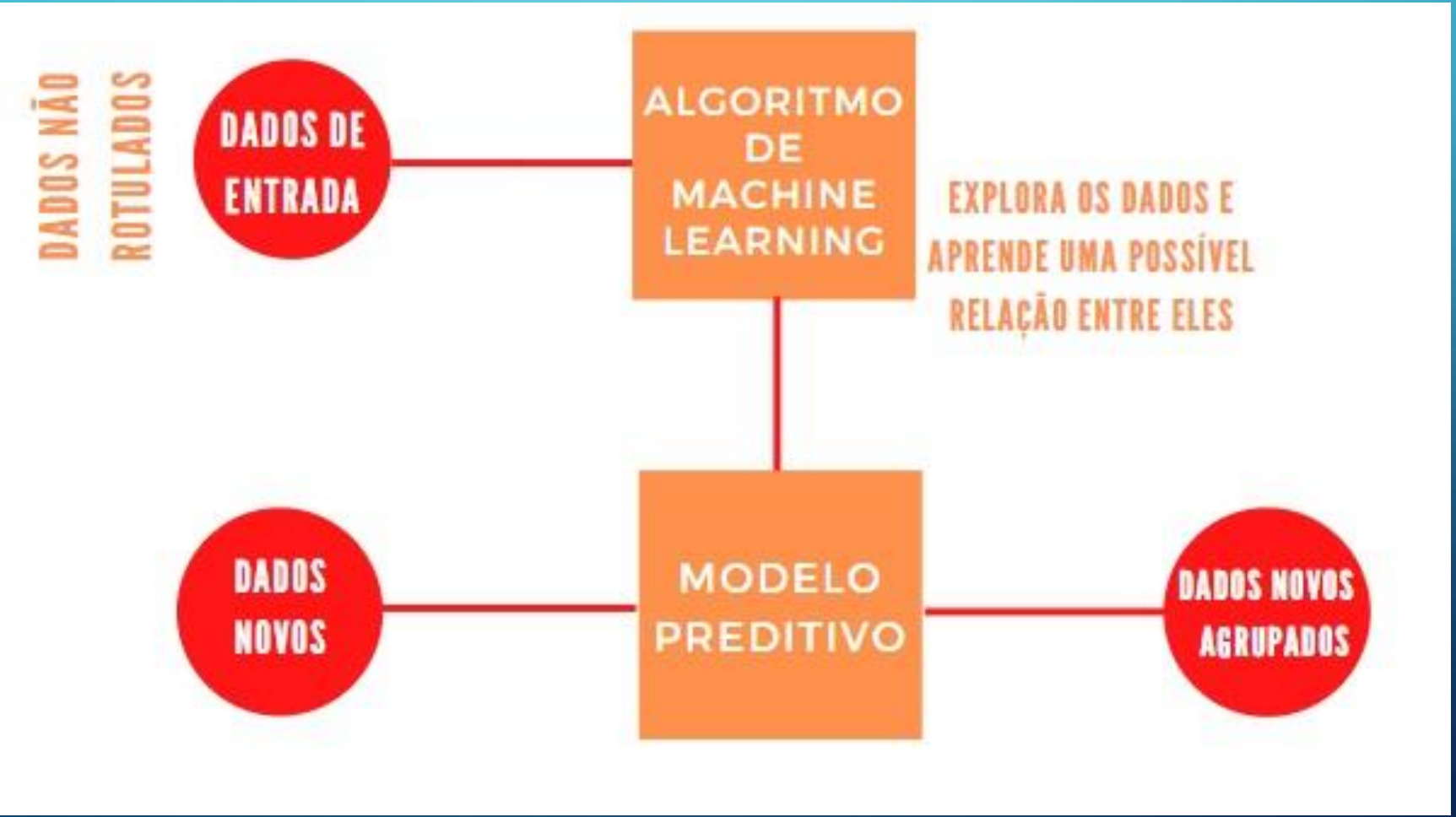
REGRESSÃO

Possui o objetivo de prever valores e comportamentos a partir da análise de dados. É usada quando o valor que está sendo previsto difere do ‘sim’ ou ‘não’. Nesse caso, o modelo pode aprender uma função para prever o preço de um produto, por exemplo.

APRENDIZADO NÃO-SUPERVISIONADO

- Nesse tipo de aprendizado, o programa pode automaticamente encontrar padrões e relações em um conjunto de dados. Diferente do aprendizado supervisionado, esse modelo aprende a executar uma tarefa a partir de dados não-rotulados (sem um resultado conhecido).

MODELO DE APRENDIZAGEM NÃO-SUPERVISIONADA

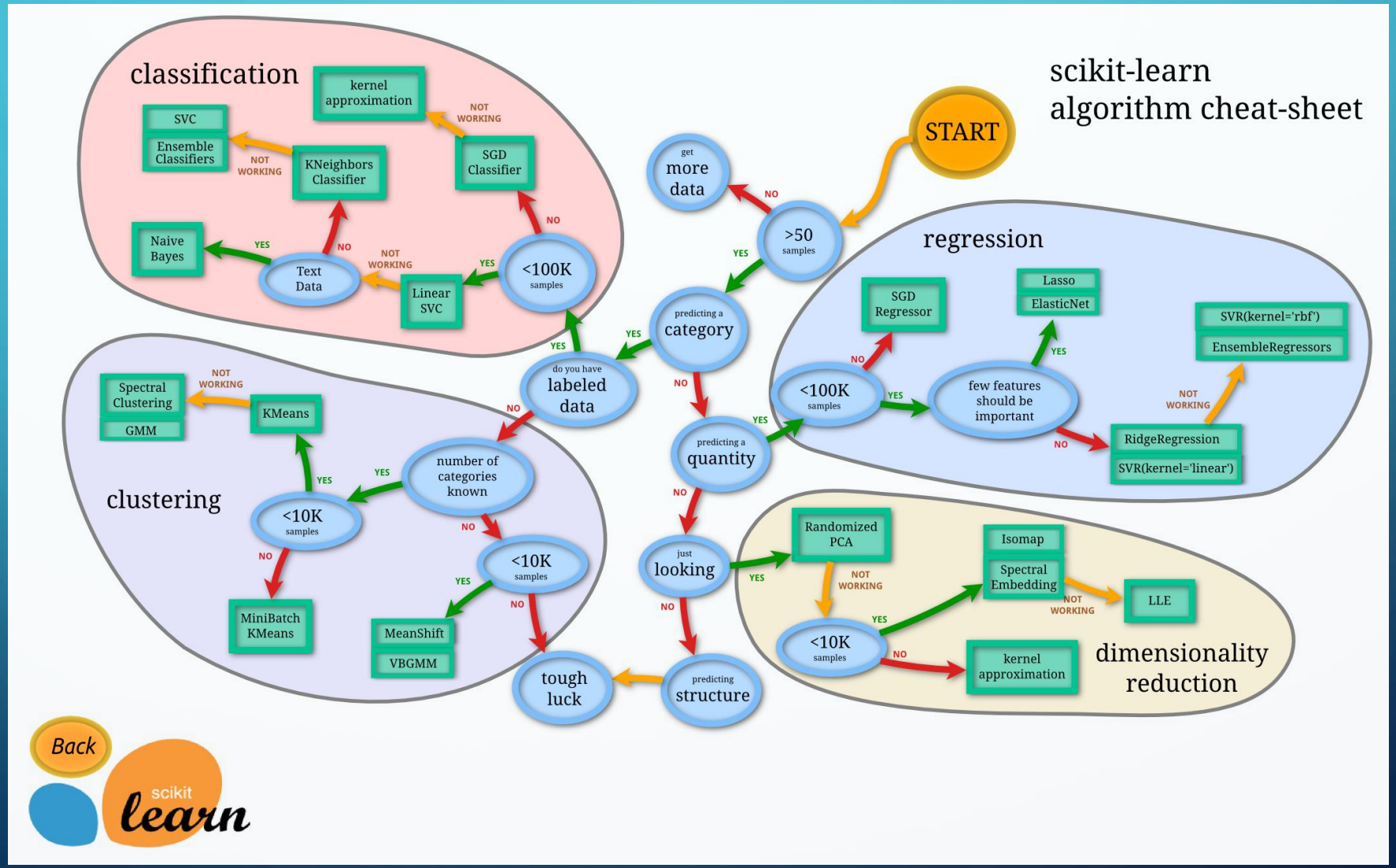


AGRUPAMENTOS (CLUSTERING)

- O algoritmo possui a tarefa de separar dados em grupos, segmentado-os por características similares. Basicamente, o algoritmo irá agrupar dados para achar um padrão.

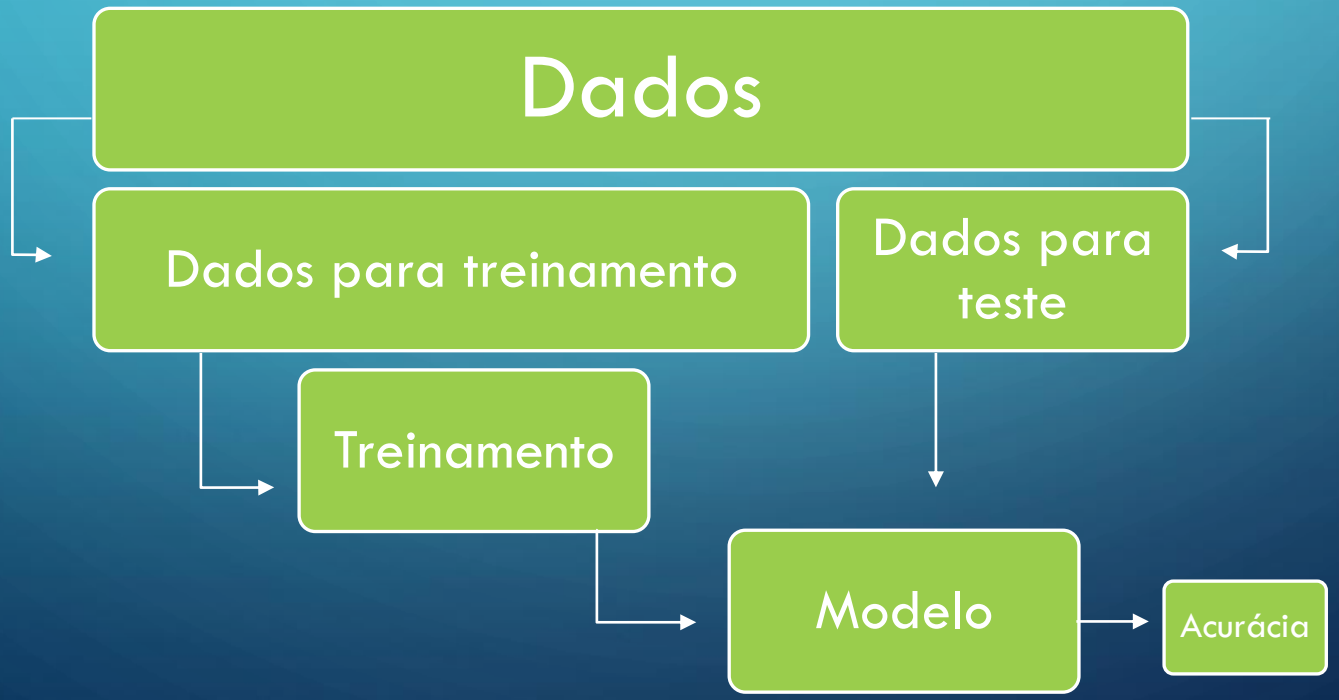
Ex.: divisão de clientes em grupos com base em suas preferências

COMO IDENTIFICAR O ALGORITMO DE MACHINE LEARNING:



TREINAMENTO, VALIDAÇÃO E TESTE

- Em qualquer processo de aprendizagem é necessário que haja treinamento, validação do conhecimento adquirido e teste.



A **acurácia** define o nível de exatidão dos resultados obtidos pelas aplicações de tecnologia.

DIVISÃO COMUM

Treinamento, validação e teste

75% a 70% - dados de treino

25% a 30% - dados de teste

DIVISÃO COM CONJUNTO DE VALIDAÇÃO

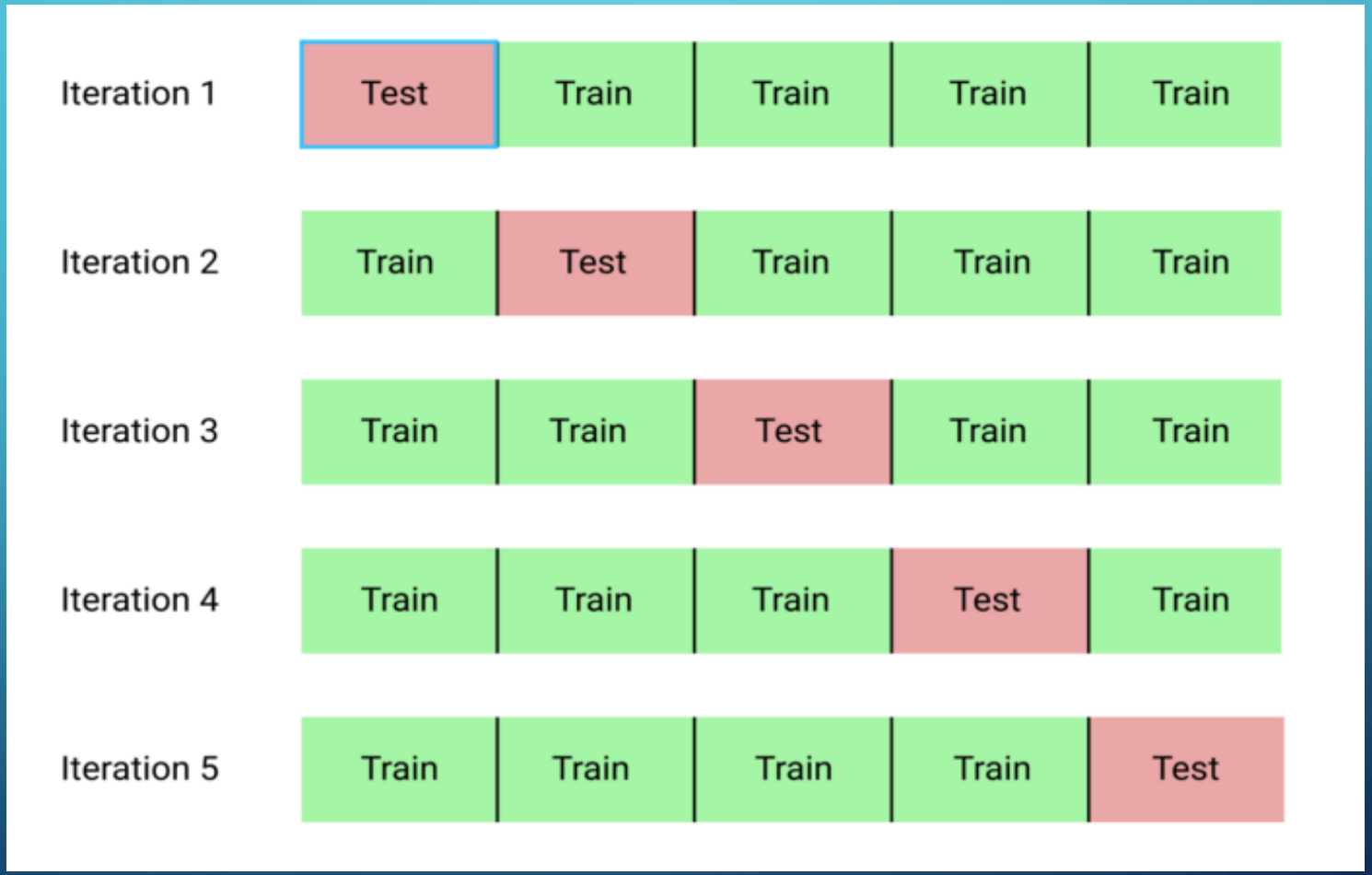
Treinamento, validação e teste

70% - dados de treino

20% - dados de validação

10% - dados de teste

CROSS-VALIDATION



(Fonte: Como saber se seu modelo de Machine Learning está funcionando mesmo | by Paulo Vasconcellos)

LINKS ÚTEIS

- Scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation (https://scikit-learn.org/stable/)
- Como saber se seu modelo de Machine Learning está funcionando mesmo | by Paulo Vasconcellos (https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-est%C3%A1-funcionando-mesmo-a5892f6468b)