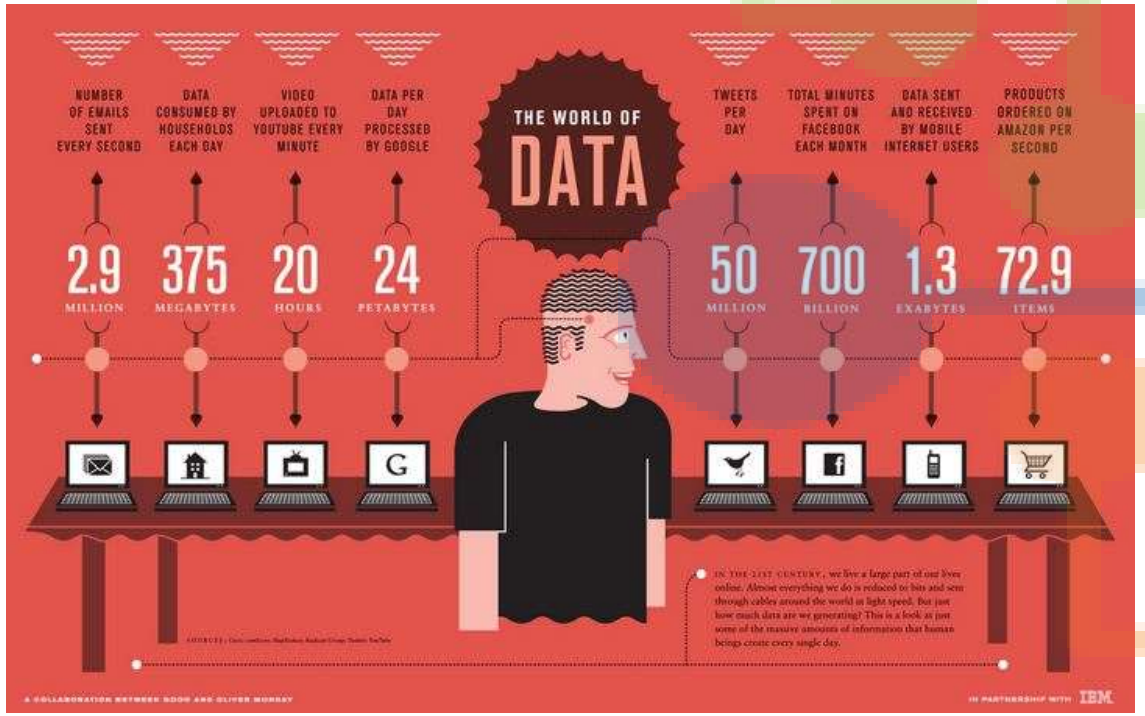
Big Data Real-Time Analytics com Python e Spark

Seja muito bem-vindo(a)!

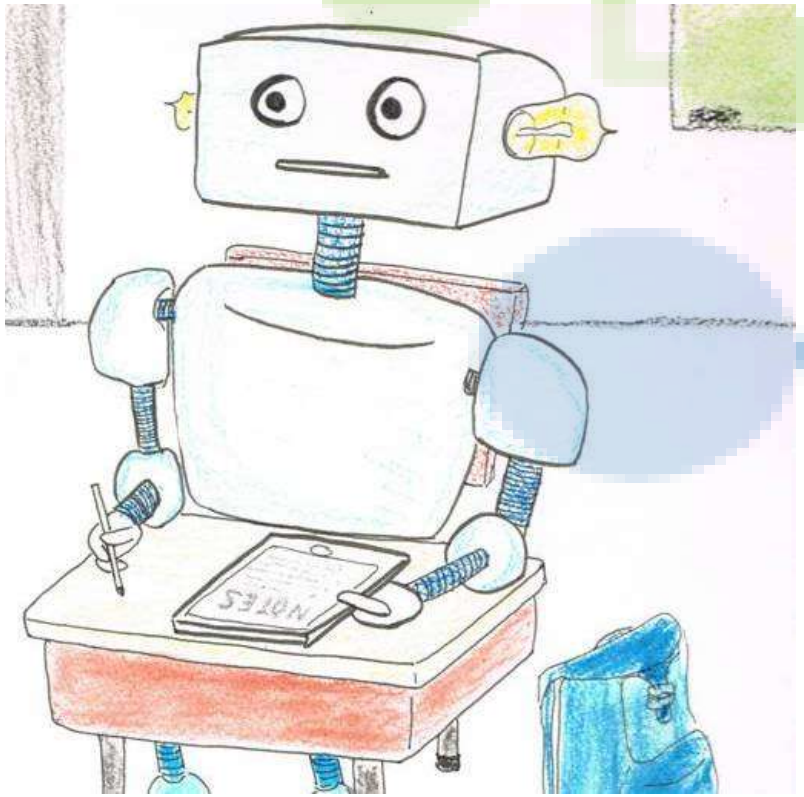
Machine Learning em Python

Introdução



Estamos no meio de uma revolução, proporcionada pelo Big Data.

Introdução

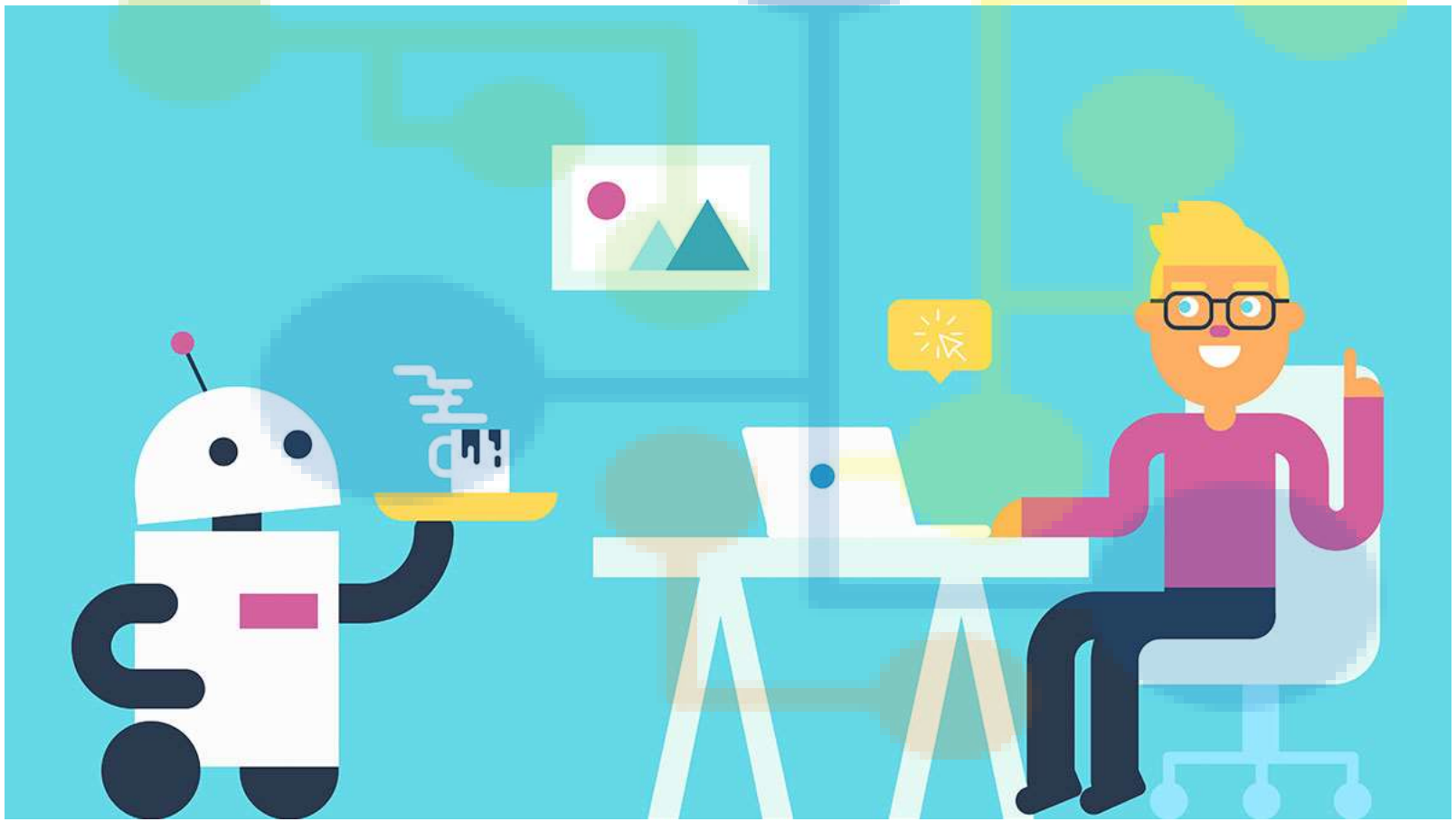


Machine Learning

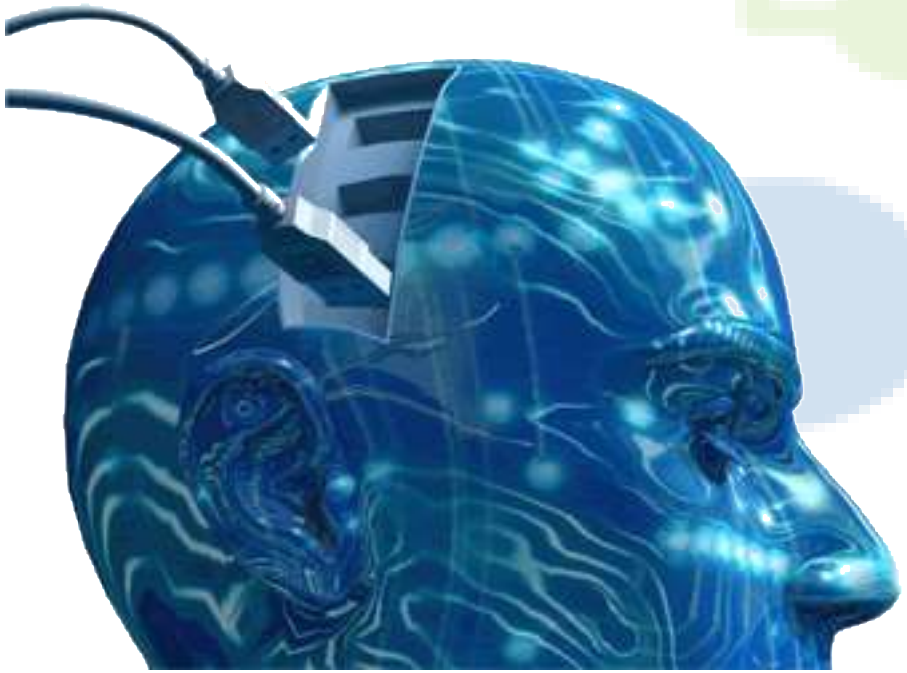
Introdução

Considerar uma carreira em Machine Learning e aprender tudo que for possível sobre esse assunto é uma das decisões mais inteligentes que você pode tomar na sua carreira professional.

Introdução



Introdução



Computação Cognitiva

Introdução

O que vamos estudar neste capítulo?

- Processo de Machine Learning
- Biblioteca Scikit-learn
- Coleta, Análise Exploratória e Pré-Processamento
- Feature Selection
- Algoritmos de Machine Learning – Classificação
- Algoritmos de Machine Learning – Regressão
- Métodos Ensemble
- Algoritmo XGBoost

Introdução

Este capítulo não é um curso de Machine Learning, mas sim uma introdução ao tema. O curso de Machine Learning é o número 4 da Formação Cientista de Dados e aborda o tema em profundidade. Nosso objetivo aqui é que você aprenda como aplicar Machine Learning, pois mais a frente no curso aplicaremos aprendizado de máquina em dados gerados em tempo real.

Algoritmos de Machine Learning

Algoritmos de Machine Learning

Hoje, como Cientista de Dados, é possível construir sistemas que trituram dados com algoritmos complexos, tudo com baixo custo e alta capacidade de processamento.

Algoritmos de Machine Learning

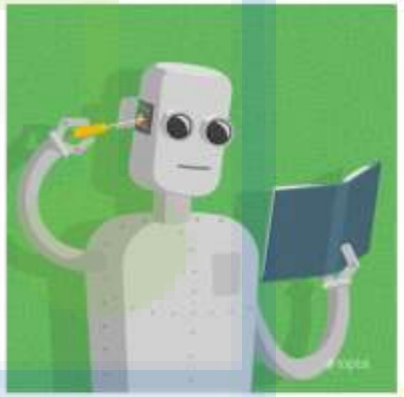
Machine Learning, ou seja a aplicação de algoritmos e ciência para extrair informações de dados, é um dos campos mais espetaculares da ciência da computação atualmente.

Algoritmos de Machine Learning

Big Data

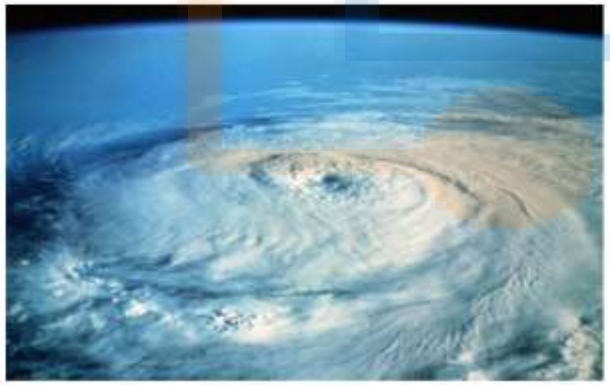
+

Machine Learning

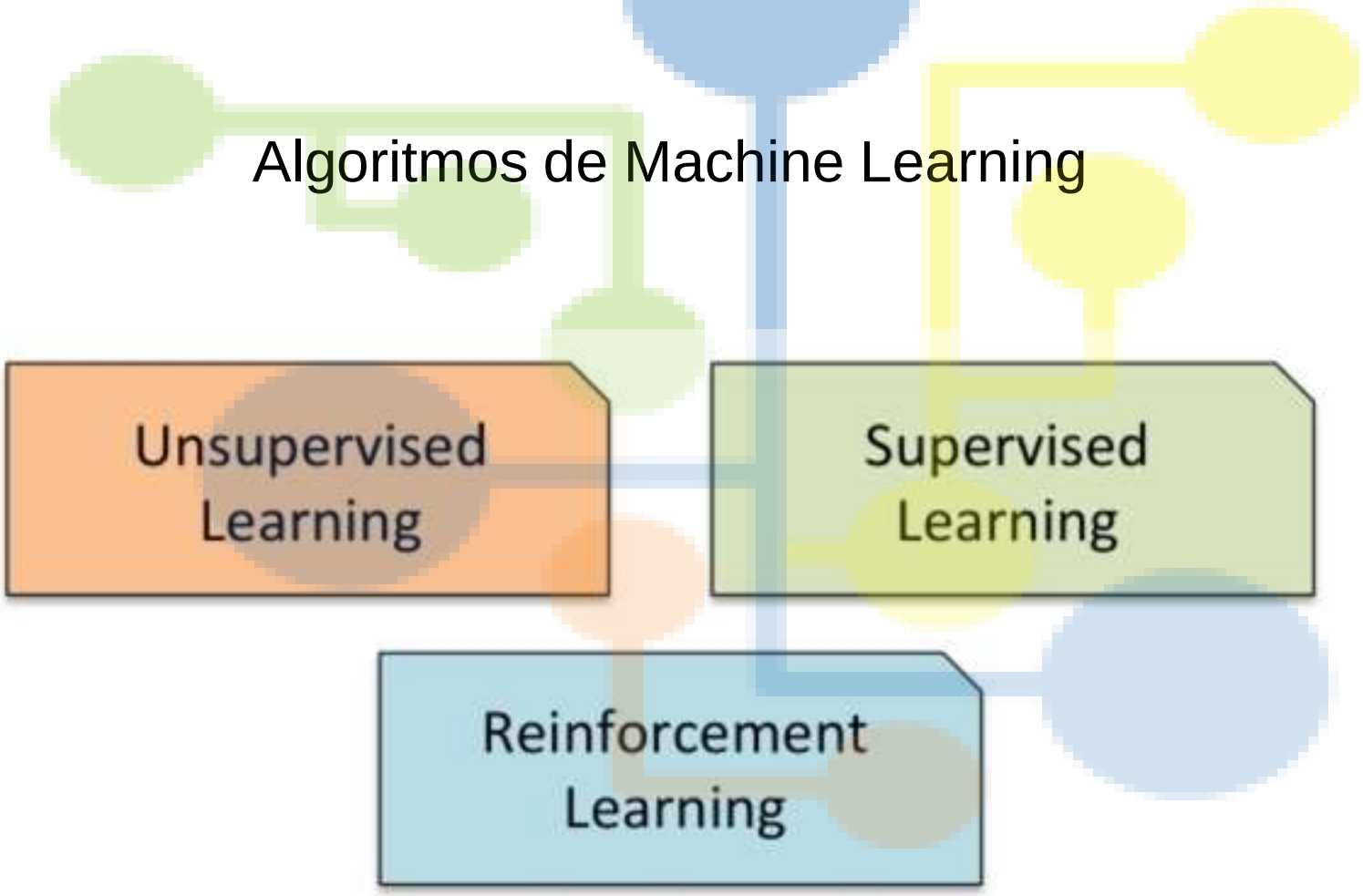


+

Capacidade Computacional



Algoritmos de Machine Learning



Aprendizagem Supervisionada

Aprendizagem Supervisionada

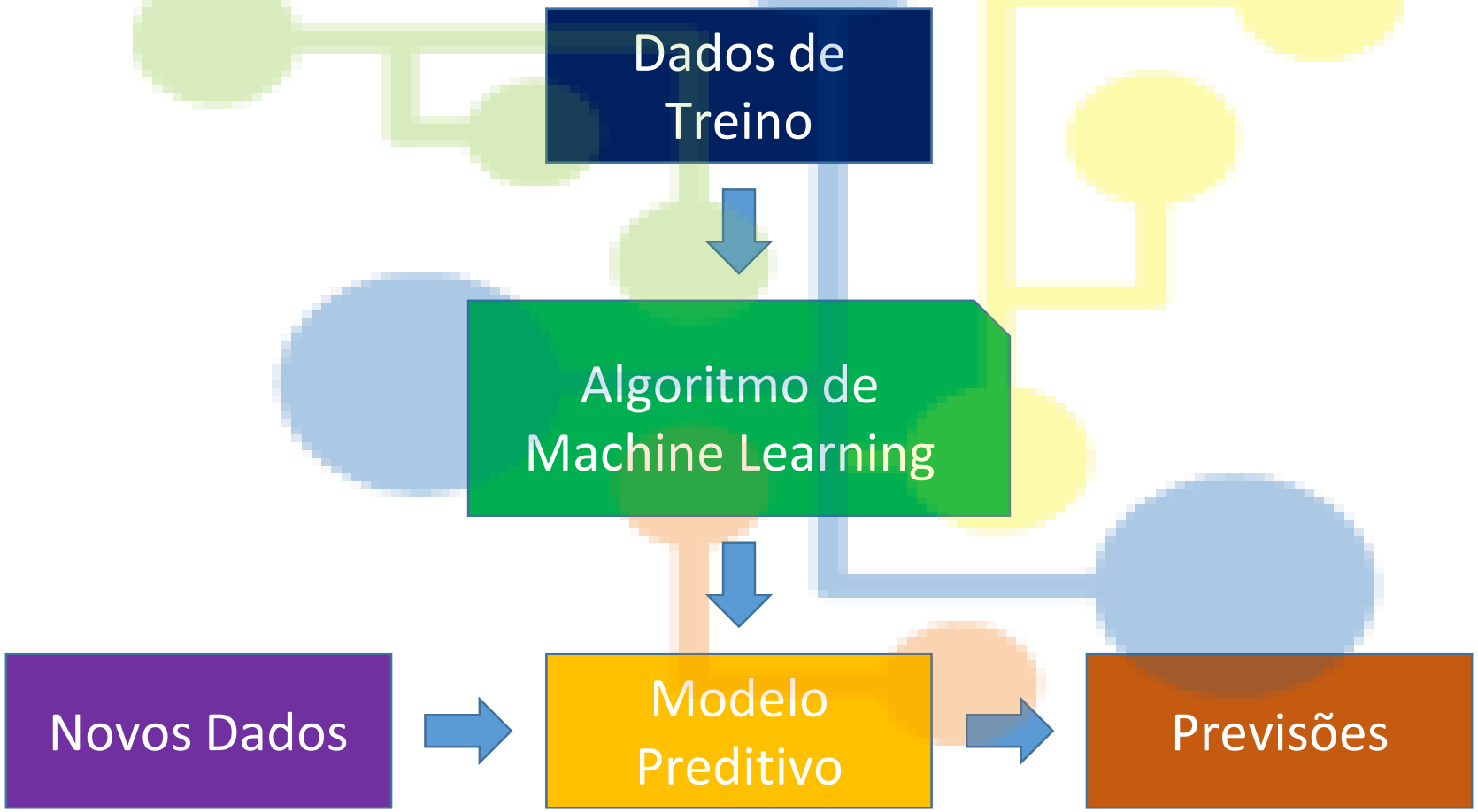
Como uma criança aprende?

Um professor apresenta imagens, textos ou objetos informando para a criança o que aquilo representa.

Por exemplo, o professor apresenta a foto de um carro, explicando suas principais características.

Mais tarde, quando a criança encontrar algo com as mesmas características será capaz de reconhecer que se trata de um carro.

Aprendizagem Supervisionada



Aprendizagem Não Supervisionada

Aprendizagem Não Supervisionada

Na aprendizagem supervisionada, temos as entradas (as características) e temos as saídas. O algoritmo então aprende o relacionamento nos dados e um modelo é criado. Quando o modelo é apresentado a novos dados de entrada, é capaz de prever as saídas.

Aprendizagem Não Supervisionada

Na aprendizagem supervisionada, temos as entradas (as características) e temos as saídas. O algoritmo então aprende o relacionamento nos dados e um modelo é criado. Quando o modelo é apresentado a novos dados de entrada, é capaz de prever as saídas.

Mas e quando não temos os dados de saída?

Aprendizagem Não Supervisionada

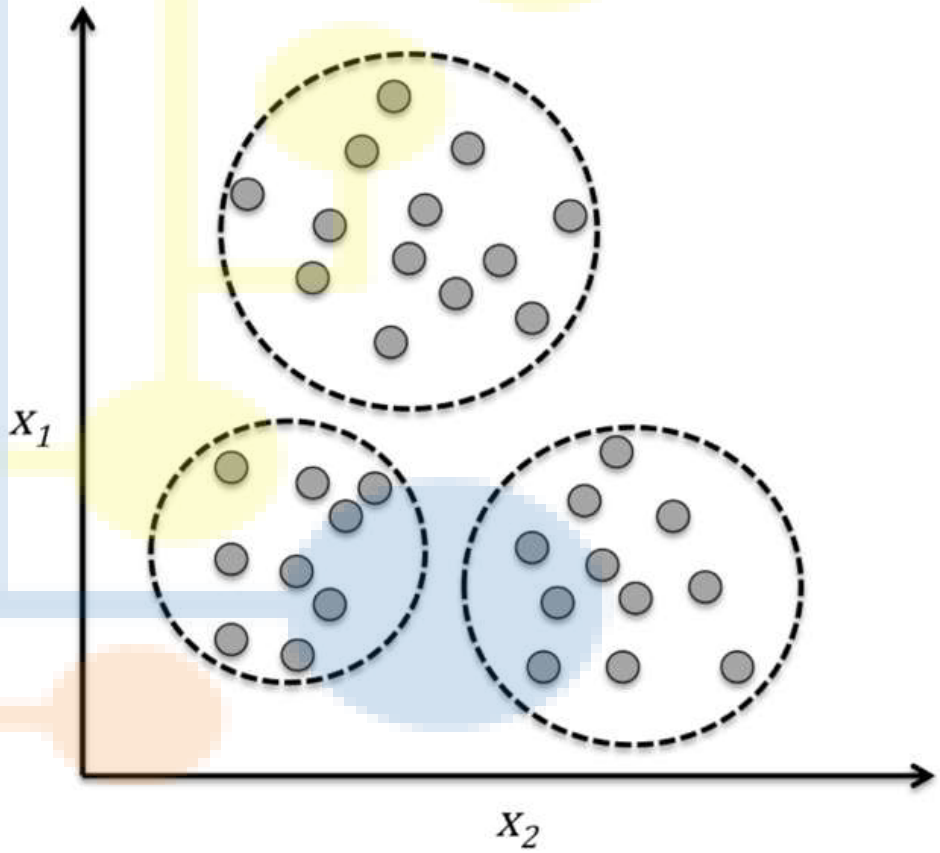
Atributo 1	Atributo 2	Atributo 3	Saída
x1	x2	x3	Carro
x4	x5	x6	Avião
x7	x8	x9	?
x10	x11	x12	?

Aprendizagem Supervisionada

Aprendizagem Não Supervisionada

Aprendizagem Não Supervisionada

Na aprendizagem não supervisionada, temos as entradas (as características) mas não temos as saídas. O algoritmo aprende o relacionamento nos dados e gera agrupamentos (clusters).



Aprendizagem Por Reforço

Aprendizagem Por Reforço

Como alguém aprende a andar de bicicleta?

Podemos usar a aprendizagem supervisionada?

E a aprendizagem não supervisionada?

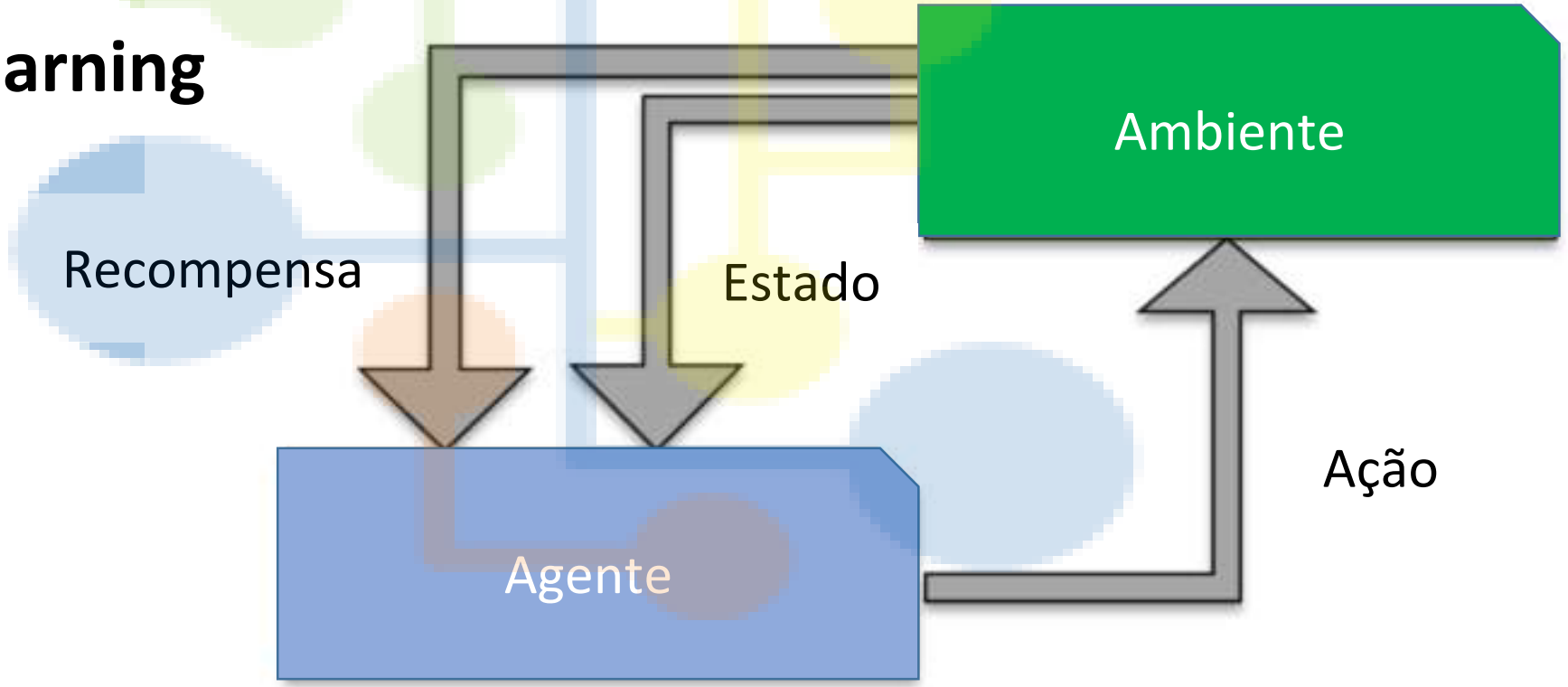
Qual seria o melhor método neste caso?

Aprendizagem Por Reforço

Tentativa e Erro

Aprendizagem Por Reforço

Aprendizagem Por Reforço ou Reinforcement Learning

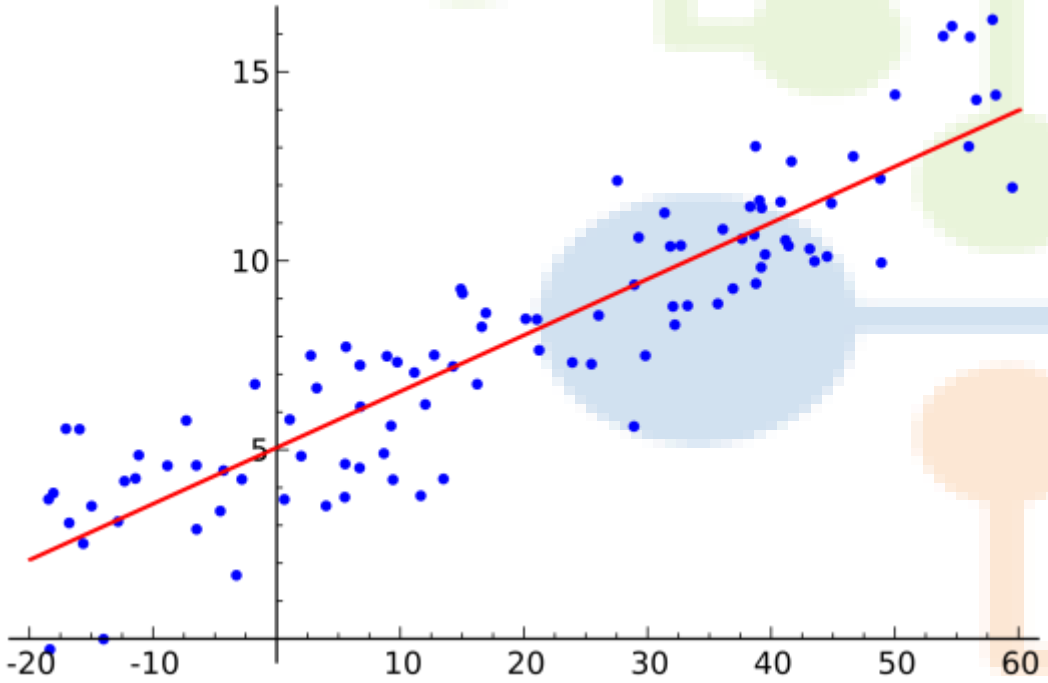


Principais Algoritmos de Machine Learning

Principais Algoritmos de Machine Learning

Quando trabalhamos com aprendizagem supervisionada temos basicamente 2 tipos de algoritmos:

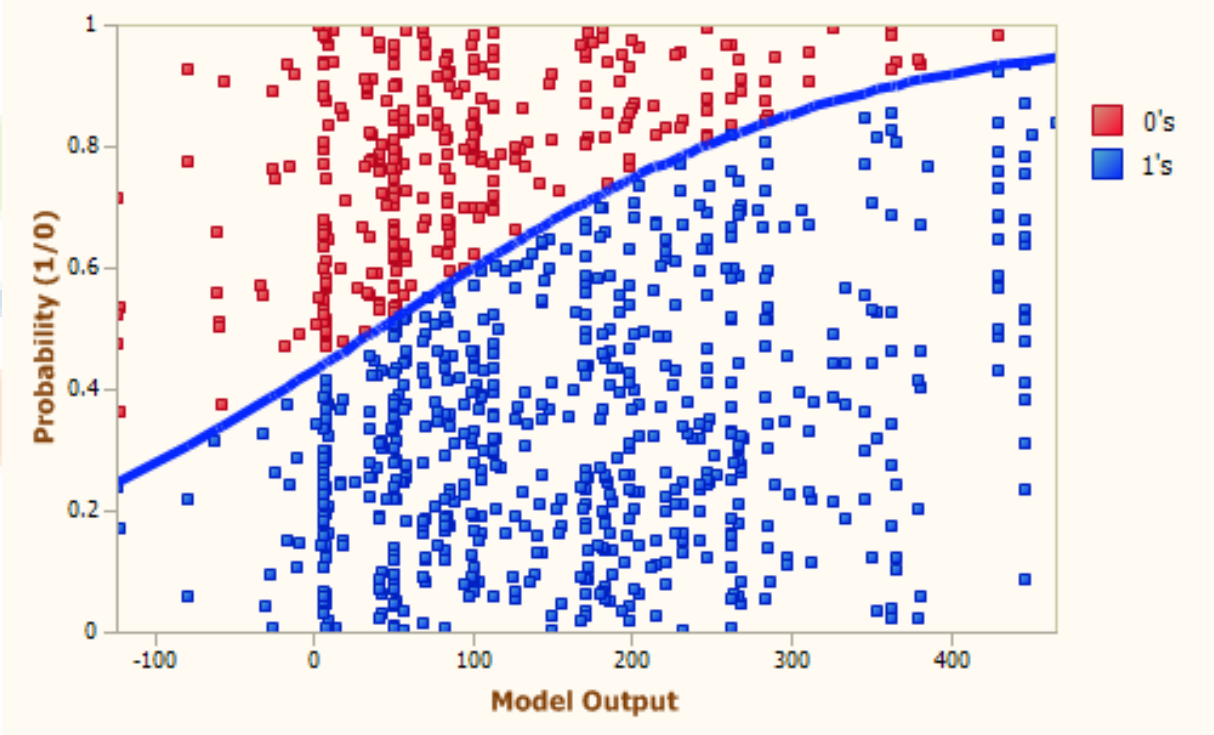
Principais Algoritmos de Machine Learning



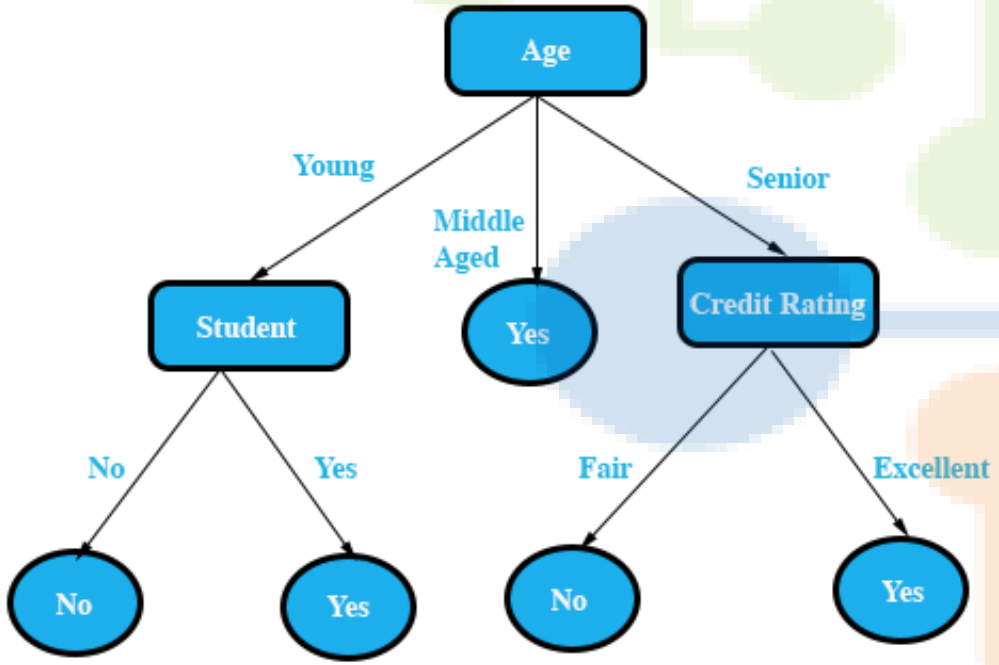
Regressão Linear

Principais Algoritmos de Machine Learning

Regressão Logística



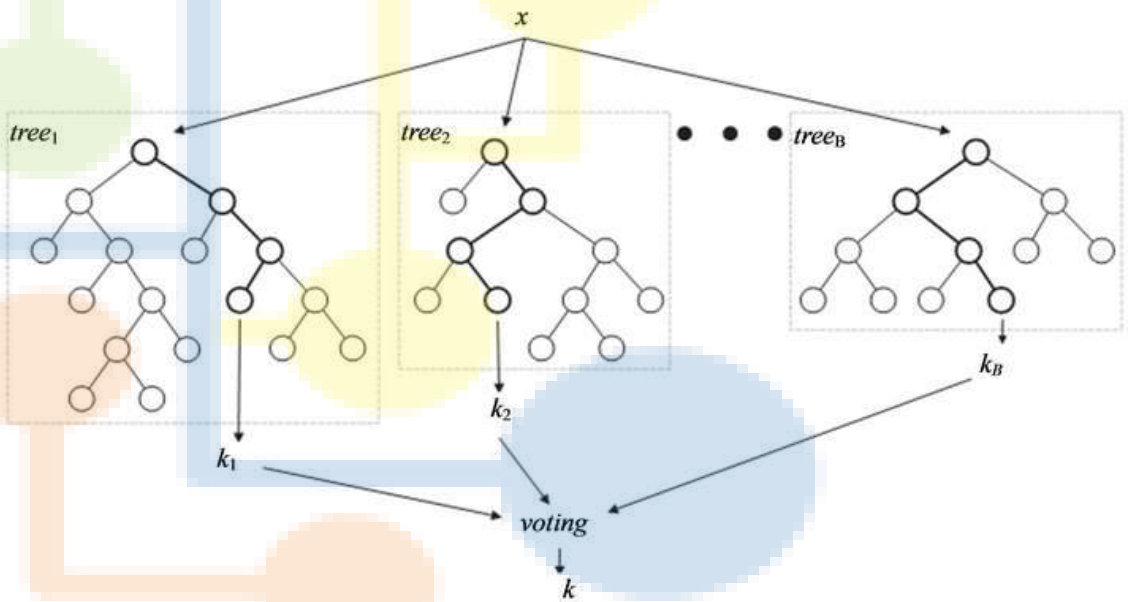
Principais Algoritmos de Machine Learning



Árvores de Decisão

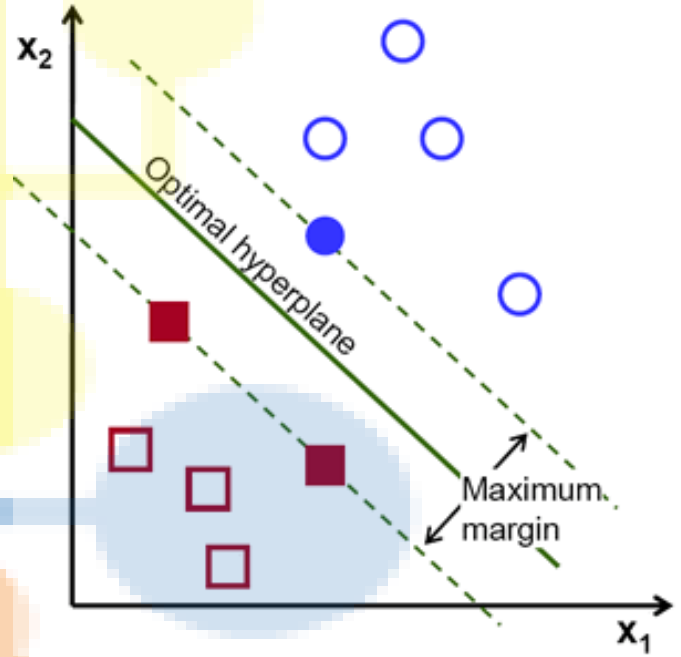
Principais Algoritmos de Machine Learning

Random Forest
(Floresta Aleatória)

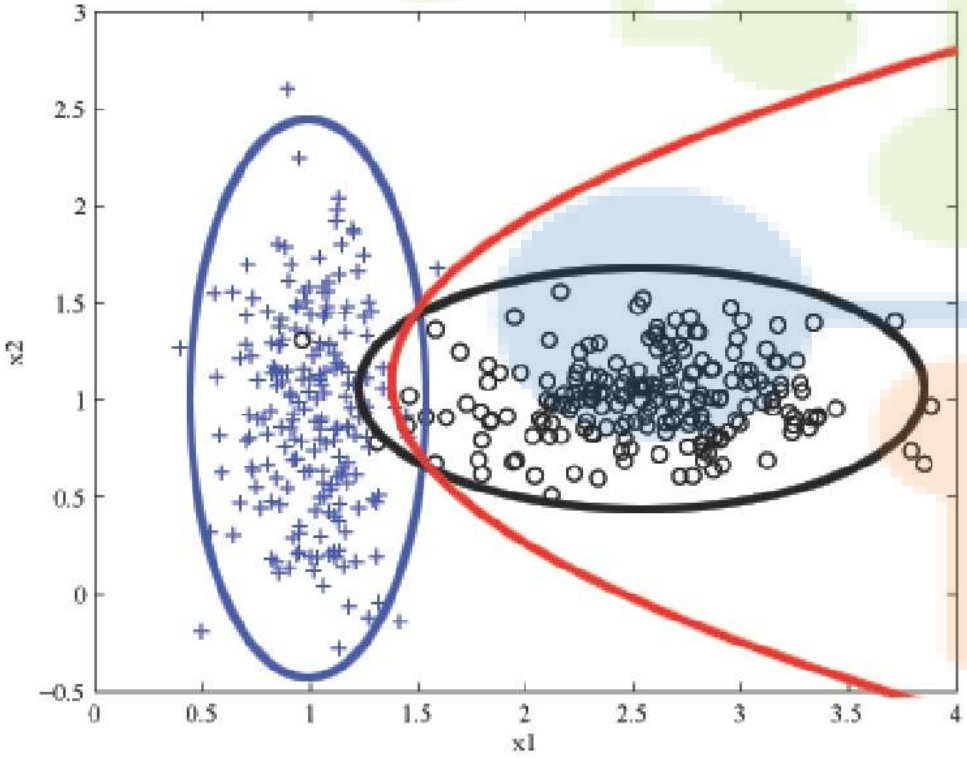


Principais Algoritmos de Machine Learning

Support Vector Machines
(Máquinas de Vetor de Suporte)



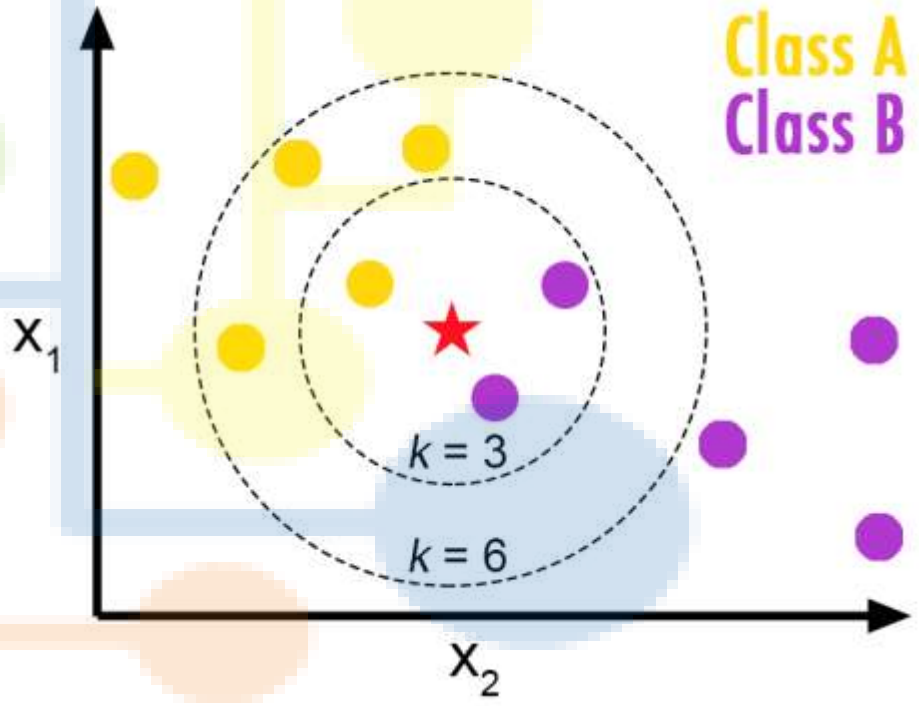
Principais Algoritmos de Machine Learning



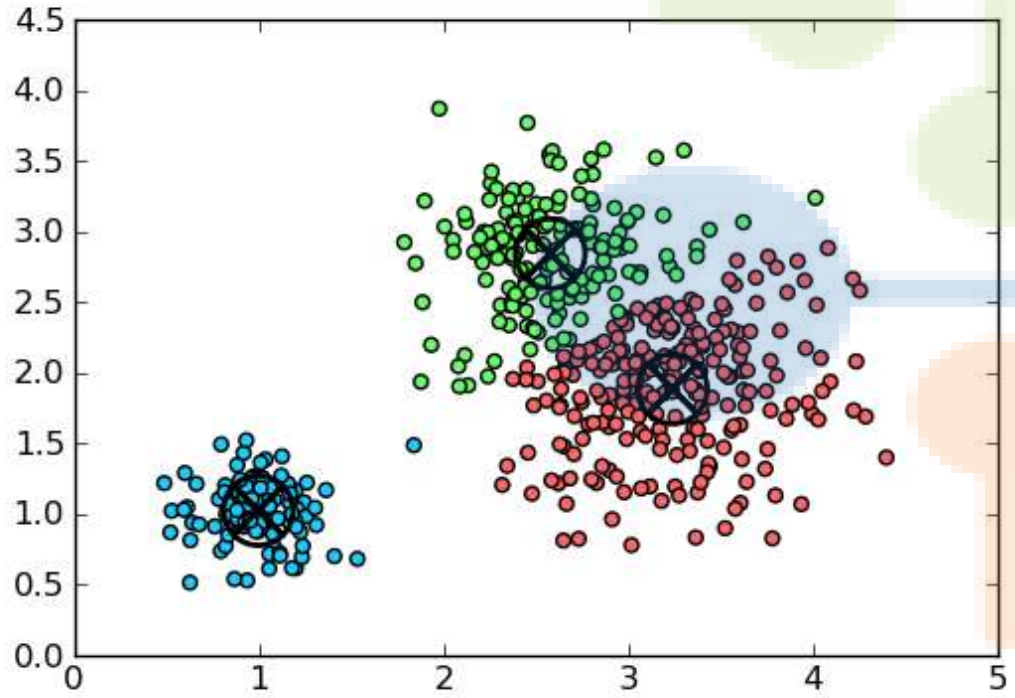
Naive Bayes

Principais Algoritmos de Machine Learning

K-Nearest Neighbors (KNN)



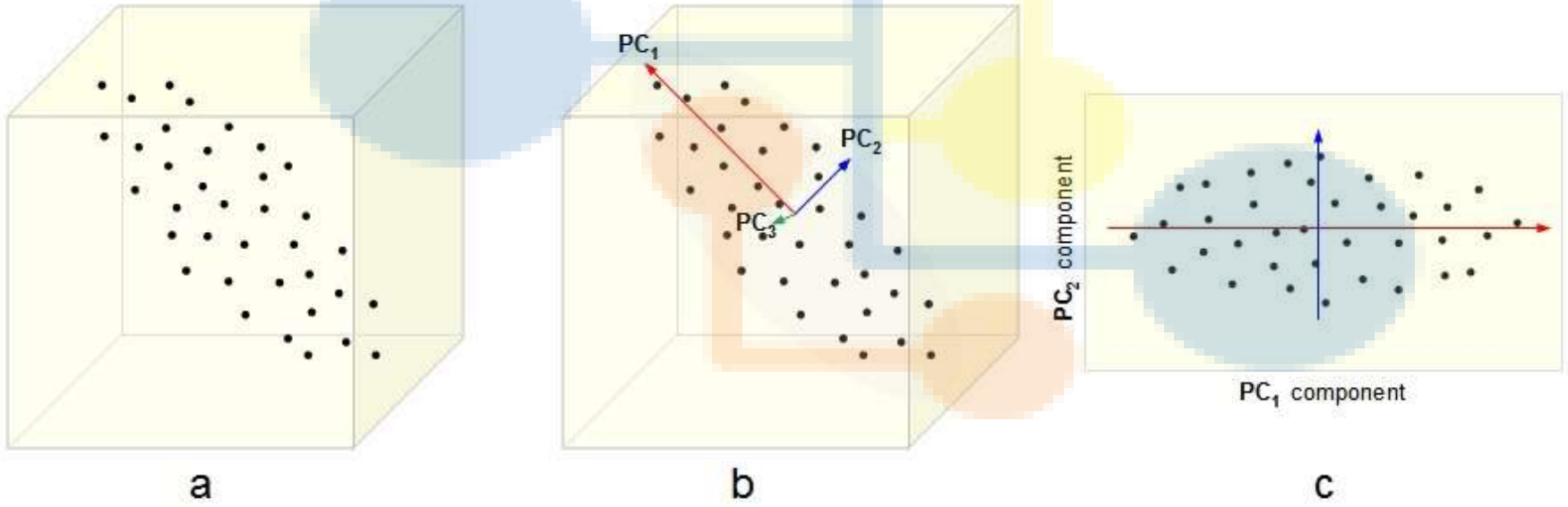
Principais Algoritmos de Machine Learning



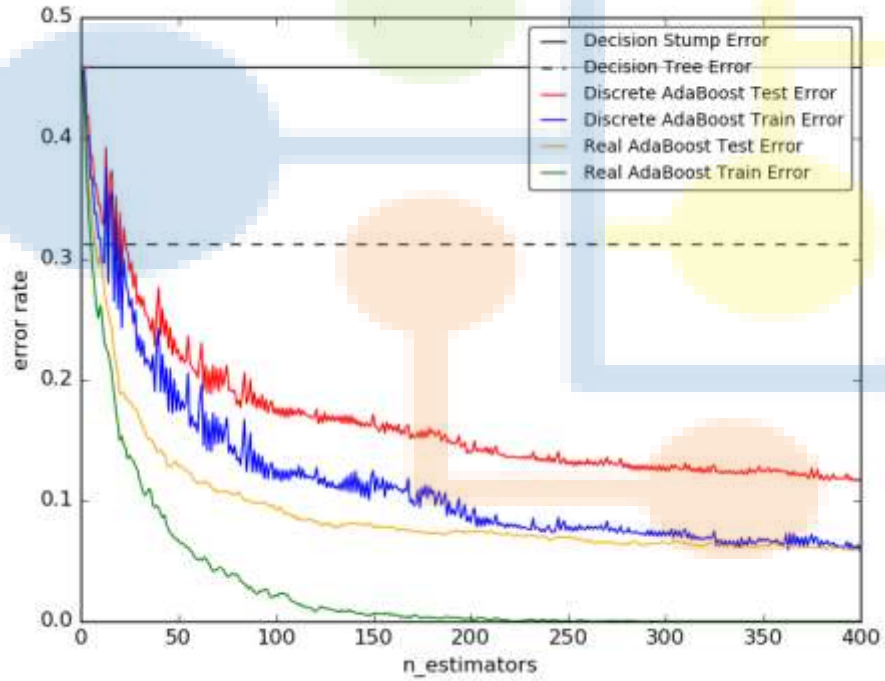
K-Means

Principais Algoritmos de Machine Learning

Algoritmo para Redução de Dimensionalidade



Gradient Boosting & AdaBoost

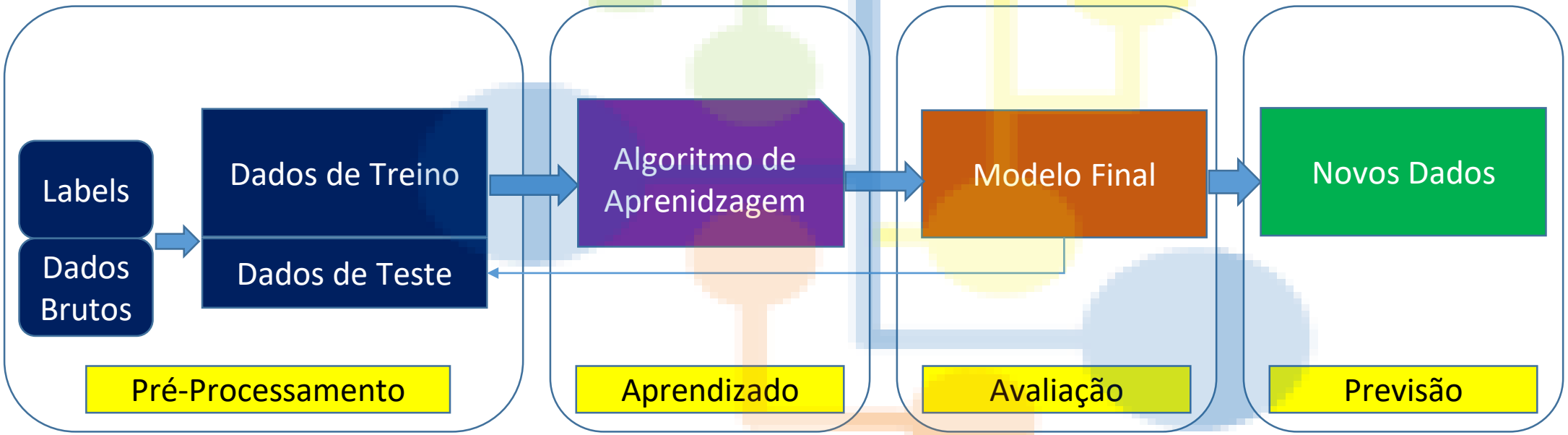


O Processo de Construção de Modelos de Machine Learning

O Processo de Construção de Modelos de Machine Learning

- Transformação de Variáveis
- Feature Selection
- Redução de Dimensionalidade
- Amostragem

- Validação do Modelo
- Otimização



- Seleção do Modelo
- Cross-Validation
- Métricas de Performance
- Otimização

Soluções de Machine Learning

Soluções de Machine Learning

Podemos construir modelos de Machine Learning de duas formas principais:

Desenvolvendo todo o algoritmo a partir do zero usando uma linguagem de programação.

Utilizando um framework pronto, onde os principais algoritmos já estão implementados.

Soluções de Machine Learning

Principais linguagens de programação para ML:

- Python
- Linguagem R
- Scala
- Java
- JavaScript
- Go
- C++ / C#

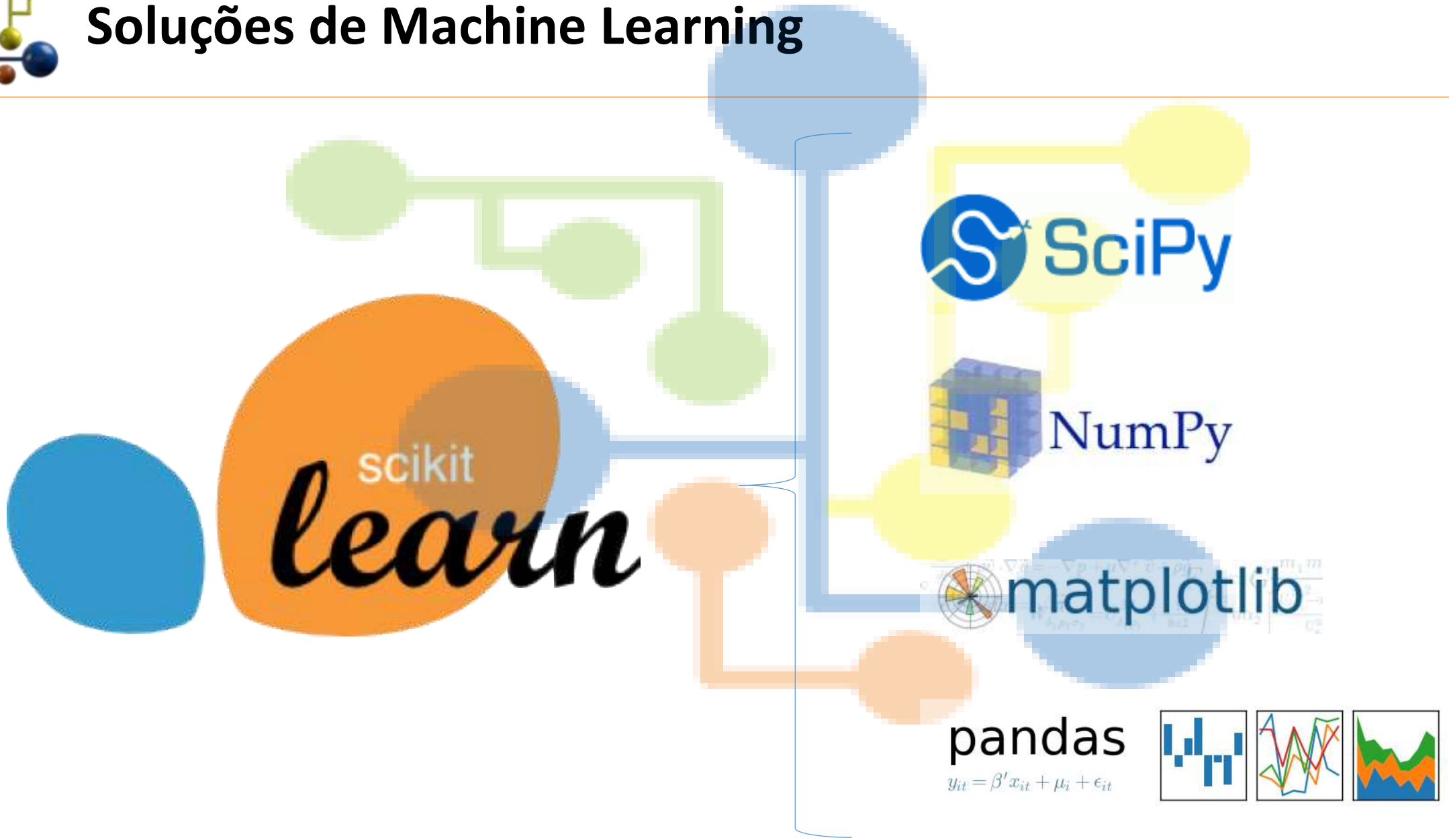
Principais frameworks para ML:

- Scikit-learn (Python)
- Caret (R)
- TensorFlow (Python, R, Java, C++)
- Apache Mahout (Python, Java)
- Spark Mllib (Scala, Java, Python, R)
- H2O (Java, Python)
- Weka (Java, Python)
- PyTorch, CNTK, MXNet (Python, C++, Java)

Soluções de Machine Learning

A linguagem Python oferece duas vantagens principais sobre todas as outras soluções. Primeiro, por se tratar de uma linguagem de uso geral, ela pode ser usada com qualquer uma destas soluções. Segundo, Python possui uma das mais poderosas soluções gratuitas de Machine Learning, o Scikit-learn, que estudaremos neste capítulo.

Soluções de Machine Learning



Soluções de Machine Learning

- Excelente documentação
- Fácil e intuitivo
- Diversos datasets
- Licença BSD (permite ser usado para fins comerciais)
- Confiável

Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy