Curso de Data Science

Charles Adriano dos Santos
Rafael Roberto Dias

ald
eia

Manhã

Horário	Assunto
09:30	Analisando Qualidade dos Dados
11:00	Variáveis Relevantes / Extração de Características
12:30	Almoço

Tarde

Horário	Assunto
13:30	Welcome to Python!
16:00	Agro XP – A Solução
17:30	Próximos Passos

Nos Episódios Anteriores...

Profissão Data Science

Estatística & Ciência da Computação

Desafio Agro XP

- Kanban
- Repositório
- Modelagem de Dados
- ETL
- Banco de Dados
- Namorando Dados SQL
- Linguagem R / R Studio

Pauta

1 – Analisando Qualidade dos Dados

2 – Variáveis Relevantes

Analisando a Qualidade dos Dados

- Objetivo nesta etapa do estudo é verificar a qualidade dos dados para entender quais tem potencial de fazer parte do estudo

- Foco maior em verificar se existem dados faltantes ou nulos que podem interferir no estudo

- Também aqui começa o entendimento de como cada variável ajuda a explicar o evento em estudo

- Aqui começam as **descobertas** do Cientista de Dados

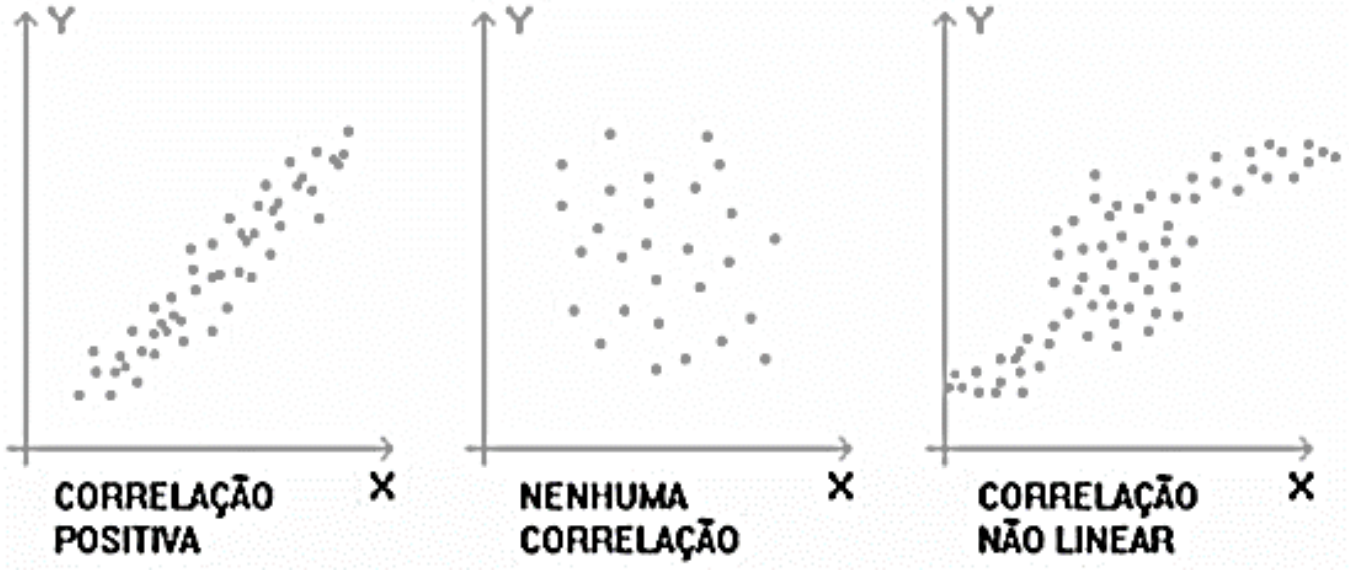
Pauta

1 – Analisando Qualidade dos Dados

2 – Variáveis Relevantes

Variáveis Relevantes

- Objetivo nesta etapa do estudo é verificar a como as variáveis se relacionam entre si
  - **Foco maior aqui é entender a correlação entre as variáveis**
- O modelo ou a metodologia que será utilizada para responder as perguntas do estudo dependem dos achados desta etapa



Pauta

1 – Welcome To Python!

2 – Agro XP Brazil - Solução

3 – Próximos Passos

Pauta

1 – Welcome To Python!

2 – Agro XP Brazil - Solução

3 – Próximos Passos

Python

Mar 2019	Mar 2018	Change	Programming Language	Ratings	Change
1	1		Java	14.880%	-0.06%
2	2		C	13.305%	+0.55%
3	4	↑	Python	8.262%	+2.39%
4	3	↓	C++	8.126%	+1.67%
5	6	↑	Visual Basic .NET	6.429%	+2.34%
6	5	↓	C#	3.267%	-1.80%
7	8	↑	JavaScript	2.426%	-1.49%
8	7	↓	PHP	2.420%	-1.59%
9	10	↑	SQL	1.926%	-0.76%
10	14	⇈	Objective-C	1.681%	-0.09%
11	18	⇈	MATLAB	1.469%	+0.06%
12	16	⇈	Assembly language	1.413%	-0.29%
13	11	↓	Perl	1.302%	-0.93%
14	20	⇈	R	1.278%	+0.15%
15	9	⇊	Ruby	1.202%	-1.54%
16	60	⇈	Groovy	1.178%	+1.04%
17	12	⇊	Swift	1.158%	-0.99%
18	17	↓	Go	1.016%	-0.43%
19	13	⇊	Delphi/Object Pascal	1.012%	-0.78%
20	15	⇊	Visual Basic	0.954%	-0.79%

Fonte: https://www.tiobe.com/tiobe-index/

Python – Me Dê Motivos

Linguagem em forte ascenção (3ª linguagem mais amada Stack Overflow)

Curva de Aprendizado Baixa

Free (Licença GLP)

Estável (1ª versão 1991)

Multiplataforma (Windows, Linux, MacOS e etc.)

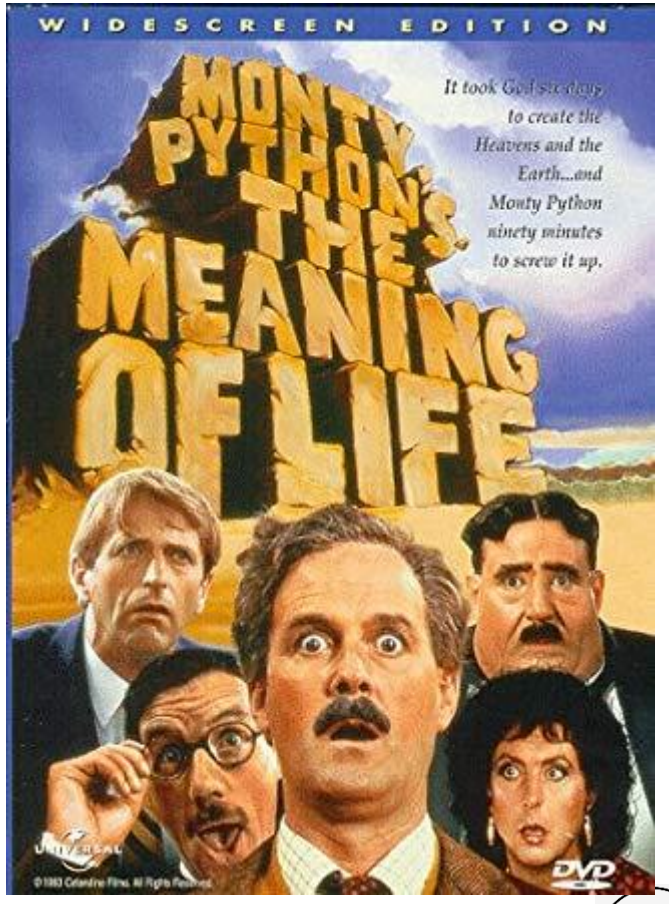
Comunidade

Data Science → Ótimos pacotes

Python – História

Pai do Python →

Guido van Rossum



A inspiração do nome →

Python – História

Versão 2 (2.7) **x Versão 3 (3.5)**

3/4 Paradigmas de Programação:

- **Programação Imperativa** → Ações/Comandos de um programa
- **Programação Orientada o Objeto** → Abstração, Encapsulamento, Herança e Polimorfismo
- **Programação Funcional** → Soluções como problemas de funções

Interpretada

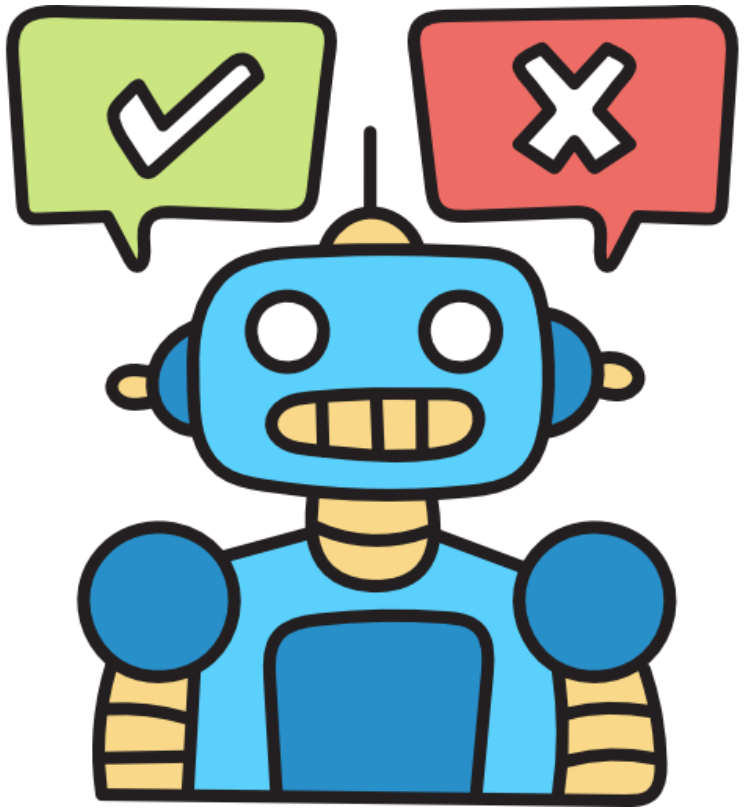
Python – Hands-on

Python – Versão 2 x Versão 3

Python 2.X	Python 3.X
There's ASCII str type and unicodetype, but no separate type to handle bytes of data	All strings (str) are Unicode strings; two byte classes are introduced: bytes and bytearray
Two types of integers: C-based integers (int) and Python long integer (long)	All integers are long but referred to by the int type
Return type of division is int if operands are integers: 5 / 4 gives 1; 4 / 2 gives 2	Return type of division is float even if operands or result are integers: 5 / 4 gives 1.25; 4 / 2 gives 2.0
round(16.5) returns a float of value 16.0	round(16.5) returns an int of value 16
Unorderable types can be compared	Comparison of unorderable types raises a TypeError
print is a statement: print "Hello World!"	print() is a built-in function: print("Hello World!")
range() returns a list of numbers while xrange() returns an object for lazy evaluation	range() returns an object for lazy evaluation similar to Python 2 xrange(); and range()method __contains__ speeds up lookups
Functions/methods map(), filter(), zip(), dict.items(), dict. keys(), dict.values() return lists	These function/methods return objects for lazy evaluation
raw_input() returns input as strand input() evaluates the input as a Python expression	input() will return a string similar to Python 2 raw_input()
Raising exceptions: raise IOError("file error") or raise IOError, "file error"	Raising exceptions: raise IOError("file error")
Handling exceptions: except NameError, err: or except (TypeError, NameError), err:	Handling exceptions: except NameError as err or except (TypeError, NameError) as err
On generators, a method or function call: g.next() or next(g)	On generators, only a function call: next(g)
Loop variables in a comprehension leak to global namespace	Loop variables are limited in scope to the comprehension

Fonte: https://devopedia.org/python-2-vs-3

Machine Learning - Conceito

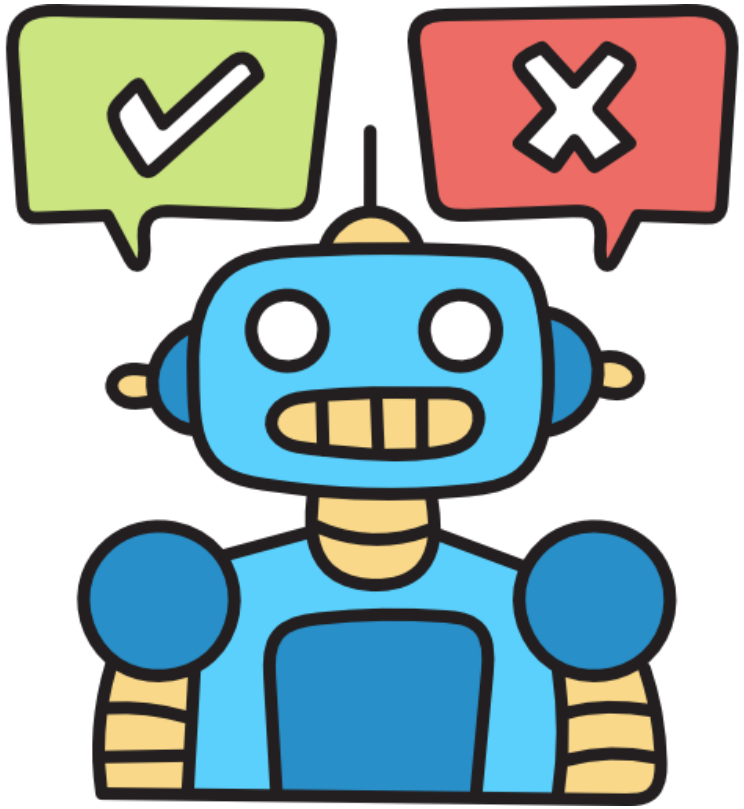


A máquina, através de algoritmos, obter padrões sobre características extraídas dos dados para, com um modelo gerado/criados, classificar as observações futuras de novos dados.

No conceito cada vez menos intervenção humana (conceito).

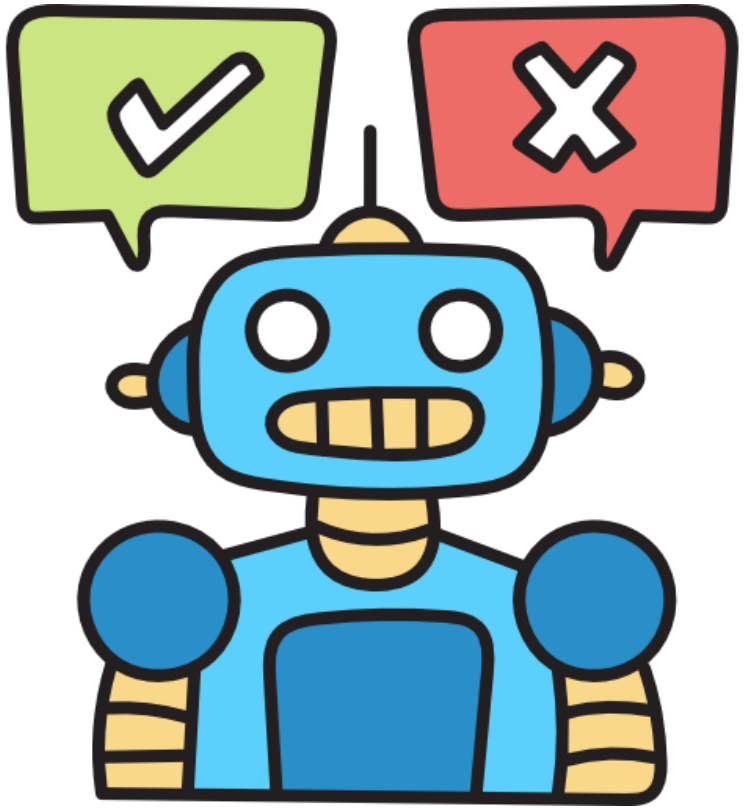
Pré-processamento e análise dos dados, além de realizar “grid” de valores para treinamento obterem maior acurácia
(na prática)

Machine Learning - História



1950 - IA: Computadores com habilidade de “pensar” -Teste de Turing. Em 2014 chatbot enganou 10/30 juízes

Machine Learning - História

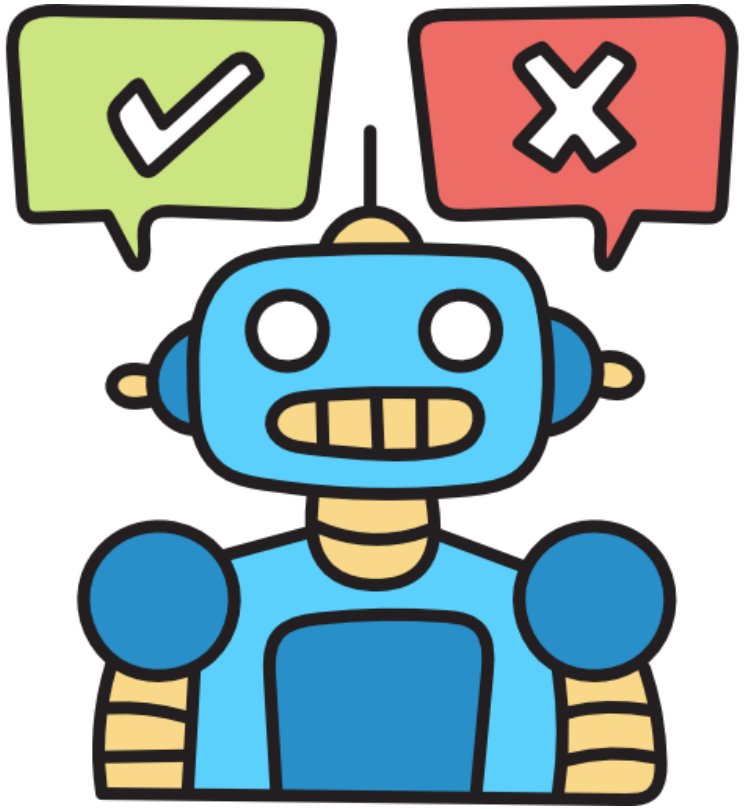


1959 - ML: Aprender a partir dos dados - Arthur Samuel

Aprender com a experiência que existe intrínseca aos dados.

Algoritmos de aprendizado de máquina analisam as correlações entre os atributos (variáveis) de um sistema (base de dados) a partir de dados amostrais (base de treinamento)

Machine Learning - História



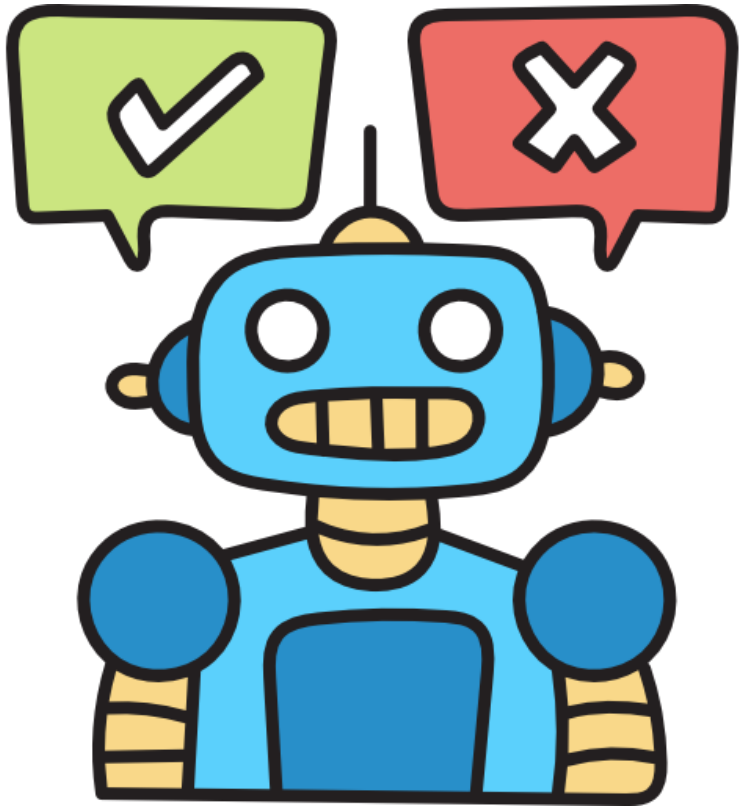
2012: DS – Entender os Dados

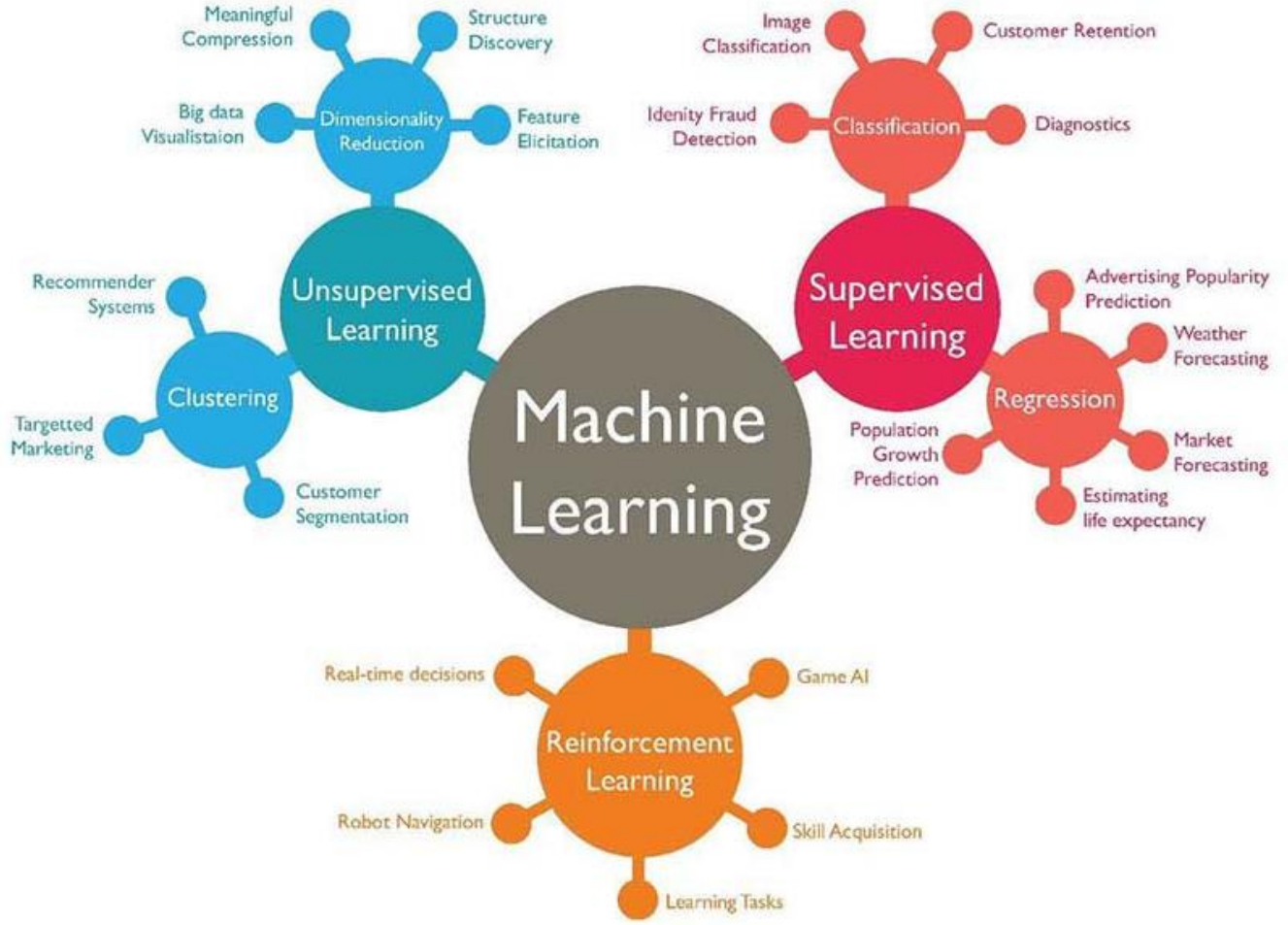
Ciência de dados utilizando probabilidade, estatística álgebra linear e computação.

Conhecimentos de IA e ML

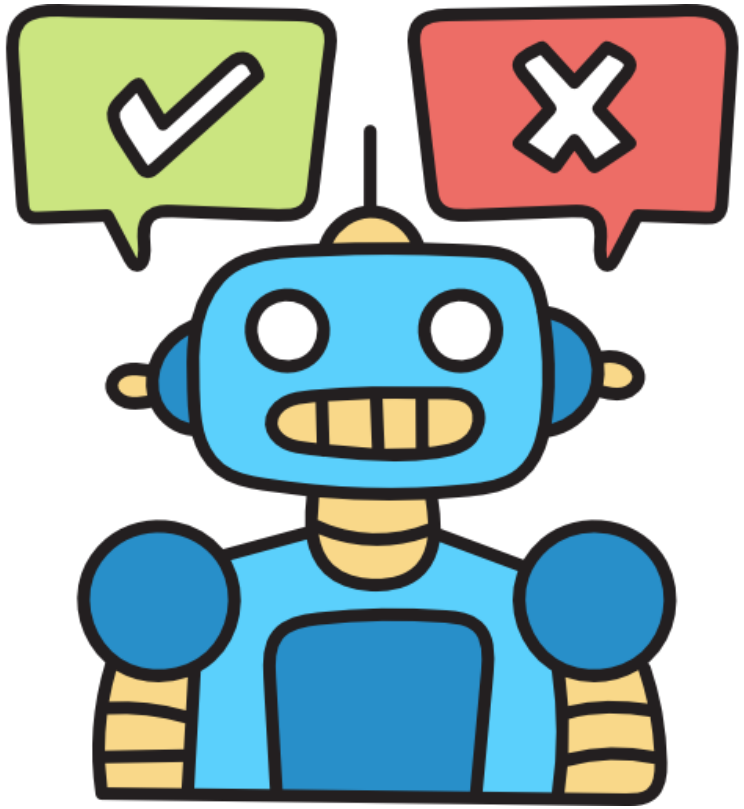
"É a ciência (e arte) de programar computadores de tal forma que eles aprendam a partir de dados"
(Aurélien Géron, 2017)

Machine Learning – Tipos de Aprendizado





Machine Learning – Tipo de Aprendizado

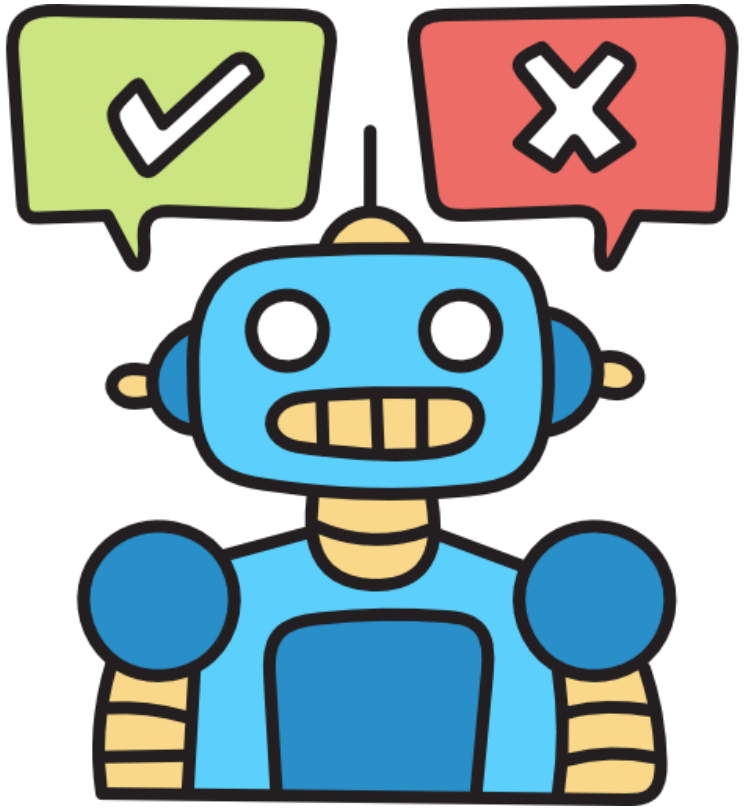


Supervisionado → rotulado com saídas esperadas. Modelo gera ao entrar com conjunto de características uma saída rotulada (Classificação) ou um valor futuro (Predição). Ex: Nosso desafio AgroXP.

Não Supervisionado → Não existe rótulo prévio. Analisa a rede de relacionamento entre os dados para agrupá-los por características similares. Ex: Categorização de Clientes

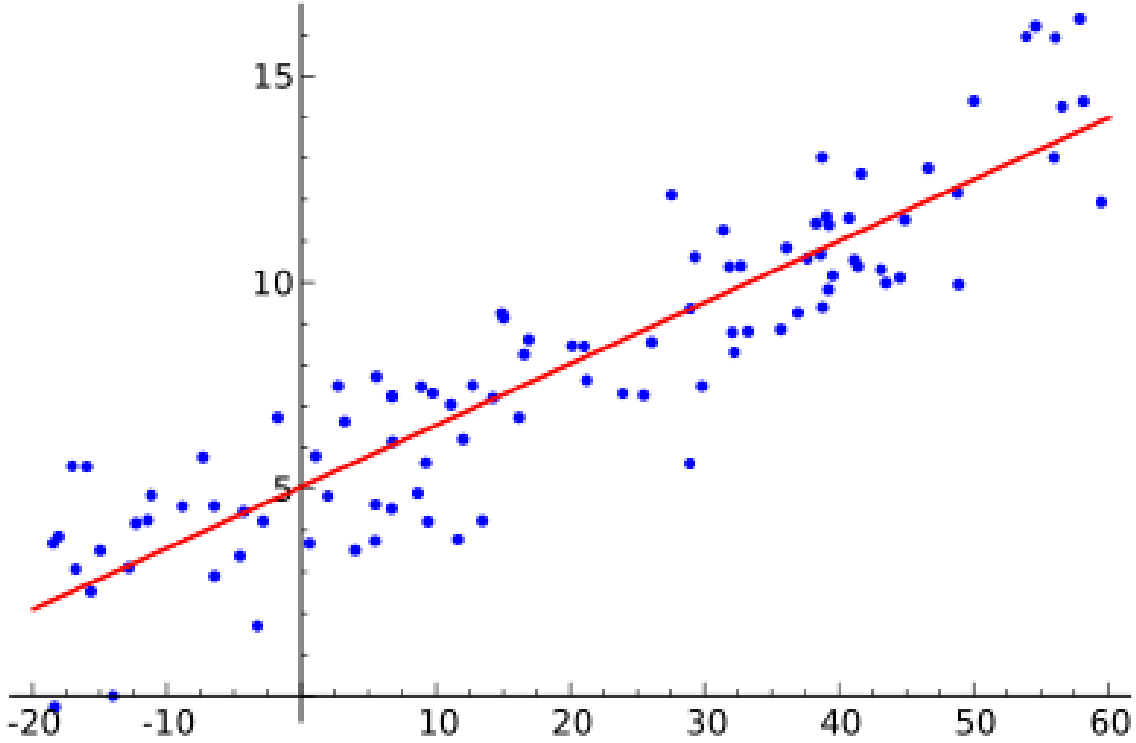
Reforço → Maximizar o resultado. Baseado em recompensa / punição. Com isso algoritmo encontrar a "política" que mapeia os dados. Ex: Personagens Jogos

Machine Learning – Exemplos Algoritmos

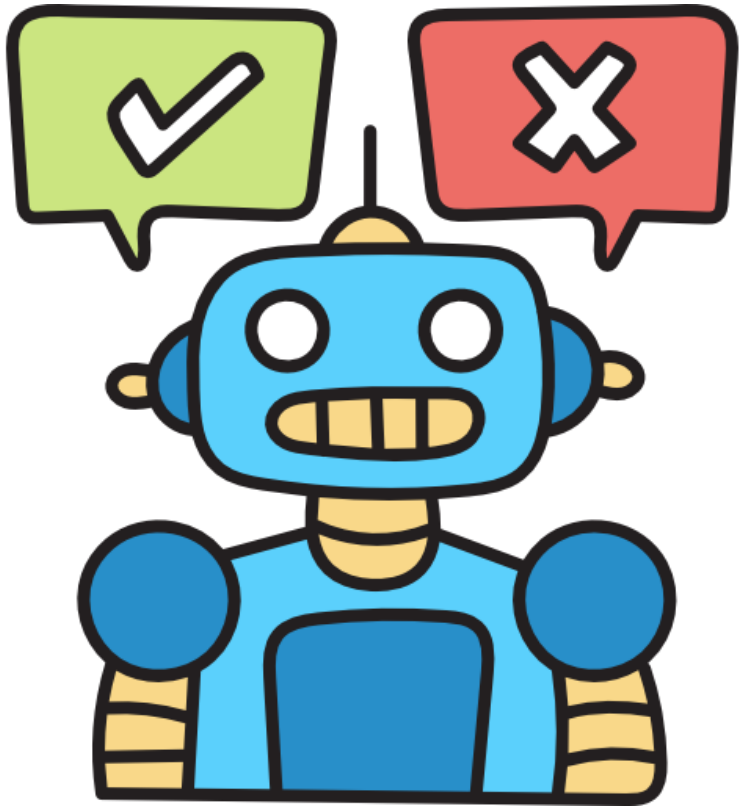


Regressão Linear (Supervisionado – Predição)

Simples... Busca uma reta para se ajustar aos dados.
Problemas de relação linear.

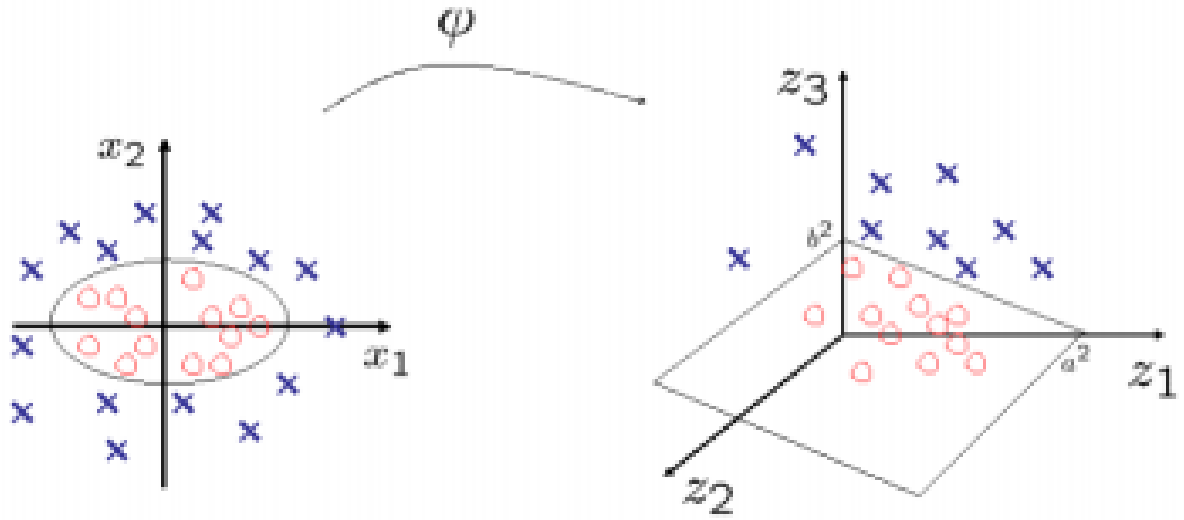


Machine Learning – Exemplos Algoritmos

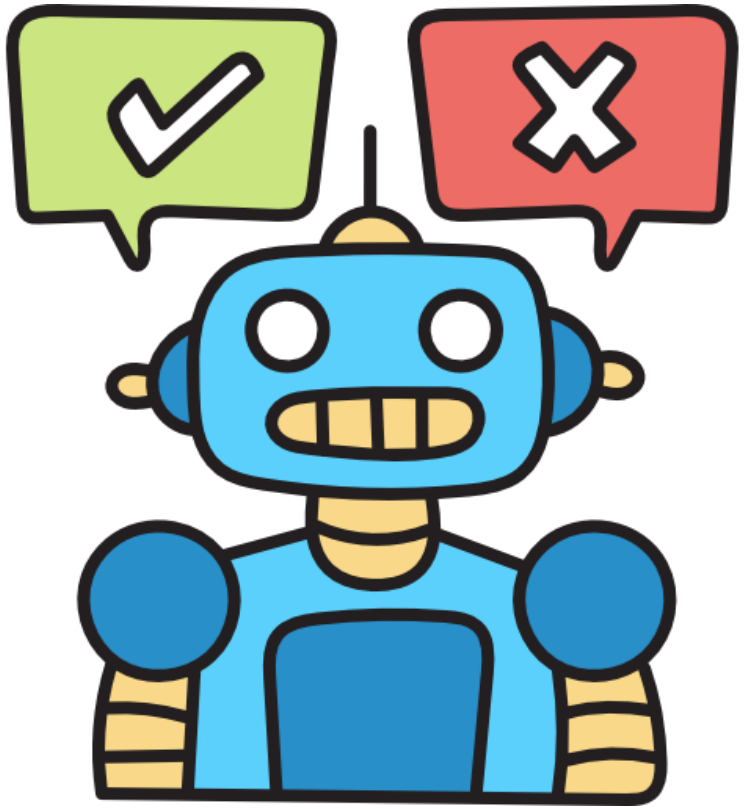


SVM - Support Vector Machine **(Supervisionado – Classificação) –** Vapnik (1963)

Distância das amostrar da linha superfície de separação. Consegue trabalhar com dados não lineares com a premissa de que em alguma dimensão os dados terão linearidade.

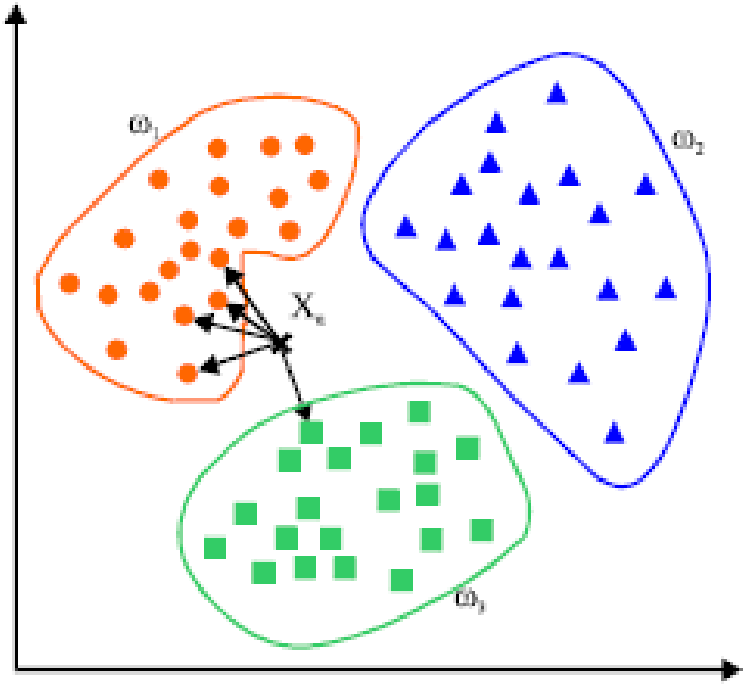


Machine Learning – Exemplos Algoritmos

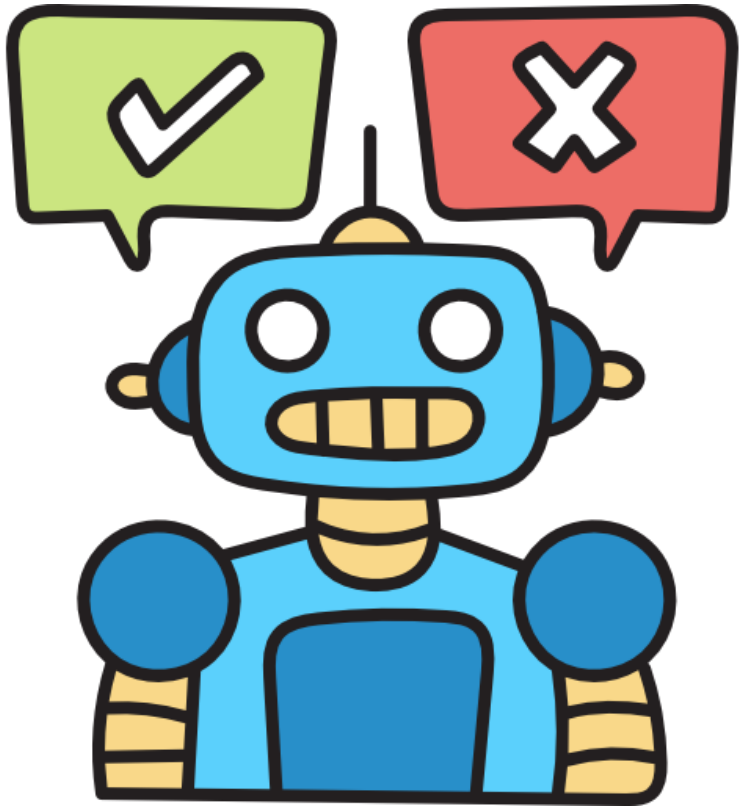


KNN – K-Nearest Neighbors (Supervisionado – Classificação)

Baseado em encontrar o valor de K que consiga através de funções básicas de distância Euclidiana encontrar a melhor superfície de separação

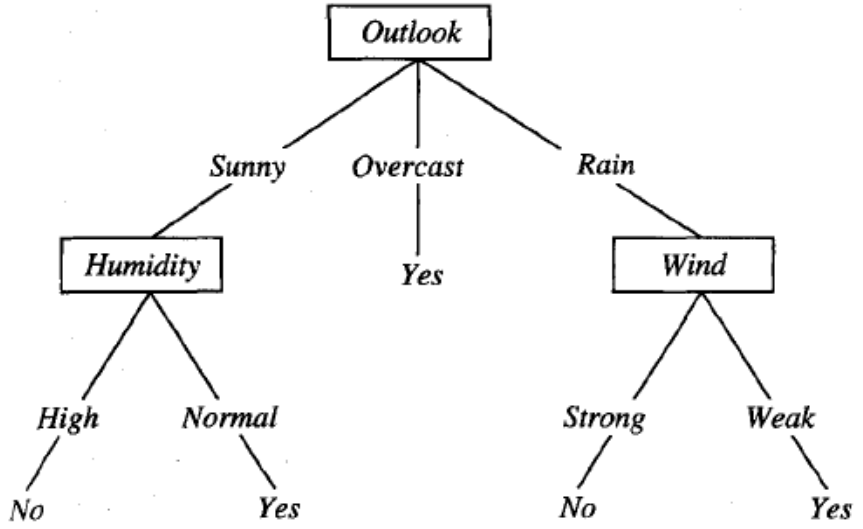


Machine Learning – Exemplos Algoritmos

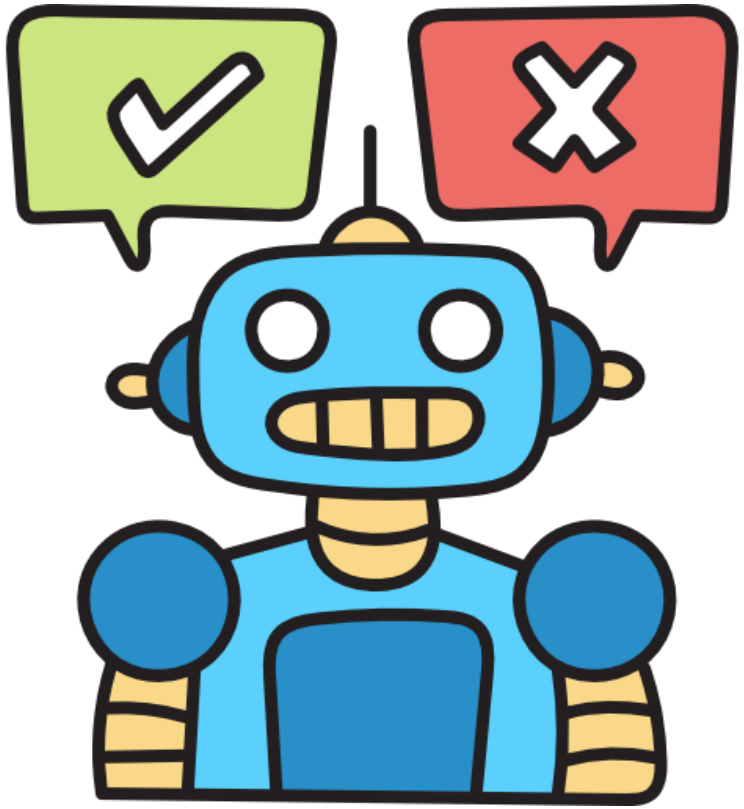


Árvore de Decisão (Supervisionado – Classificação)

De fácil explicação do modelo obtido, este algoritmo utiliza a categorização utilizando técnicas referente a Ganho de Informação dos atributos (o quanto a variável sozinha classifica os exemplos de treinamento). Pode ser utilizado para dados numérico ou simbólicos.

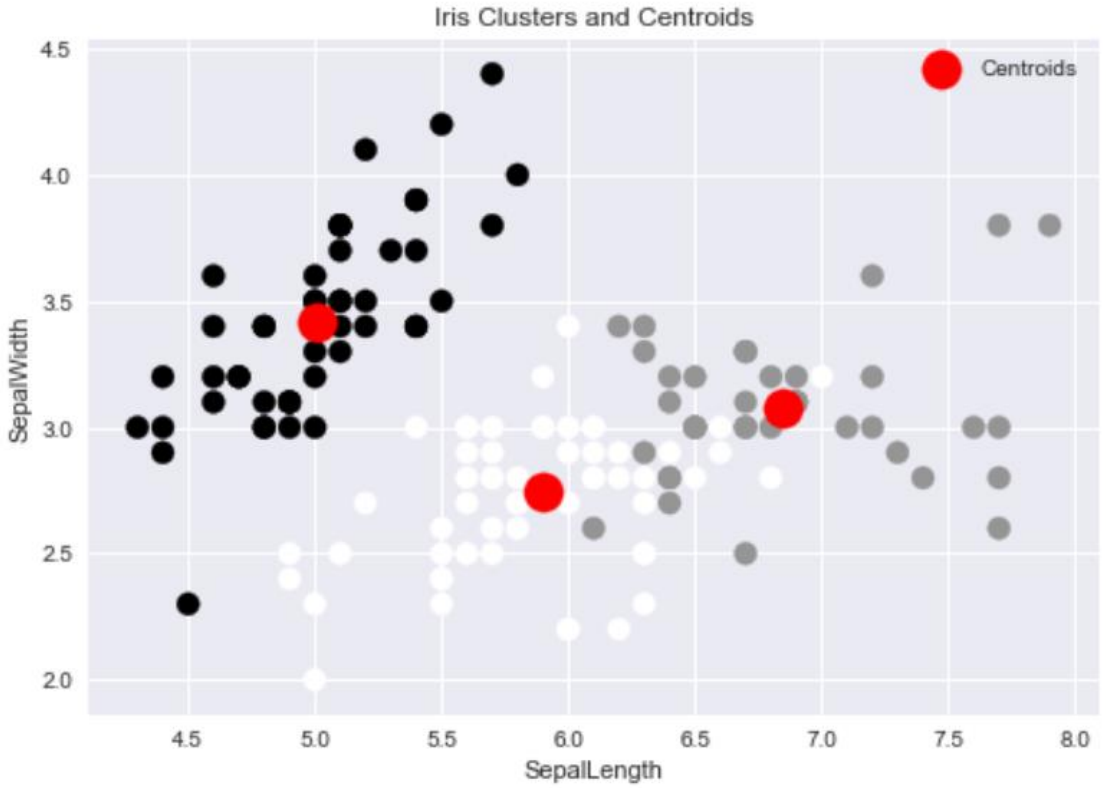


Machine Learning – Exemplos Algoritmos

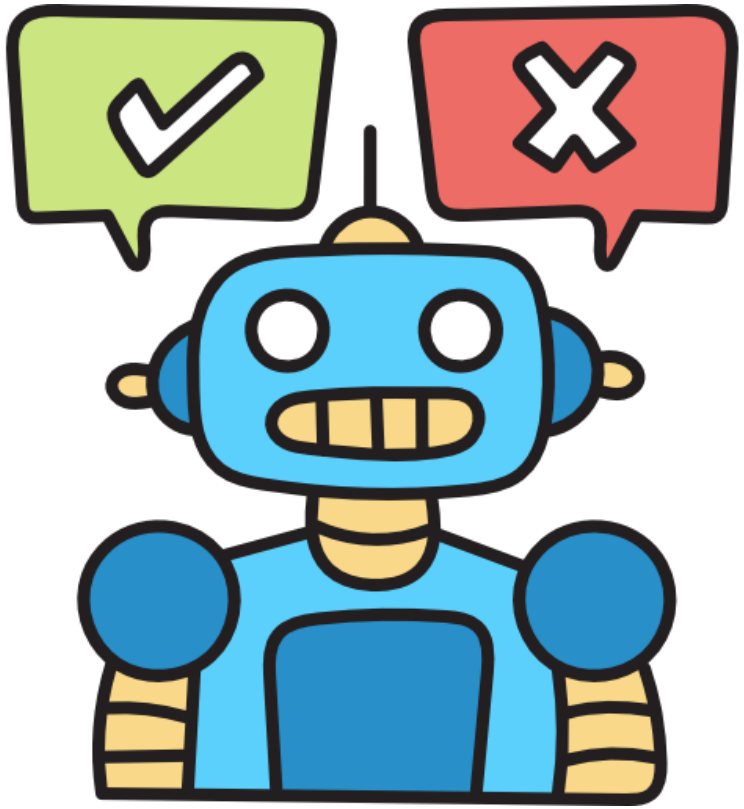


K-Means – (Não Supervisionado)

Forma clusters que contêm pontos homogêneos aos dados.

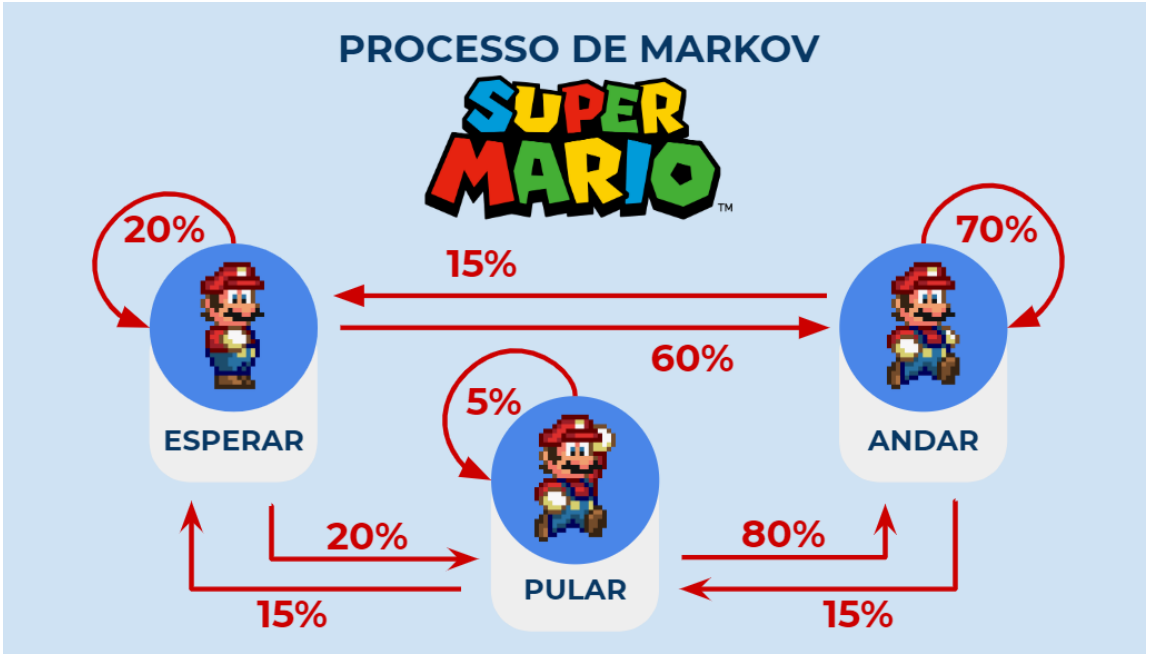


Machine Learning – Exemplos Algoritmos

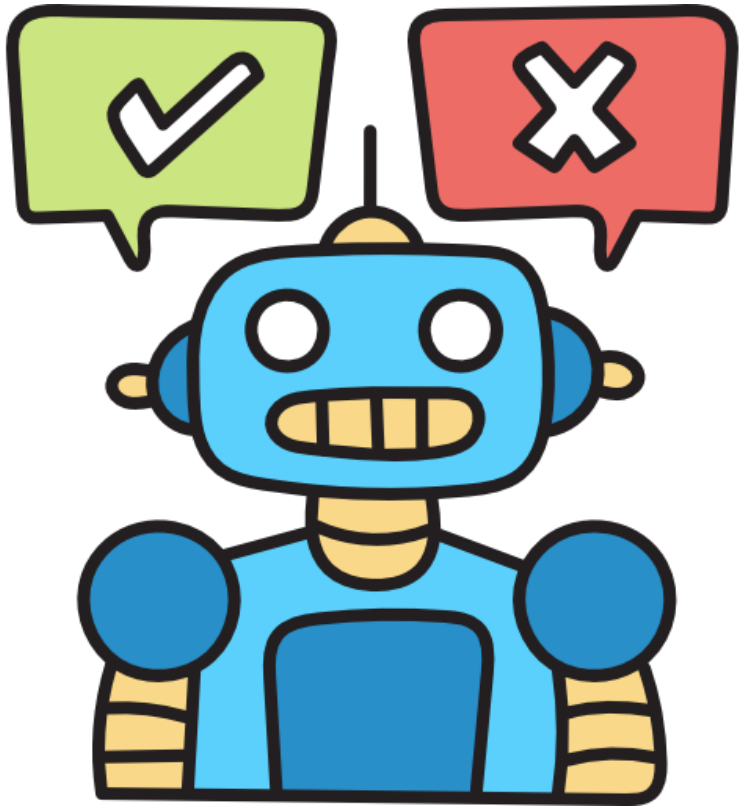


Cadeia de Markov (Reforço)

Processo estocástico (futuro ← estado atual). Com base na cadeia e suas probabilidades o algoritmo toma uma decisão e, se houver recompensa, reforça a decisão tomada. Se houver uma punição rechaça.

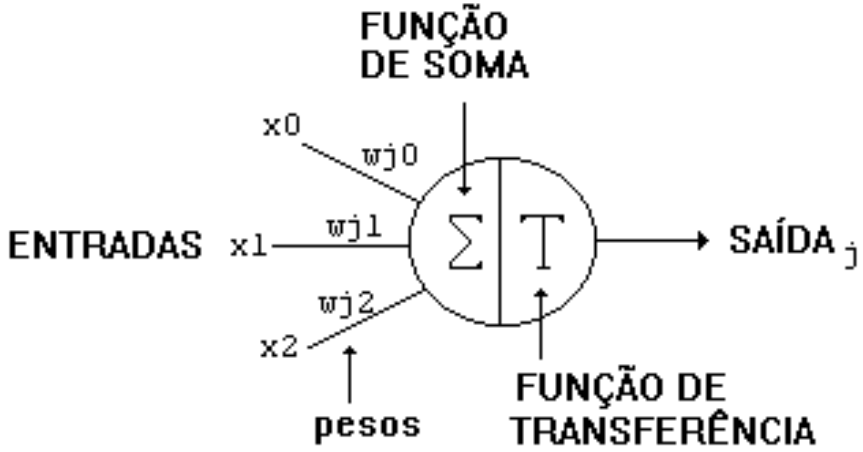


Machine Learning – Exemplos Algoritmos

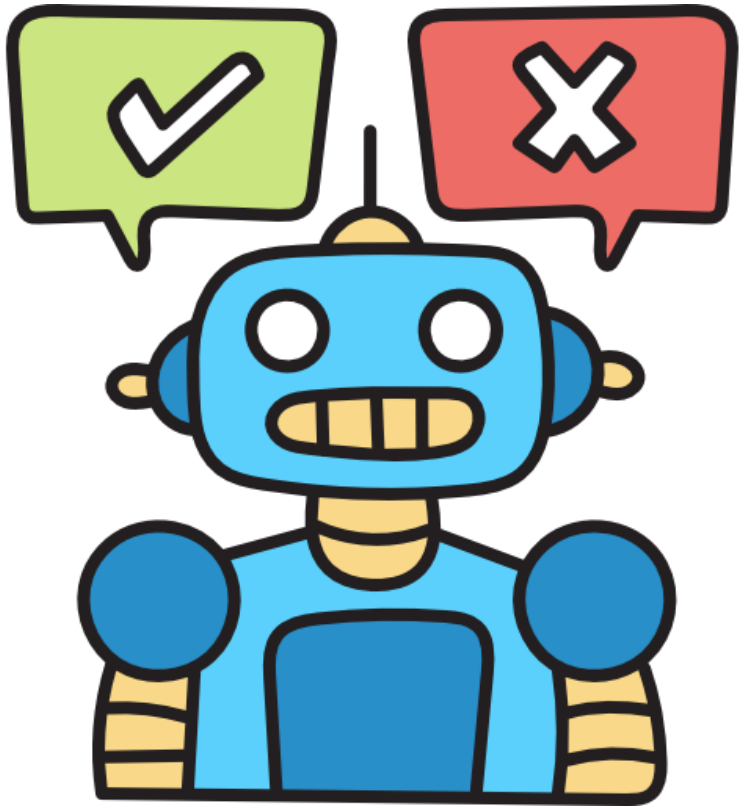


Redes Neurais (Supervisionado – Classificação)

Baseado no conceito matemático e computacional (1943) que visa descrever o modelo artificial para um neurônio biológico.
Responde “ligando/desligando”
os vários neurônios interligada e com isso classifica as características de entrada no rótulo predito pelo modelo.

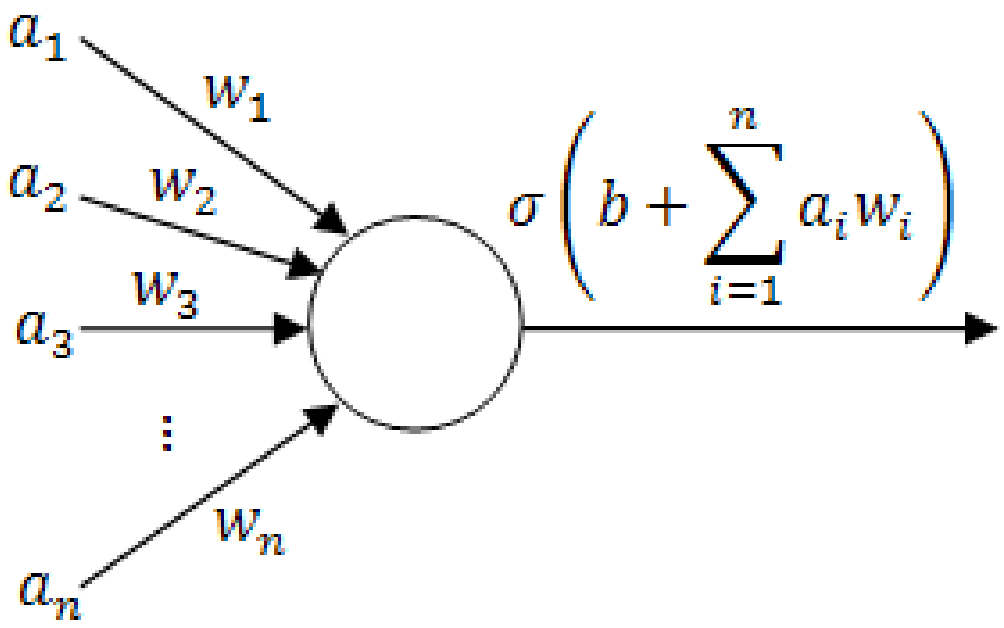


Machine Learning – Exemplos Algoritmos

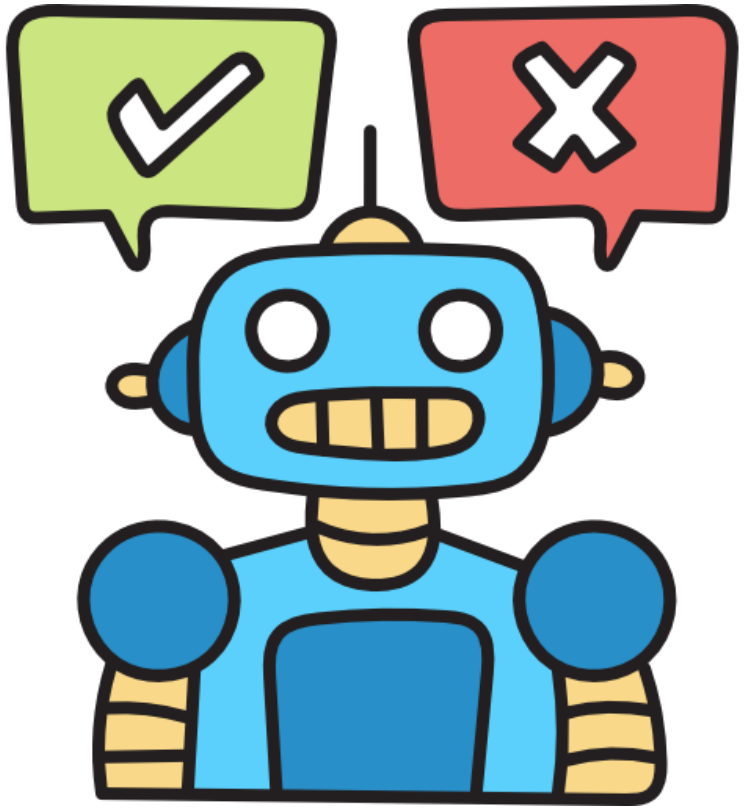


Redes Neurais (Supervisionado – Classificação)

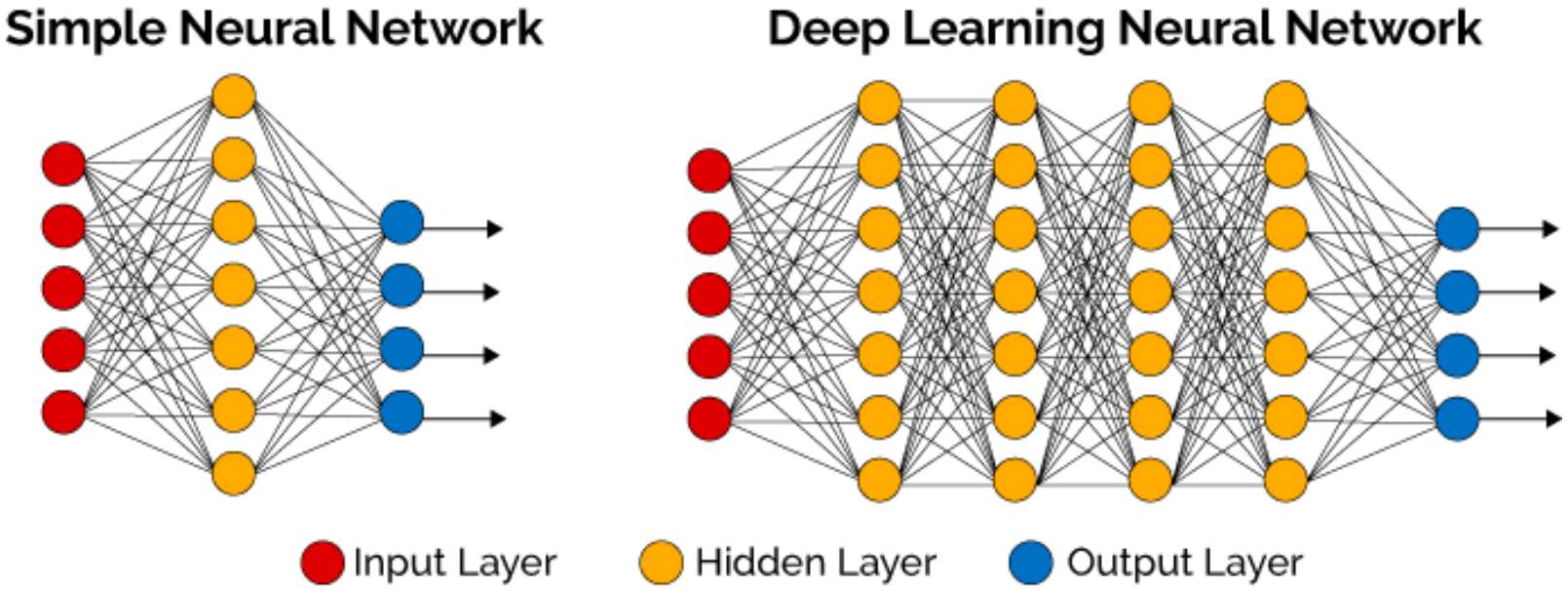
Perceptron → Tipo básico de rede neural. Demonstrou em 1957 a possibilidade de simulação de um neurônio biológico.



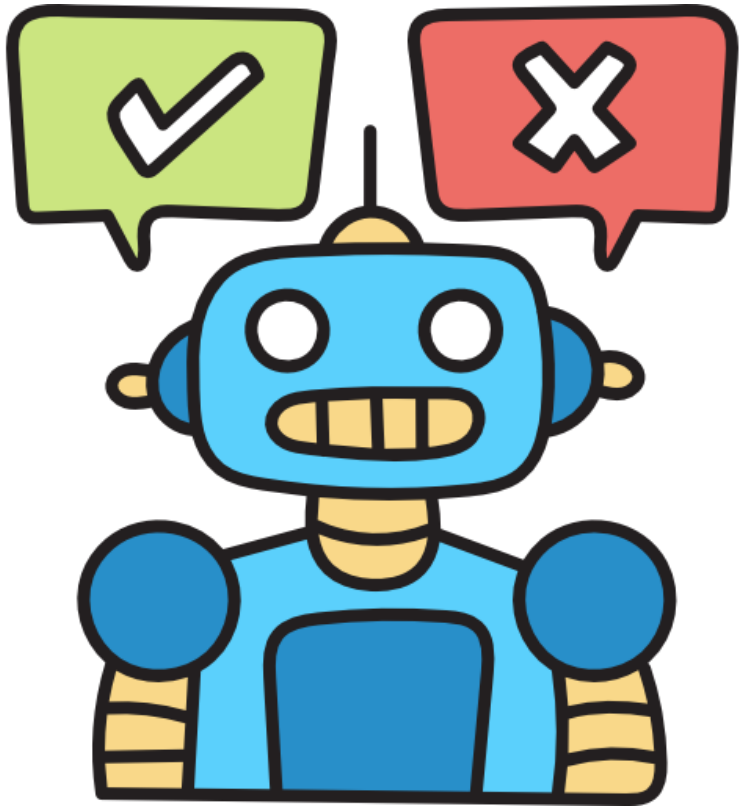
Machine Learning – Exemplos Algoritmos



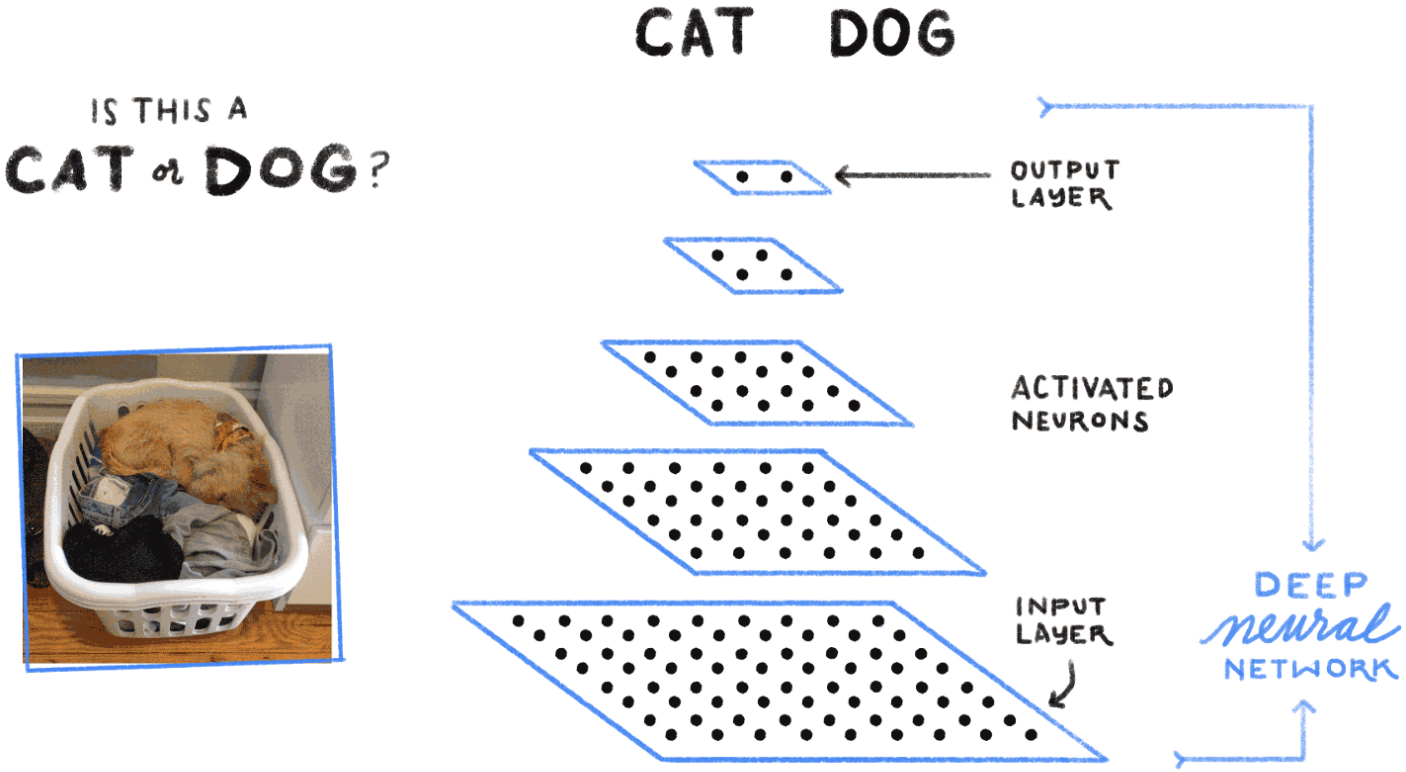
Redes Neurais (Supervisionado – Classificação)



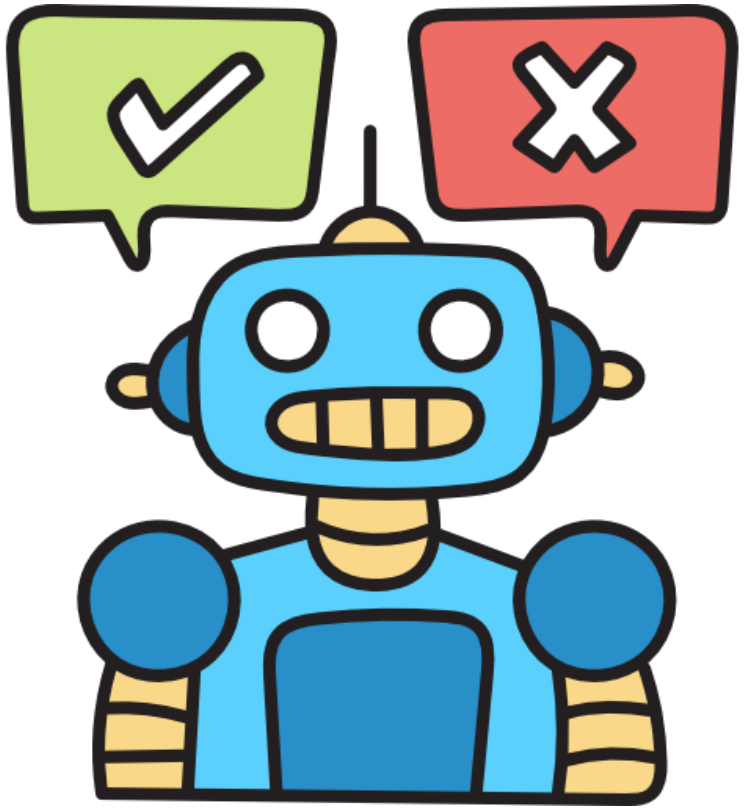
Machine Learning – Exemplos Algoritmos

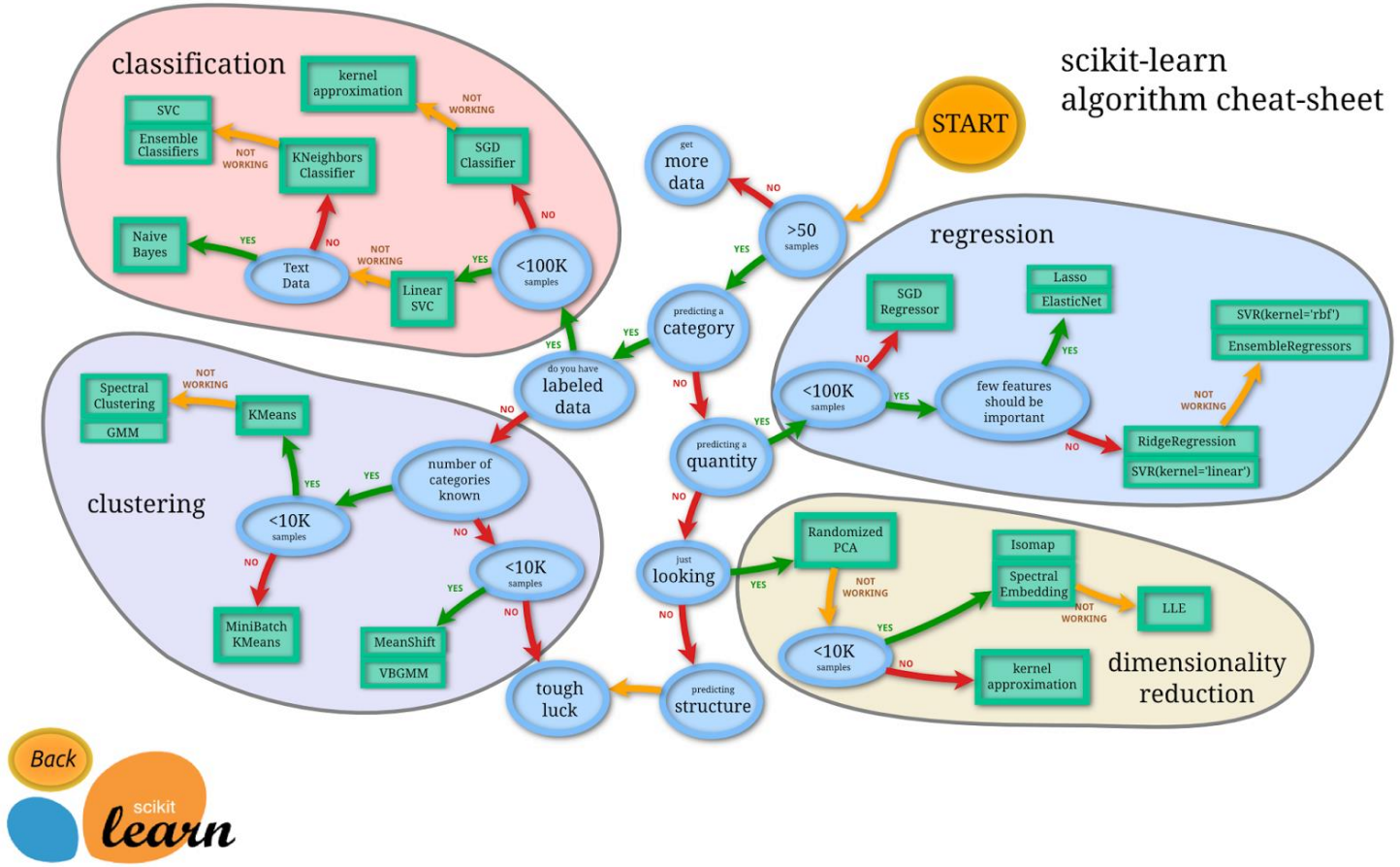


Redes Neurais (Supervisionado – Classificação)



Machine Learning – Algoritmo x Características Dados





Fonte: https://scikit-learn.org/stable/tutorial/machine_learning_map/

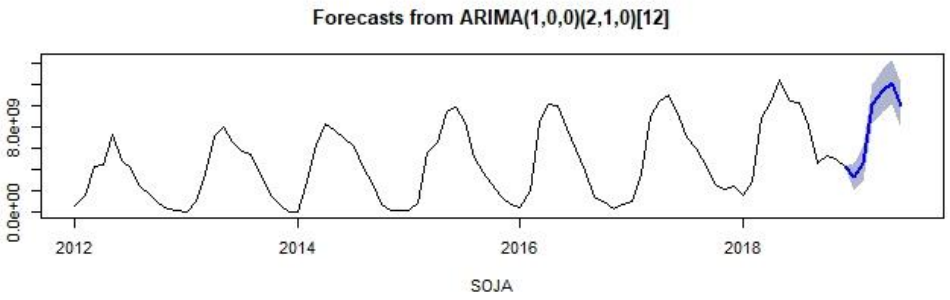
Pauta

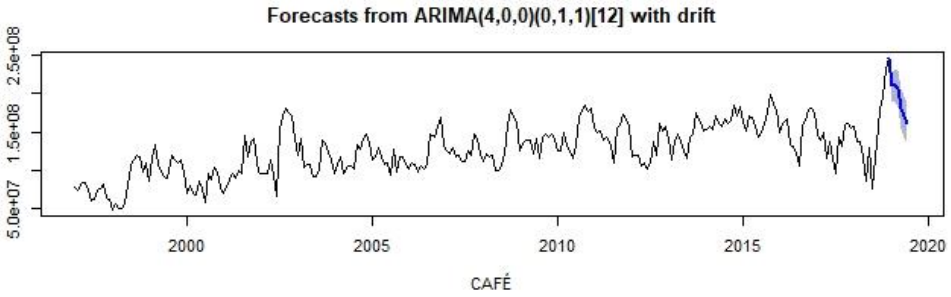
1 – Welcome To Python!

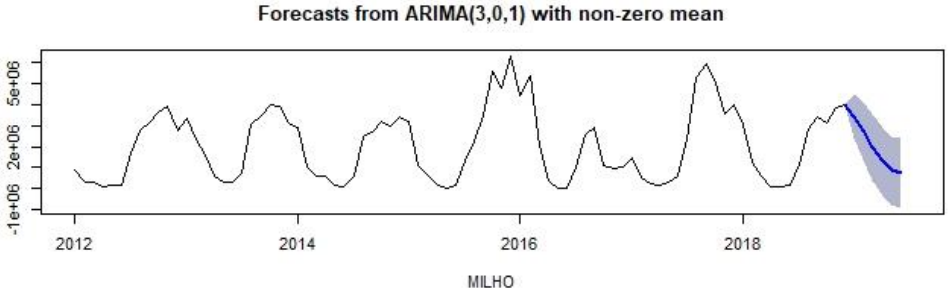
2 – Agro XP Brazil - Solução

3 – Próximos Passos

Agro XP Brazil - Solução







- **Proposta**: Verificar qual é a previsão para os próximos 4 meses para cada um dos grãos
- E decidir em qual commodities iremos investir no 1º semestre/2019
- Utilizaremos técnicas de **Séries Temporais**

Pauta

1 – Welcome To Python!

2 – Agro XP Brazil - Solução

3 – Próximos Passos

Obrigado !

Charles Adriano dos Santos
charles.a.santos@caelis.it
chadri
41 99144 6663

Rafael Roberto Dias
rafael.dias@madeiramadeira.com.br
rafael-roberto-dias-00b39123
41 99672 7170