



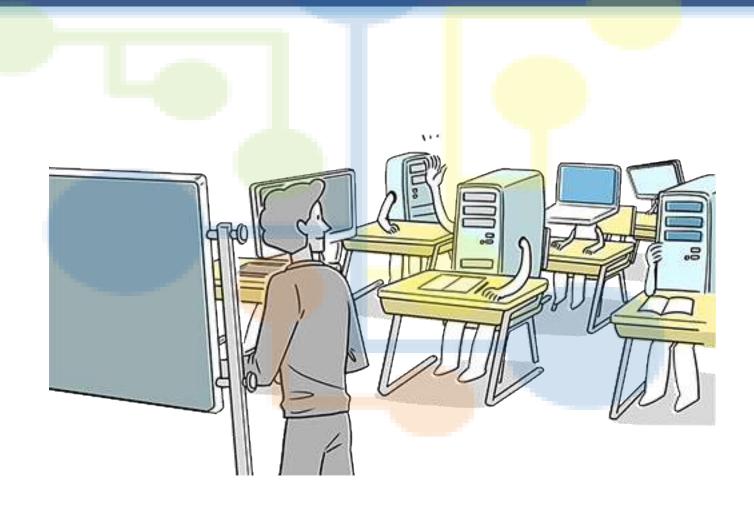
# Big Data Analytics com R e Microsoft Azure Machine Learning

# Machine Learning em Linguagem R

Seja Bem-Vindo(a)!











#### O que veremos neste capítulo:

- Definição de Machine Learning
- Frameworks de Machine Learning
- Processo de Aprendizagem
- Treinamento, Validação e Teste
- Modelos Preditivos
- Algoritmos de Machine Learning
- Regressão e Classificação Através de Projetos
- Lista de Exercícios com a Construção de Modelos



# Projetos Inteiros de Regressão e Classificação

Prevendo despesas hospitalares Prevendo a ocorrência de câncer



Embora tenhamos aqui uma grande quantidade de conteúdo, Machine Learning ainda será estudado em muito mais detalhes nos demais cursos da Formação Cientista de Dados.





## Big Data Analytics com R e Microsoft Azure Machine Learning

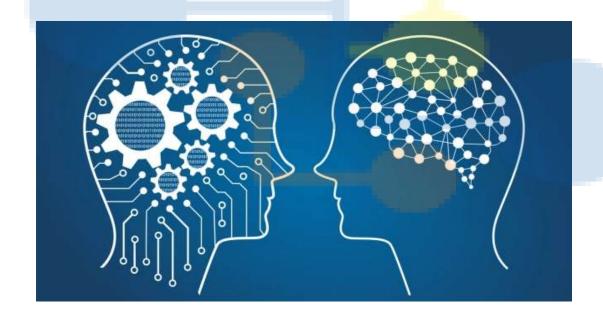
# Introdução ao Aprendizado de Máquina (Machine Learning)

Seja Bem-Vindo(a)!





O termo Machine Learning (ou aprendizado de máquina em português) possui atualmente as mais variadas definições, especialmente depois de tantos filmes sobre robôs e Inteligência Artificial, que transformaram Machine Learning em algo que realmente não é.





O que é Machine Learning? (Aprendizado de Máquina)



# Machine Learning é o método de análise de dados que automatiza a construção de modelos analíticos.





E como as máquinas aprendem?







# Machine Learning pode realizar análises preditivas mais rápido que qualquer humano seria capaz de fazer!

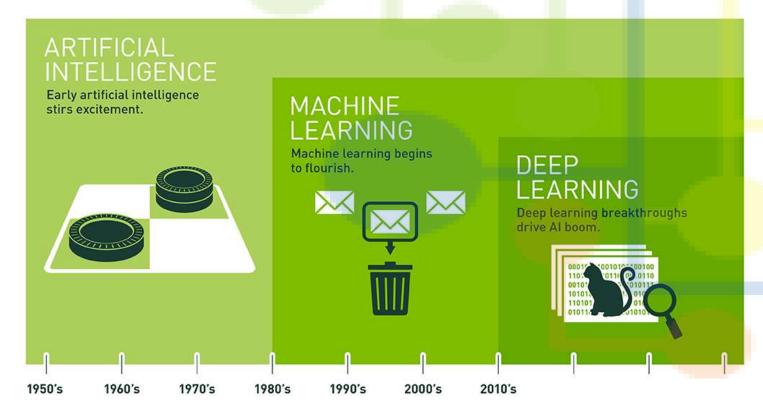




# Então Machine Learning e IA são conceitos diferentes?







Machine Learning é um subconjunto da Inteligência Artificial

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.



# Inteligência Artificial inclui Machine Learning, mas Machine Learning por si só não define Inteligência Artificial.



Inteligência Artificial é baseada em Machine Learning e Machine Learning é essencialmente diferente de Estatística.





# Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e Introdução ao Aprendizado de Máquina (Machine Learning)

Técnica	Estatística	Machine Learning
Entrada de Dados	Os parâmetros interpretam fenômenos da vida real e trabalham a magnitude.	Os dados são randomizados e transformados para aumentar a acurácia de análises preditivas.
Tratamento de Dados	Modelos são usados para previsões em amostras pequenas.	Trabalha com Big Data na forma de redes e grafos. Os dados são divididos em dados de treino e dados de teste.
Resultado	Captura a variabilidade e a incerteza dos parâmetros.	Probabilidade é usada para comparações e para buscar as melhores decisões.
Distribuição dos Dados	Assumimos uma distribuiç <mark>ão bem definida</mark> dos dados.	A distribuição dos dados é desconhecida ou ignorada antes do processo de aprendizagem.
Objetivos	Assumimos um determinado resultado e então tentamos prová-lo.	Os algoritmos aprendem a partir dos dados.



# Machine Learning se baseia em alguns importantes conceitos da Matemática, Estatística e Ciência da Computação:

Manipulação de Matrizes

Teoria da Probabilidade e Inferência Estatística

Programação

Armazenamento e Processamento de Dados





# Big Data Analytics com R e Microsoft Azure Machine Learning

O Que São Algoritmos?

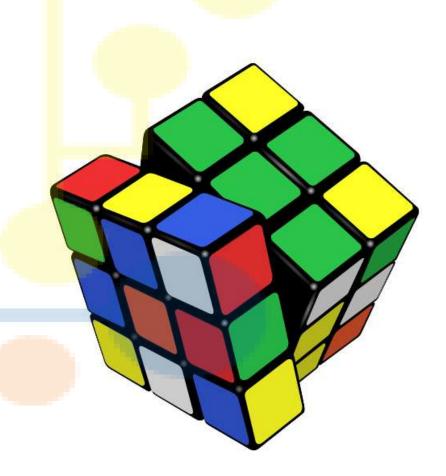
Seja Bem-Vindo(a)!



# Machine Learning usa algoritmos para analisar grandes conjuntos de dados!

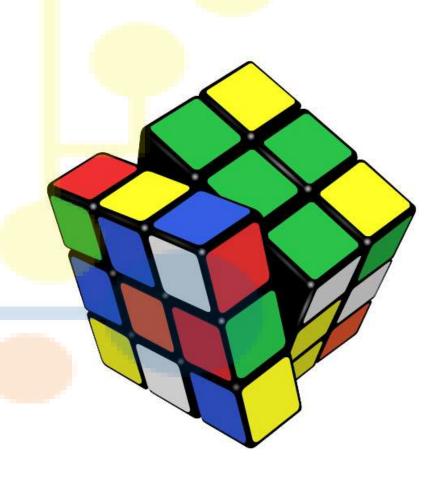


Ok entendi, mas o que são algoritmos?





Algoritmos são procedimentos ou fórmulas usados para resolver problemas.





Algoritmos são procedimentos ou fórmulas usados para resolver problemas.





O tipo de problema a ser resolvido, determina o tipo de algoritmo a ser utilizado.







#### Algoritmos - Exemplo

Algoritmo: Sacar dinheiro

#### INÍCIO

- 1. Ir até o caixa eletrônico.
- 2. Colocar o cartão.
- Digitar a senha.
- 4. Solicitar o saldo.
- Se o saldo for maior ou igual à quantia desejada, sacar a quantia desejada; caso contrário sacar o valor do saldo.
- Retirar dinheiro e cartão.

FIM.



# Falhas são mais comuns que sucesso em processos de Machine Learning.







# Big Data Analytics com R e Microsoft Azure Machine Learning

# Machine Learning Frameworks

Seja Bem-Vindo(a)!



# Para criar modelos de Machine Learning você tem duas opções:

Desenvolver os algoritmos a partir do zero

Utilizar Frameworks prontos



- Um framework é um conjunto de softwares que produzem um resultado específico. Um framework nos permite focar mais no problema de negócio e menos na parte de codificação.
- Frameworks de Machine Learning permitem que você trabalhe em um problema, sem ter que saber muito sobre programação (embora seja altamente recomendável que você conheça bem sobre programação).
- O framework cuida da gestão de infraestrutura, enquanto você pode focar mais na parte inteligente da sua aplicação.



# E por que usar Machine Learning Frameworks?



# Principais Machine Learning Frameworks





Linguagem R (Pacote caret)





# Microsoft Azure Machine Learning



Azure machine learning



Scikit-Learn (Linguagem Python)





Apache Spark MLlib





Google Tensor Flow







## Machine Learning Frameworks









Keras

Caffe

CNTK

**M**xnet





## Machine Learning Frameworks





## Big Data Analytics com R e Microsoft Azure Machine Learning

# Tipos de Aprendizagem em Machine Learning

Seja Bem-Vindo(a)!



O Processo de Aprendizagem ocorre de diferentes formas e podemos dividir os algoritmos de Machine Learning em 3 grupos principais:





Aprendizagem Supervisionada Aprendizagem Não Supervisionada

Aprendizagem
Por Reforço



Aprendizagem Supervisionada É o termo usado sempre que o algoritmo é "treinado" sobre um conjunto de dados históricos contendo entradas e saídas.

Baseado no treinamento com os dados históricos, o modelo pode tomar decisões precisas quando recebe novos dados.





Aprendizagem Não Supervisionada

A aprendizagem não supervisionada ocorre quando um algoritmo aprende com exemplos simples, sem qualquer resposta associada, deixando a cargo do algoritmo determinar os padrões de dados por conta própria.





Aprendizagem Por Reforço

O conceito de Aprendizagem Por Reforço (Reinforcement Learning) é como aprender por tentativa e erro. Os erros ajudam a aprender, porque eles têm uma grande penalidade associada a eles (custo, perda de tempo e assim por diante), ensinando que um determinado curso de ação tem menor probabilidade de êxito do que outros.



#### Aprendizagem Supervisionada

A aprendizagem supervisionada ocorre quando um algoritmo aprende a partir de dados históricos de exemplo, com entradas (inputs) e possíveis saídas (outputs), que podem consistir em valores quantitativos ou qualitativos, a fim de prever a resposta correta quando recebe novos dados.











## Big Data Analytics com R e Microsoft Azure Machine Learning

# O Processo de Aprendizagem em Machine Learning

Seja Bem-Vindo(a)!





## O Processo de Aprendizagem









## Processo de Aprendizagem





Um algoritmo de ML, como um algoritmo de classificação por exemplo, funciona da mesma forma. Ele constrói suas capacidades cognitivas através da criação de uma formulação matemática que inclui todas as características dadas sobre um determinado fenômeno.



#### Data Science Academy

## O Processo de Aprendizagem em Machine Learning

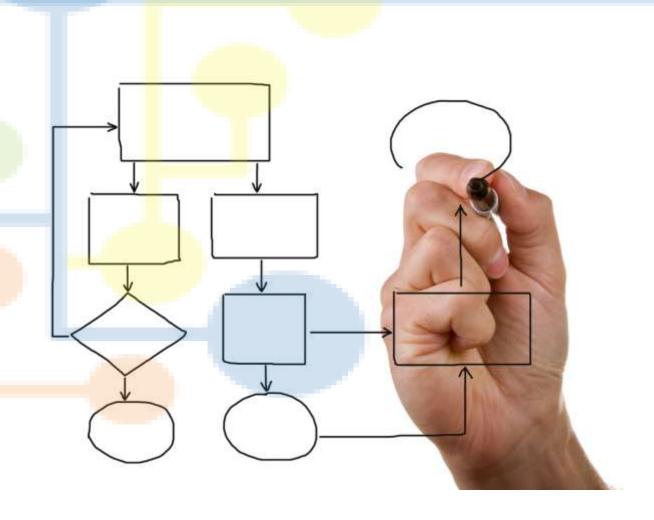
O Processo de Aprendizagem ocorre de diferentes formas e podemos dividir os algoritmos de Machine Learning em 3 grupos principais:

Aprendizagem Supervisionada, Aprendizagem Não Supervisionada e Aprendizagem Por Reforço.





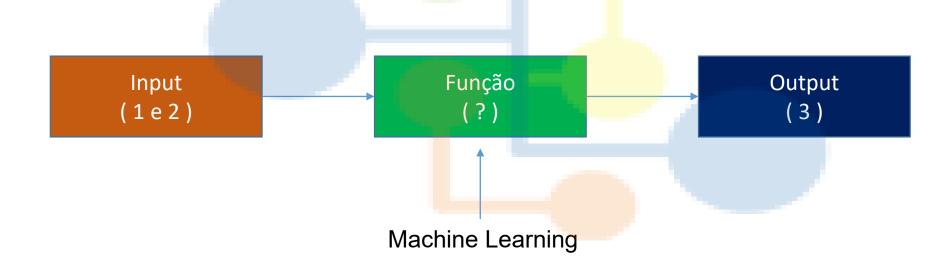
Do ponto de vista matemático, você pode expressar o processo de representação no aprendizado de máquina utilizando o mapeamento equivalente.







## Processo de Aprendizagem





## Big Data Analytics com R e Microsoft Azure Machine Learning

## O Processo de Aprendizagem em Detalhes

Seja Bem-Vindo(a)!





## Um componente chave do processo de aprendizagem é a generalização!





E para poder generalizar a função que melhor resolve o problema, os algoritmos de Machine Learning se baseiam em 3 componentes:















Os algoritmos de aprendizagem possuem diversos parâmetros internos (valores separados em vetores e matrizes).



Esses parâmetros funcionam como uma espécide de memória para o algoritmo, permitindo que o mapeamento ocorra e as características analisadas sejam conectadas.



As dimensões e tipos de parâmetros internos delimitam o tipo de funções-alvo que um algoritmo pode aprender. O engine de otimização no algoritmo muda os valores iniciais dos parâmetros durante a aprendizagem para representar função-alvo.

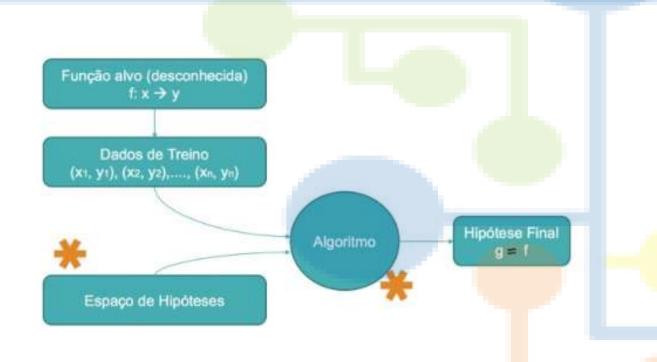












Espaço de Hipóteses

$$\mathcal{H} = \{h\}$$
  $g \in \mathcal{H}$ 

Algoritmo de Aprendizagem

Espaço de Hipóteses

Redes Neurais

Algoritmo de Aprendizagem

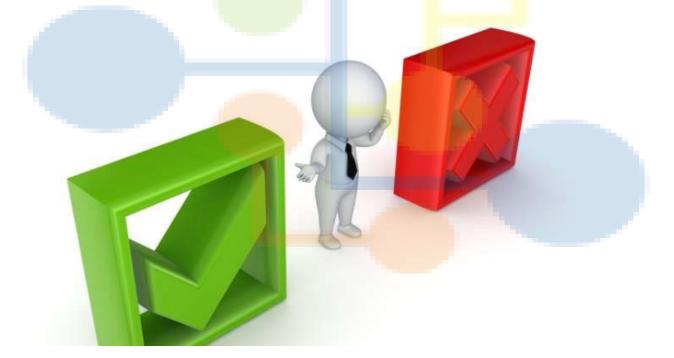
**Back Propagation** 

Modelo de Aprendizagem





## Falso Positivo





Big Data é uma grande mistura de dados. Um bom algoritmo de Machine Learning deve ser capaz de distinguir os sinais e mapear as funções alvo de forma eficiente.





**Cost Function** 

Hypothesis:  $h_{\theta}(x) = \theta_0 + \theta_1 x$ 

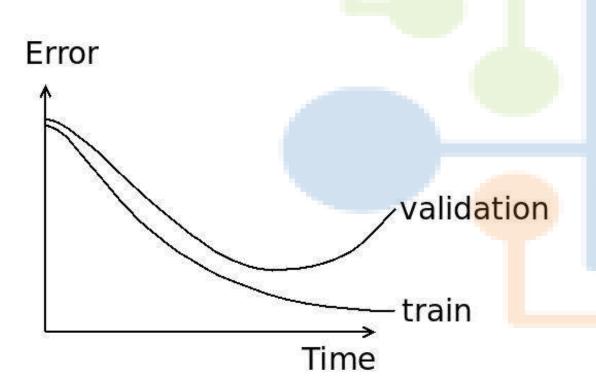
Parameters:  $\theta_0, \theta_1$ 

Cost Function:  $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^2$ 

Goal:  $\min_{\theta_0,\theta_1} \operatorname{minimize} J(\theta_0,\theta_1)$ 







Definindo o Erro

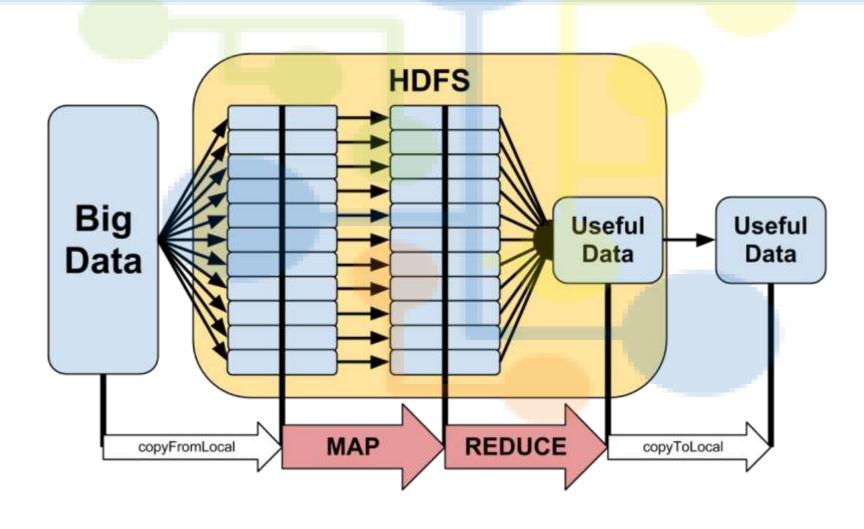
Cost Function → Nível de erro



As técnicas de aprendizagem de máquina baseadas em algoritmos estatísticos utilizam Cálculo e Álgebra Linear e os dados precisam estar carregados em memória.



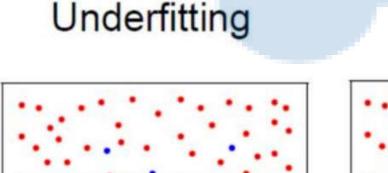


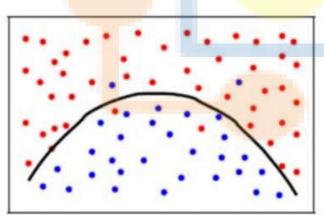


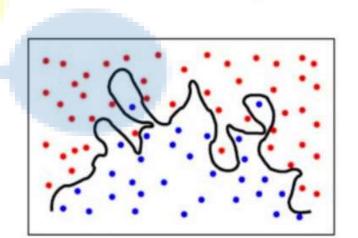




## O modelo pode aprender demais (overfitting) ou aprender de menos (underfitting).





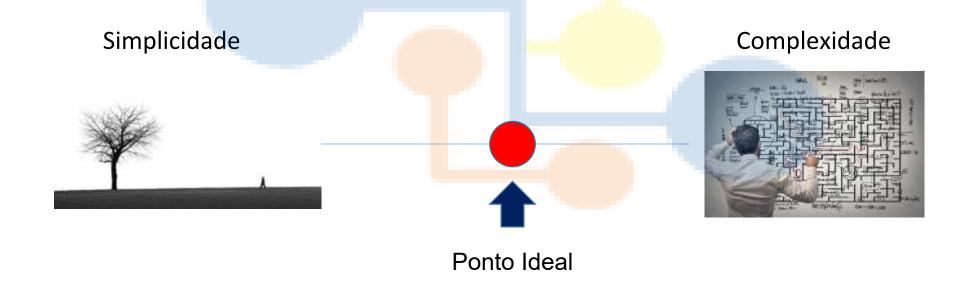


Overfitting





Para atingir o equilíbrio e criar grandes soluções de Machine Learning, você terá que fazer escolhas.







Para visualizar se os seus algoritmos de Machine Learning estão sofrendo algum tipo de força tendenciosa, você pode usar um gráfico chamado *Curva de Aprendizagem.* 





# O Processo de Aprendizagem em Detalhes

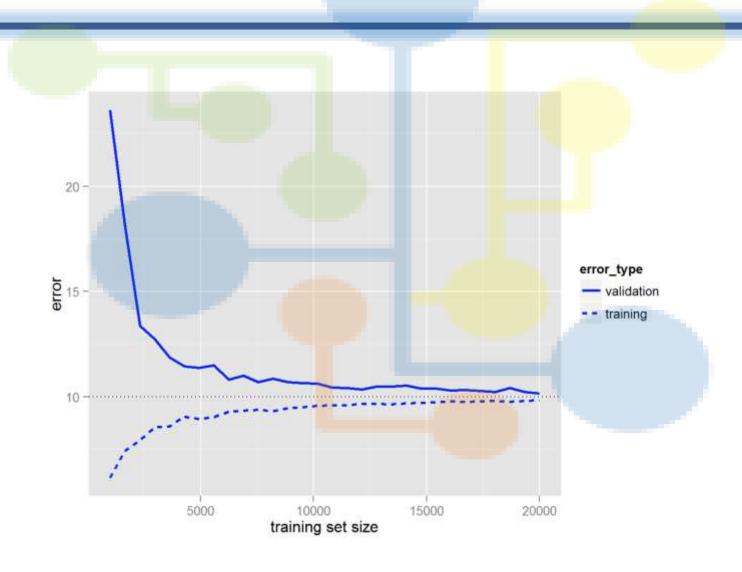
### Para usar uma curva de aprendizagem, você precisa:

- 1- Dividir seus dados em amostras, chamadas dados de treino e dados de teste (uma divisão 70/30 funciona bem). Dados de validação podem ser usados durante o treinamento.
- 2- Criar porções dos seus dados de treino, com tamanhos diferentes a cada passagem de treino.
- 3- Treinar seus modelos com os diferentes subsets. Registrar a performance.
- 4- Gerar um gráfico com os resultados. Atenção aos intervalos de confiança e ao desvio padrão.





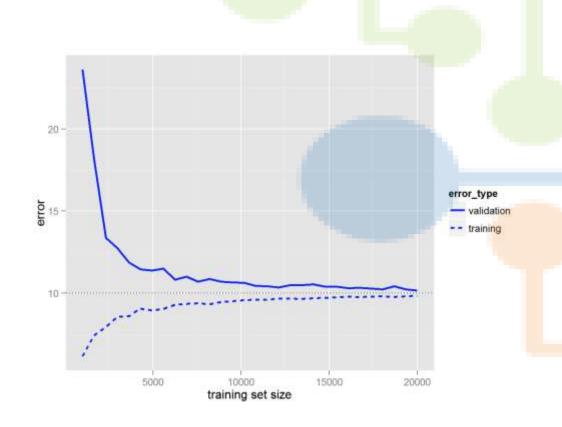
# Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4c O Processo de Aprendizagem em Detalhes







# O Processo de Aprendizagem em Detalhes



Podemos criar curvas de aprendizagem em R de diversas formas, usando os pacotes mlr, caret ou mesmo o ggplot2.

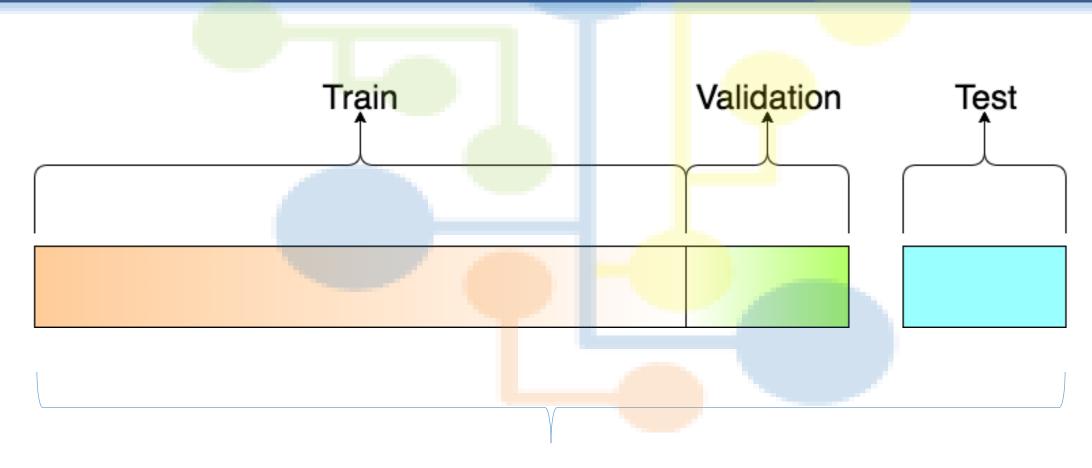


# Big Data Analytics com R e Microsoft Azure Machine Learning

Treinamento, Validação e Teste

Seja Bem-Vindo(a)!





Conjunto de Dados Completo



# Treinamento, Validação e Teste

75 a 70% - dados de treino

25 a 30% - dados de teste



# Treinamento, Validação e Teste

70% - dados de treino

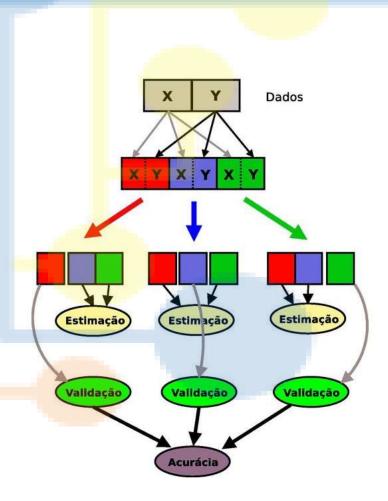
20% - dados de validação

10% - dados teste





Treinamento, Validação e Teste







Treinamento, Validação e Teste

n > 10.000







# **Cross-Validation**





# **Cross-Validation**

Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 1
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 2
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 3
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 4
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 5

Training data

Test data



### Cross-Validation

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.



# Big Data Analytics com R e Microsoft Azure Machine Learning

O Que é um Modelo de Machine Learning?

Seja Bem-Vindo(a)!

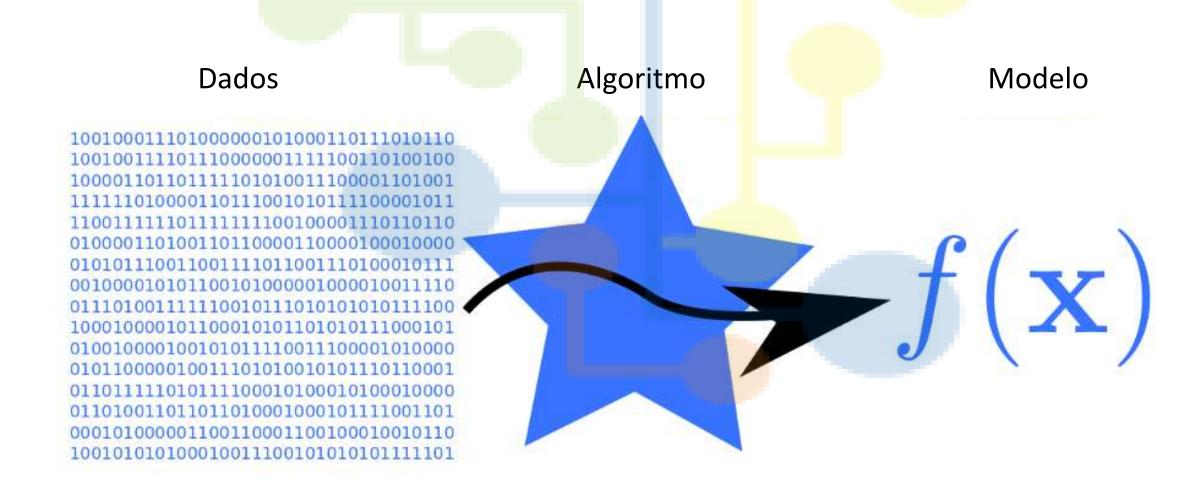


Como já vimos, a aprendizagem de máquina é um subcampo da Inteligência Artificial que evoluiu a partir do estudo de reconhecimento de padrões e teoria da aprendizagem computacional.



Machine Learning é um campo de estudo que dá ao computador a capacidade de aprender, sem ser programado de forma explícita.







### Modelo

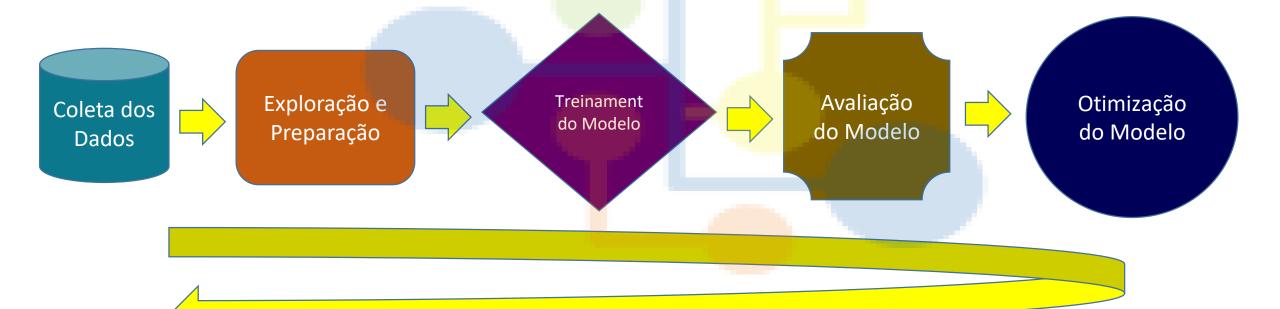
Existem muitos tipos diferentes de modelos. Você pode já estar familiarizado com alguns. Os exemplos incluem:

- Equações matemáticas
- Diagramas relacionais
- Agrupamentos de dados, conhecidos como clusters





# Criação do Modelo





#### Data Science Academy

# O Que é um Modelo de Machine Learning?



Este é um trabalho iterativo e assim como um surfista está sempre em busca da onda perfeita, seu trabalho como Cientista de Dados é buscar sempre o melhor modelo possível para suas previsões.





# Machine Learning na Prática

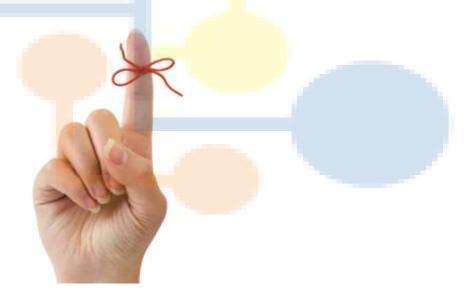








# Lembre-se: um modelo de Machine Learning será usado para resolver um problema específico!





Não caia na tentação de querer aplicar seu modelo a tudo que você vê pela frente.

Cada problema de negócio, cada conjunto de dados, pode requerer um modelo diferente.



# Big Data Analytics com R e Microsoft Azure Machine Learning

# Algoritmos de Machine Learning

Seja Bem-Vindo(a)!



### Aprendizagem Supervisionada

- Classificação
- Regressão

# Aprendizagem Não Supervisionada

- Clustering
- Segmentação
- Redução de Dimensionalidade

# Aprendizagem por Reforço

- Sistemas de Recomendação
- Sistemas de Recompensa
- Processo de Decisão



# Não Supervisionada

 Clustering & Dimensionality Reduction

o SVD

o PCA

K-means

Association Analysis

Apriori

> FP-Growth

Hidden Markov Model

#### <u>Supervisionada</u>

- Regression
  - Linear
  - Polynomial
- Decision Trees
- Random Forests

Classification

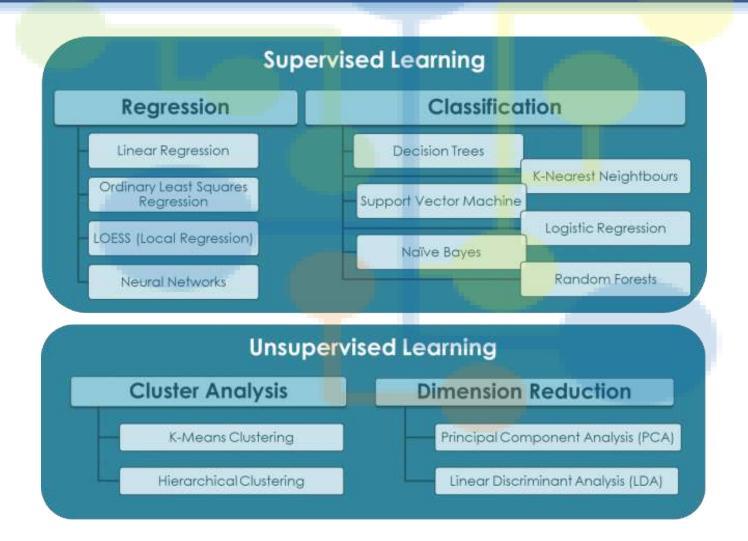
- KNN
- Trees
- Logistic Regression
- Naive-Bayes
- ⇒ SVM

Categórica

Contínua

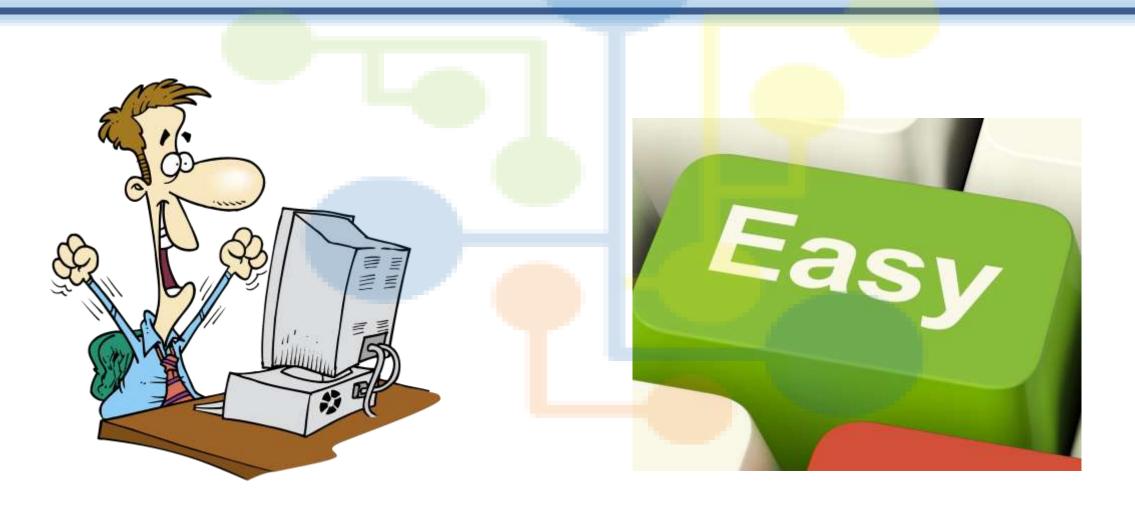














# Data Science Data Science Academy

# Algoritmos de Machine Learning

Há tantas algoritmos disponíveis com tantos métodos diferentes, que somente o processo de escolha de qual deve ser usado, já vai consumir bastante do seu tempo como Cientista de Dados.



Podemos categorizar os algoritmos de Machine Learning em 2 grupos principais:

Estilo de Aprendizagem Similaridade (Funcionamento)









### Algoritmos de Regressão

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)



### Algoritmos Regulatórios

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)



### Algoritmos Baseados em Instância (Instance-based)

- k-Nearest Neighbour (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)



# Algoritmos de Árvore de Decisão

- Classification and Regression Tree (CART)
- Conditional Decision Trees
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5



### **Algoritmos Bayesianos**

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)



### Algoritmos de Clustering

- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering



### Algoritmos Baseados em Regras de Associação

- Apriori algorithm
- Eclat algorithm



#### Redes Neurais Artificiais

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)



### **Deep Learning**

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders



### Algoritmos de Redução de Dimensionalidade

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)



### Algoritmos Ensemble

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (blending)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest



### **Outros Algoritmos**

- Support Vector Machines
- Computer Vision (CV)
- Natural Language Processing (NLP)
- Recommender Systems
- Graphical Models





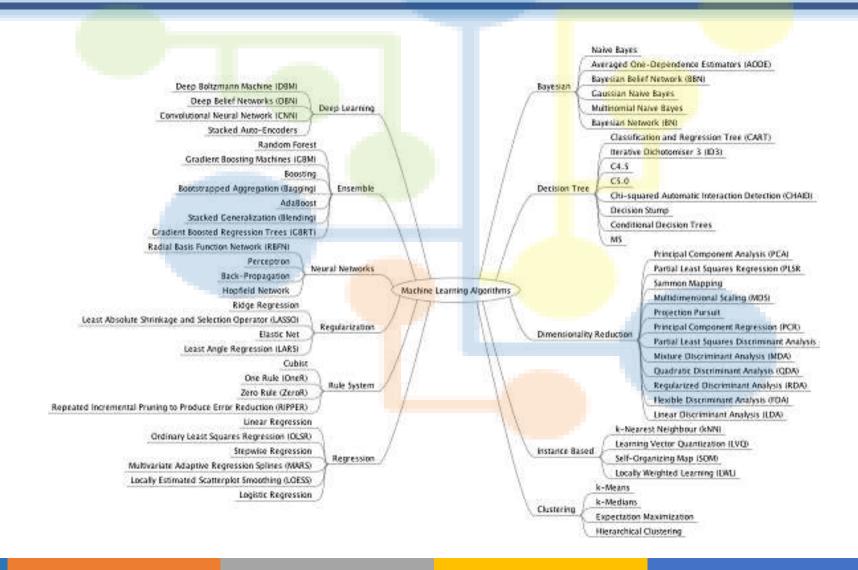


Sim, eu sei...muita coisa não??????

Mas espere, ainda não acabou!

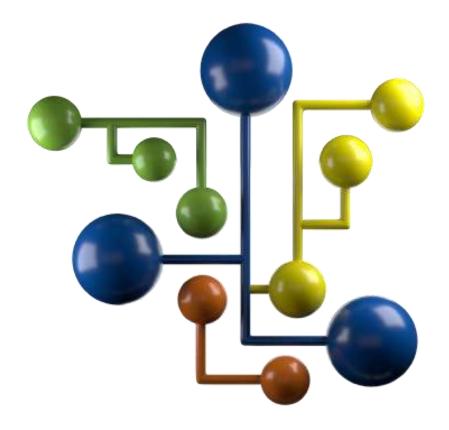








### Muito Obrigado por Participar!



Tenha uma Excelente Jornada de Aprendizagem.

**Equipe Data Science Academy** 

