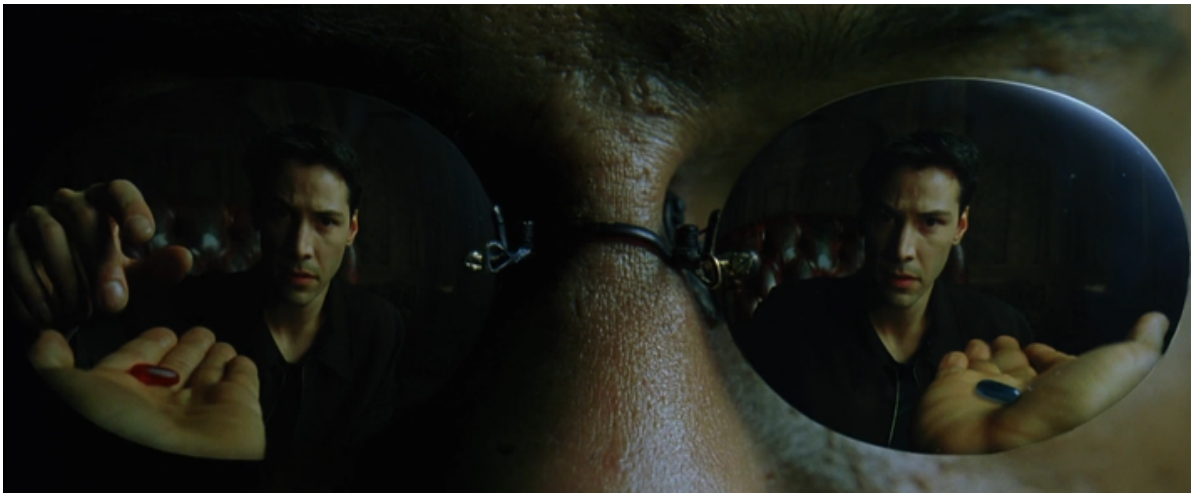


Capítulo 10 - Machine Learning - DSA

10.1 - Introdução a Machine Learning

10.2 - O que é Aprendizado de Máquina?

Conceitos de Aprendizagem

> Koogan/Houaiss:
> "Aprendizagem-Ação de aprender..."
> "Aprender- adquirir o conhecimento de, ficar sabendo, instruir-se ..."

Aprendizado é a capacidade de se adaptar, modificar e melhorar seu comportamento e suas respostas, sendo uma das propriedades mais importantes dos seres ditos inteligentes, sejam eles humanos ou não.

"...psicol.-Método que consiste em estabelecer conexões entre certos estímulos e determinadas respostas, cujo resultado é aumentar a adaptação do ser ao seu ambiente"

"Aprendizagem é o processo cognitivo que se estabelece entre o organismo e os estímulos emitidos pelo meio ambiente.
Assimilação de informações: aprendizado."
Filme: O Enigma de Kaspar Hauser".

Adaptação >> **Correção** >> **Otimização** >> **Representação** >> **Interação**

- Adptação: Mudança de comportamento de forma a evoluir, melhorar segundo algum critério;
- Correção: Correção dos erros cometidos no passado, de modo a não repeti-los no futuro;

- Otimização: A melhoria da performance do sistema como um todo, implica numa mudança do comportamento do sistema buscando uma melhoria;
- Representação: Uma forma de representar este conhecimento. Uma maneira de guardar conhecimento em regras gerais;
- Interação: Trocar experiência com o meio que o cerca para adquirir novos conhecimentos.

Estamos tentando reproduzir o processo de aprendizado de seres humanos em máquinas, através de algoritmos de Machine Learning.

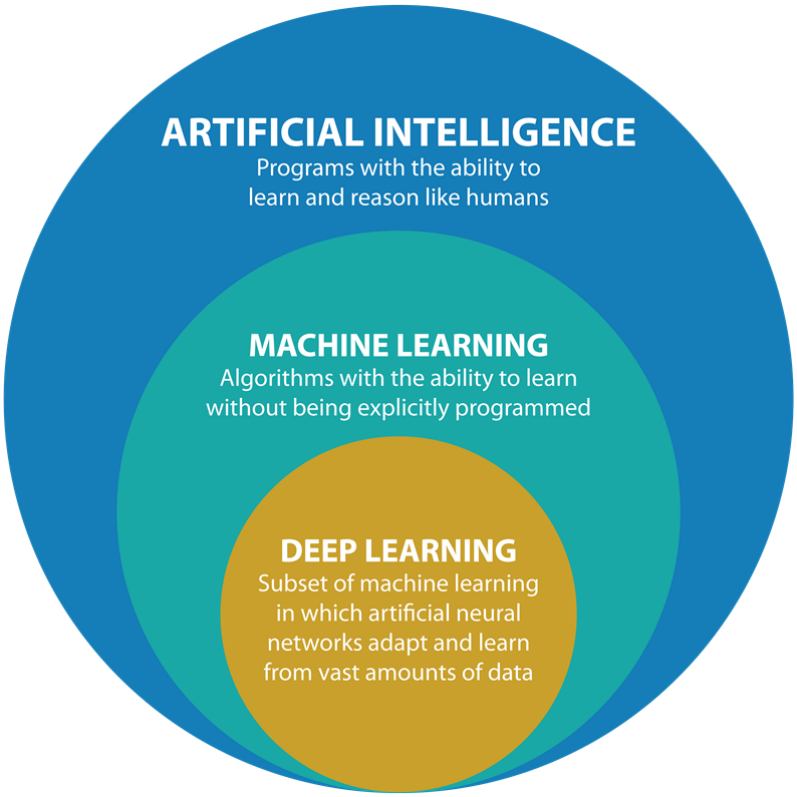
10.3 - O que é Aprendizado de Máquinas?

Machine Learning é Matemática, Estatística e Programação de Computadores!

> Machine Learning é um subcampo da Inteligência Artificial. É o estudo e construção de algoritmos que podem aprender a partir de dados e fazer previsões. O aspecto iterativo do aprendizado de máquina é importante porque, conforme os modelos são expostos a novos dados, eles são capazes de se adaptar de forma independente.

> Machine Learning ou Aprendizado de Máquina é um metódo de análise de dados que automatiza o desenvolvimento de modelos analíticos. Usando algoritmos que aprendem iterativamente a partir de dados, o aprendizado de máquina permite que os computadores encontrem insights através do reconhecimento de padrões.

10.4 - Inteligência Artificial x Machine Learning x Deep Learning.



Machine Learning é um subconjunto da Inteligência Artificial

Dentro de Machine Learning temos vários algoritmos.

Dentre estes algoritmos temos uma categoria de destaque:

Deep Learning

Conjunto de redes neurais artificiais

10.5 - Tipos de Aprendizagem de Máquina

- Aprendizado de Máquina:
  - Aprendizagem Supervisionada;
    - Previsão de valores ou classes;
    - Os dados de treino precisam conter os valores de entrada e saída, para que o modelo aprenda como, a partir de novos dados de entrada, gerar a saída correta.
  - Aprendizagem Não Supervisionada;
    - Identificação de grupos (clusters) de dados;
    - Os dados de treino contém apenas entrada;
  - Aprendizagem Por Reforço;

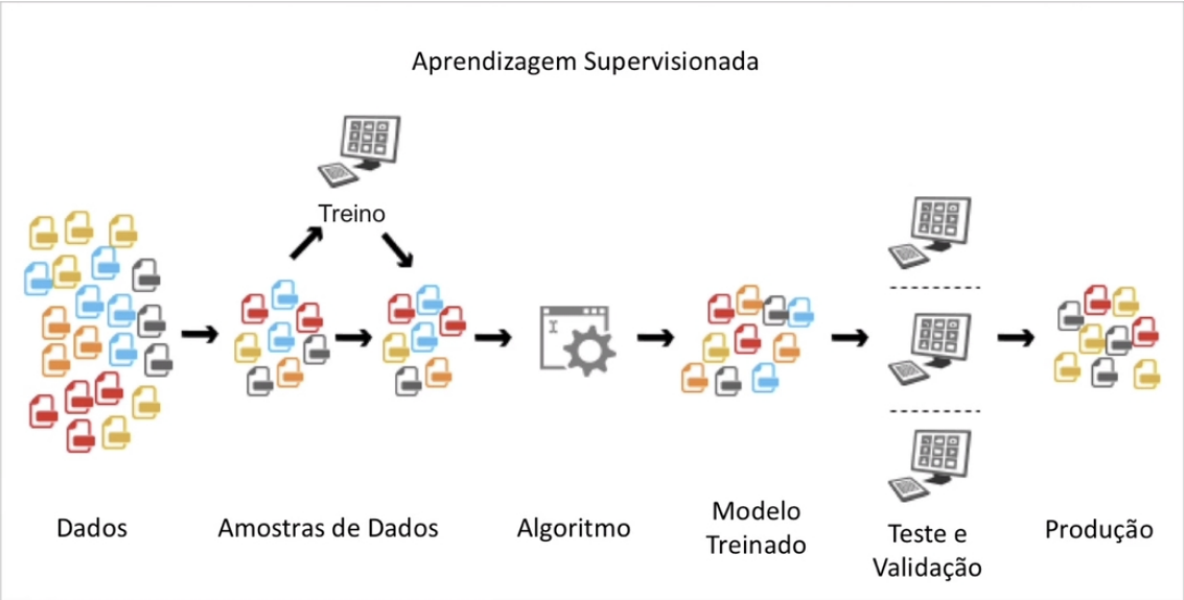
Aprendizagem Supervisionada	**Aprendizagem não Supervisionada**
Previsão de valores ou classes;	Identificação de grupos (clusters) de dados;
Os dados de treino precisam conter os valores de entrada e saída, para que o modelo aprenda como, a partir de novos dados de entrada, gerar a saída correta.	Os dados de treino contém apenas entrada;

10.6 - Aprendizagem Supervionada - Parte 1/2

10.7- Aprendizagem Supervionada - Parte 2/2

O Algoritmo aprende a partir dos Dados de exemplo, (dados de entrada) e possíveis resultados (dados de saída), podendo ser valores quantitativos e valores qualitativos. A fim de prever a resposta correta quando recebe novos conjuntos de dados.

Semelhante a abordagem Humana mediante a supervisão de um professor.



Exemplo

Atributos (tamanho, números de quartos, ano de constução) = **dados de entrada** - *Input*

Preço da Casa = **dados de saída** - *Output*

A **Aprendizagem supervisionada** se divide em duas categorias:

- **Classificação**: tem como alvo variáveis Qualitativas (categoricas), pegando a entrada e atribuindo uma classe, categoria.

  Ex.: "sim" ou "não", Mapeamento de uma imagem facial, é masculino ou feminino?

  Se forem duas 2 opções --> binária (2 classes)

  Se forem mais de duas opções --> Classificação Multiclasses
- **Regressão**: O alvo é um valor numérico, valor diferente de um booleano.

  EX.: Quanto custa? Quantos existem?

Aprendizagem Supervisionada

É o termo usado sempre que o programa é "treinado" sobre um conjunto de dados pré-definido.

Os algoritmos de aprendizado supervisionado fazem previsões com base em um conjunto de exemplos. Dados históricos de uma ação na bolsa (exemplo);

Análise de Sentimentos é um tipo de classificação, ou seja, aprendizagem supervisionada.

é usada em aplicações aonde dados históricos preveem eventos futuros;

10.8 - Aprendizagem Não Supervisionada

O **Algoritmo** aprende com os dados de entrada mas sem qualquer resposta associada;

Alguns sistemas de recomendação que você encontra na internet sob a forma de automação de marketing são baseados neste tipo de aprendizagem. (algoritmo de automação)

não dizemos ao algoritmo qual é a resposta certa ele as infere através dos dados de entrada e detecta padrões.

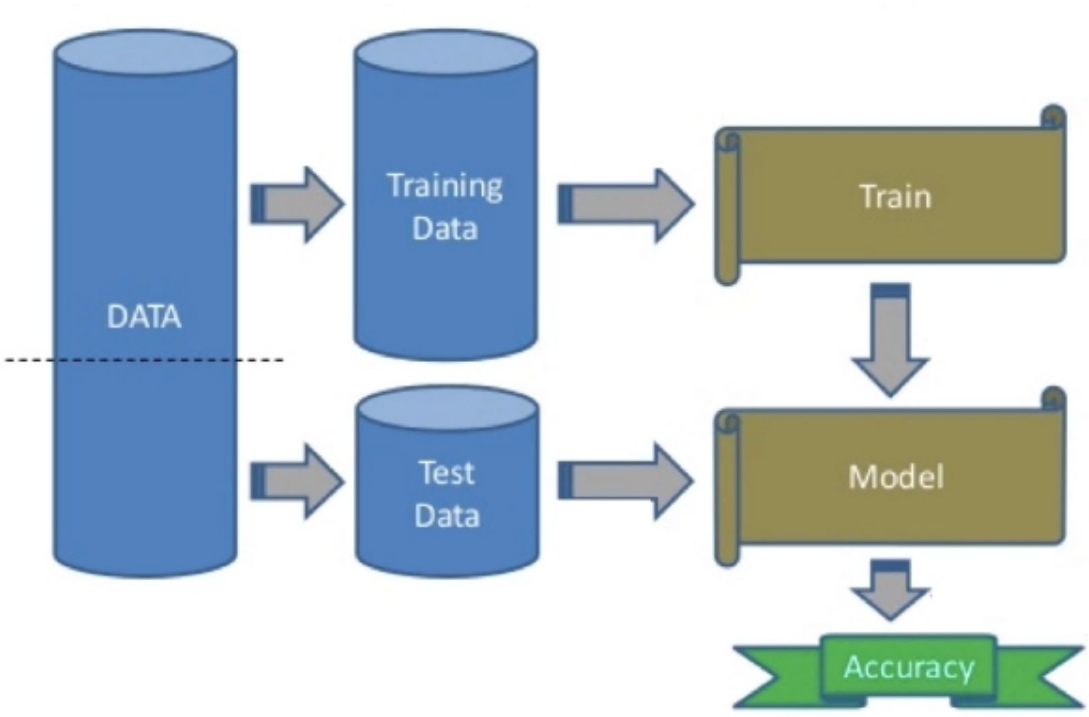
O aprendizado não supervisionado é usado com dados que não possuem rótulos históricos, ou seja, nós não temos variáveis target (as variáveis de saída) para serem estimadas.

O objetivo é explorar os dados e encontrar algum rótulo neles.

técnicas mais populares: Mapas auto-organizáveis, agrupamentos e decomposição em valores;

10.9 - Treinamento, Validação e Teste

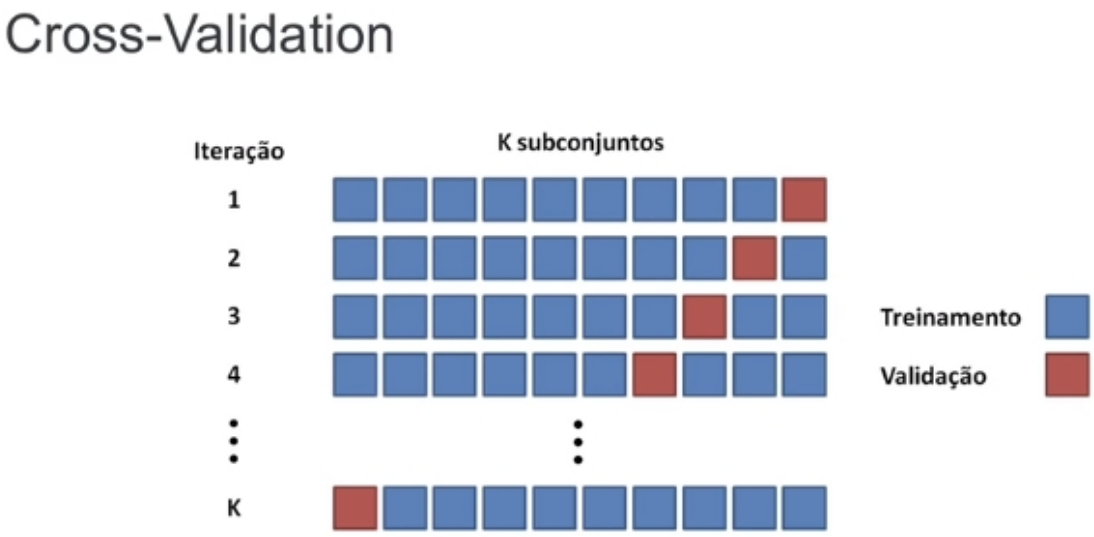
O processo de aprendizagem dos algoritmos de Machine Learn começa com a criação de subsets dos seus dados, são os chamados dados de treino e dados de teste.



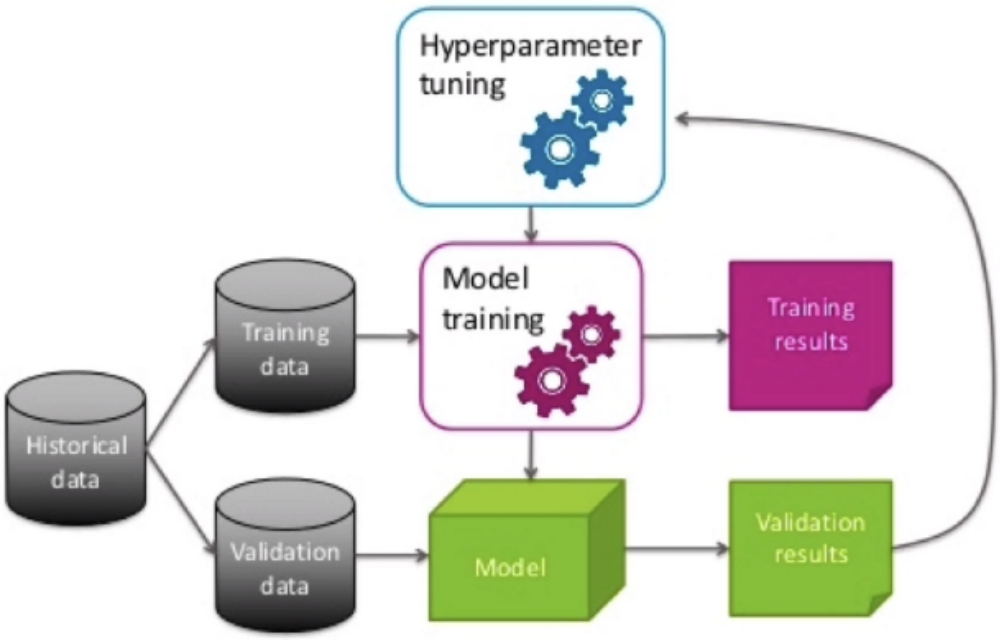
- A divisão comum é de:
  - 75 a 70% - dados de treino;
  - 20% - dados de validação;
  - 10% - dados de teste;

  é recomendável realizar a separação de forma aleatória, independente da ordenação inicial dos dados.

10.10 - Cross-Validation.



10.11 - O que é um modelo preditivo?



Etapas Repetidas em ciclos:

- Coleta de Dados.
- Exploração e Preparação.
- Treinamento do Modelo.

  O processo de "fit" do modelo a um dataset é chamado de treinamento do modelo.
- Avaliação do Modelo.
- Otimização do Modelo.

10.12 - Modelo Preditivo Um Pouco de Matemática

Modelo Preditivo é uma função matemática que, aplicada a uma massa de dados, consegue identificar padrões ocultos e prever o que poderá ocorrer.

- Regressor ou Classificador: são funções matemáticas;

O que é um processo estocástico?

é um Fenômeno que varia em algum grau, de forma imprevisível, à medida que o tempo passa!

exemplos:

- Variação do tráfego em um cruzamento;
- Variação diária no tamanho do estoque de uma empresa;

- Variação minuto a minuto do índice IBOVESPA;
- Variação no estado de um sistema de potência;
- Variação no número de chamadas feitas a uma central telefônica.

O objetivo do aprendizado de máquina é aprender a aproximação da função f que melhor representa a relação entre os atributos de entrada (chamadas variáveis preditoras) com a variável de saída (chamada de variável target).

10.13 - O Processo de Aprendizagem

Um componente chave do processo de aprendizagem é a generalização.

> Se um algoritmo de Machine Learning não for capaz de generalizar uma função matemática que faça previsões sobre novos conjuntos de dados, ele não está aprendendo nada e sim memorizando os dados, o que é bem diferente.

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

- **Representação**
    - Cira um modelo que produz um resultado para um conjunto específico de Inputs;
    - A **Representação** é um conjunto de modelos que o algoritmo pode aprender;
- **Avaliação**
    - Determina que modelo funciona melhor para criar o resultado esperado
    - O algoritmo de machine learning faz a avaliação dos modelos gerados por ele mesmo e atribui pontos pois mais de um modelo pode resolver o mesmo problema. Cada modelo recebe uma pontuação - score - que ajuda a determinar o melhor modelo ser utilizado.
- **Otimização**
    - O conjunto de modelos que produzem o resultado correto dado o conjunto de dados de entrada.
    - O processo de treinamento busca entre os modelos o que melhor resolve o problema em questão, sendo o melhor modelo utilizado.
    - Existem diversas técnicas de utilização do modelo proposto para o aprendizado

Os algoritmos de Machine Learning possuem diversos parâmetros internos (valores separados em vetores e matrizes), que funcionam como referência para o algoritmo, permitindo que o mapeamento ocorra.

Todas as funções estão no que chamamos de " espaço de hipótese" que precisa conter todas as variações nos parâmetros no algoritmos de ML, precisa conter a função que resolve o problema em questão.

O modelo de Machine Learning é composto por Espaços de hipóteses e o algoritmo de aprendizagem;

Espaço de Hipótese

Durante a fase de otimização os algoritmos buscam as possíveis variações e combinações entre os parâmetros de forma a encontrar a melhor combinação para o correto mapeamento durante a fase de treino. encontrar o relacionamento matemático entre os dados de entrada e saída.

10.13 - Cost Function

Função interna do algoritmo ou função de custo, função de perda, função objetivo ou apenas função de erro. Determina o quão bem o algoritmo executa aquilo que foi proposto.

10.14 - Overfitting x Underfitting

Quando o mapeamento de dados se tornam mais complexos erros nas previsões (overfitting) e (Underfitting);

Para atingir o equilibrio devemos escolher entre a simplicidade e a complexidade;

10.15 - Elementos do Processo de Aprendizagem

Para que ocorra a aprendizagem, é preciso que:

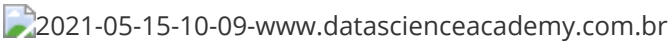
- Um padrão exista
- Não exista um único modelo matemático que explique esse padrão
- Dados estejam disponíveis

Variáveis preditoras: Dados de entrada;

Exemplo aprovação de Crédito de um Indivíduo.

Atributo	**Valor**
Sexo	Masculino
Idade	34
Salário Mensal	R$ 18.000,00
Anos no Emprego Atual	3
Anos de Residência	7
Saldo Bancário	R$ 32.671,94

- Elementos do processo de aprendizagem:
  - **Input** *x* {Dados do cliente} - variavel preditora
  - **Output** *y* {Decisão -> Crédito: Sim/Não} - variável target
  - **Função Alvo** f: *x* -> *y* {Representação do relacionamento}{Função matemática desconhecida}
  - **Dados** $(x^1, y^1), (x^2, y^2), \ldots , (Xn, Yn)$ {Dados históricos}
  - **Hipótese** g: *x* -> *y* {Função a ser descoberta pelo algoritmo}



10.16 - Espaço de hipóteses

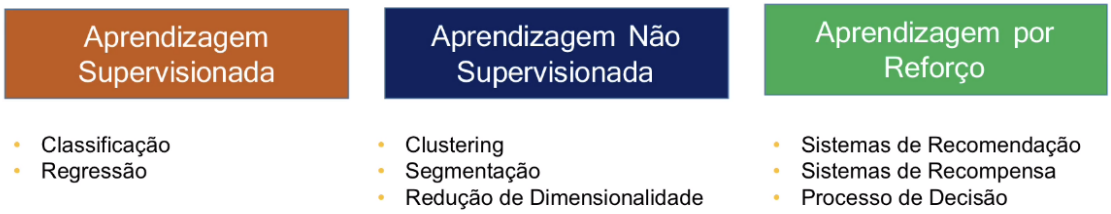
O modelo de aprendizagem possui duas partes:

- Espaço de Hipóteses: H = {h}   g pertence H
- Algoritmo de Aprendizagem:

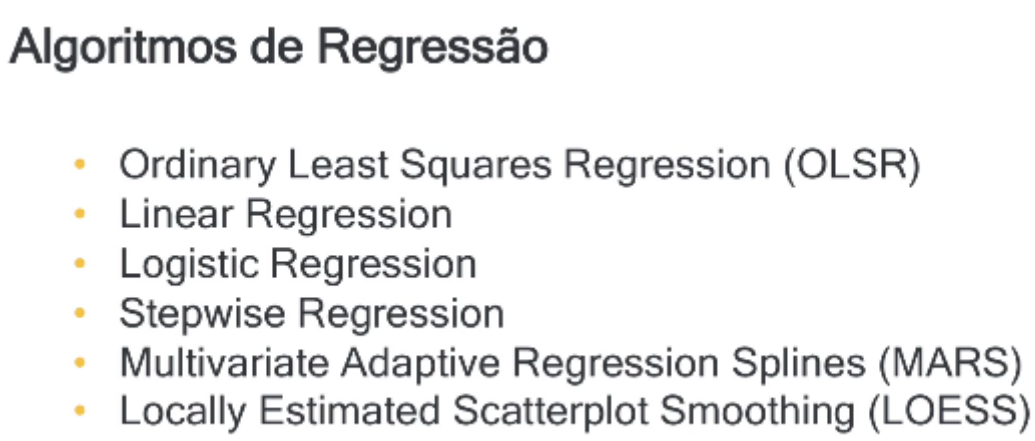
Espaço de Hipóteses + **Algoritmo de Aprendizagem** = **Modelo de Aprendizagem**

 2021-05-15-10-25-www.datascienceacademy.com.br

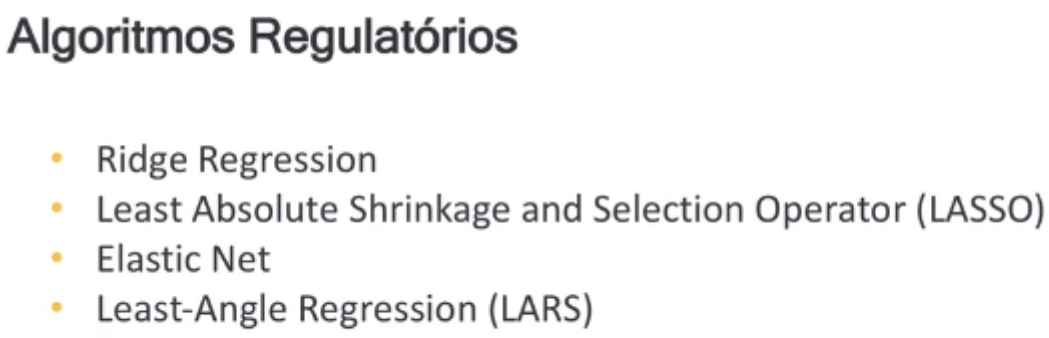
10.17 - Algoritmos de Machine Learning



Algoritmos de Regressão



Algoritmos Regulatórios



Algoritmos Baseados em Instâncias (Instance-based)

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

Algoritmos de Árvore de Decisão

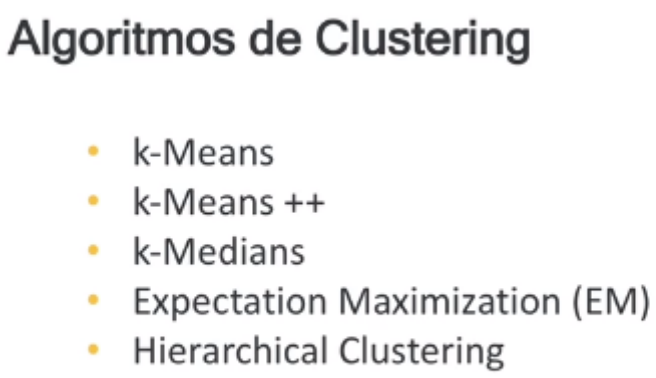
- Classification and Regression Tree (CART)
- Conditional Decision Trees
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5

Algoritmos Bayesianos

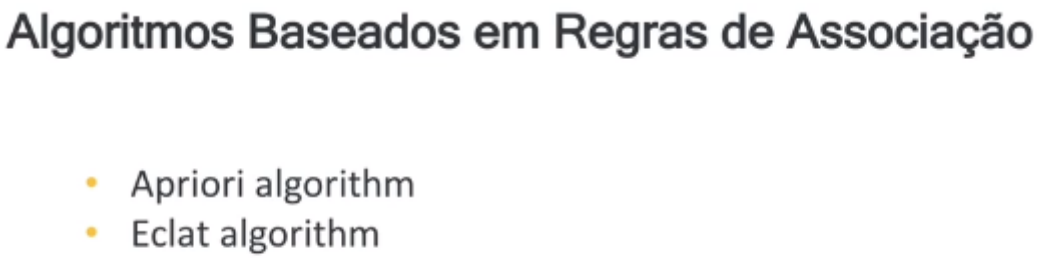
Algoritmos Bayesianos

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

Algoritmos de Clustering



Algoritmos Baseados em Regras de Associação



Redes Neurais Artificiais

correspôndencia de padrões

Deep Learning

Conceituação e métodos modernos para redes neurais maiores e mais complexas

Deep Learning

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

Algoritmos de Redução de Dimensionalidade

de forma não supervisionada, para resumir com menos informações

Algoritmos de Redução de Dimensionalidade

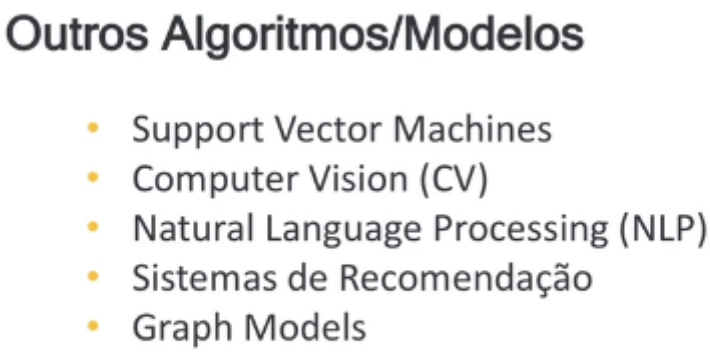
- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Multidimensional Scaling (MDS)
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)

Algoritmos Ensemble

Algoritmos Ensemble

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

Outros Algoritmos e Modelos



10.18 - Machine Learning Workflow

Fluxo de trabalho - Aquilo que precisa ser feito para se trabalhar de forma efetiva com aprendizado de máquina

É um conjunto de Etapas que sistematicamente transforma e processa dados a fim de criar soluções preditivas

As etapas necessárias dentro deste Fluxo de Trabalho:

1. Business Problem (problema de negócio)
2. Preparação de Dados
3. Seleção do Algoritmo
4. Treinamento do modelo (dividindo os dados em dado de treino e dado de teste)
5. Teste e Avaliação do Modelo

Dicas:

1. A etapa de preparção dos dados é uma das mais importantes. Lembre-se: a qualidade dos seus *outputs* será equivalente a qualidade dos seus *inputs*.
2. À medida que você caminha pelo processo, percebe a necessidade de modificar etapas anteriores. Isso é normal e esperado.

3. Os Dados raramente virão prontos, organizados e limpos. É o nosso trabalho fazer isso.
4. Mais dados = Melhores resultados - Por isso BIG DATA está revolucionando o mundo!!!
5. Não perca tempo com uma solução ruim. Avalie, otimize e se perceber que o resultado esperado não será alcançado, descarte e comece novamente!

10.19 - Business Problem - Definindo o Problema de Negócio

Definindo os Objetivos:

- Definir o escopo;
- Definir os níveis de performance do modelo preditivo;
- Definir o contexto;
- Definir como a solução será criada;

"Se você não sabe para onde vai, qualquer caminho serve!"

Exemplo proposto:

Prever se uma pessoa irá desenvolver diabetes

- Definir as fontes de dados;
- Compreender os atributos dos dados coletados;
- Selecionar as ferramentas de análise mais adequadas;
- Definir o resultado esperado. Neste caso: Verdadeiro ou Falso;
- Definir o nível de acurácia: 70% de precisão.

Fim da primeira parte;