

# Aprendizado de Máquina (machine learning) parte I

por: Rafael Stoffalette João

Data: 29/02/2020

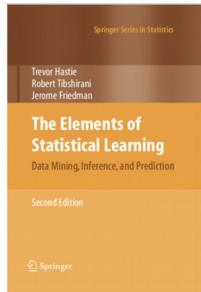
**UNIversidade Paulista (UNIP) - Araçatuba**

# Apresentação

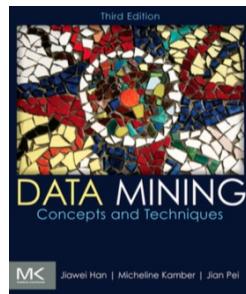
**Oi, eu sou o Rafael...**

# Bibliografia

MITCHELL, Tom M. **Machine Learning**. Nova Iorque: McGraw-Hill, 1997.



HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. Nova Iorque: Springer, 2009.



HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3. ed. Vermont: Elsevier, 2011 (The Morgan Kaufmann Series in Data Management Systems).

3

<http://minerandodados.com.br/index.php/2017/04/02/ciencia-dados-weka/>

# Objetivos da disciplina

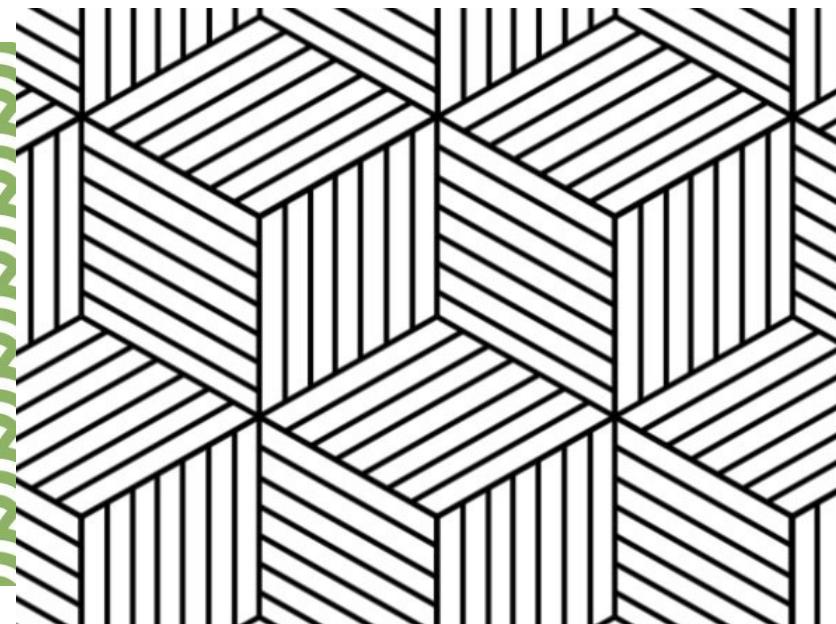
"Estudar os principais conceitos e paradigmas para a aprendizagem de máquina e apresentar a variedade de algoritmos e técnicas disponíveis para tal"

# Agenda da disciplina

- O que é o aprendizado de máquina;
- Seleção de características;
- Pré-processamento de dados;
- A ferramenta Weka;
- Treinamento supervisionado vs. Não supervisionado;
- Algoritmos de classificação;
- Árvores de decisão;
- Classificador Bayesiano;

# Primeiros passos da análise de dados

O que é um padrão?



# Primeiros passos da análise de dados

O que é um padrão?



# Primeiros passos da análise de dados

Imagen de um  
Brócolis por um  
microscópio

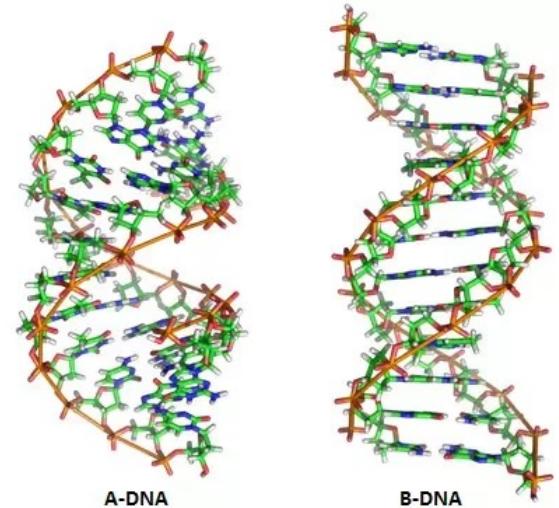


# Primeiros passos da análise de dados

**A I.A. foi, inicialmente projetada, para prever valores e situações.**

**A previsão só é possível pois algo é recorrente.**

- padrão



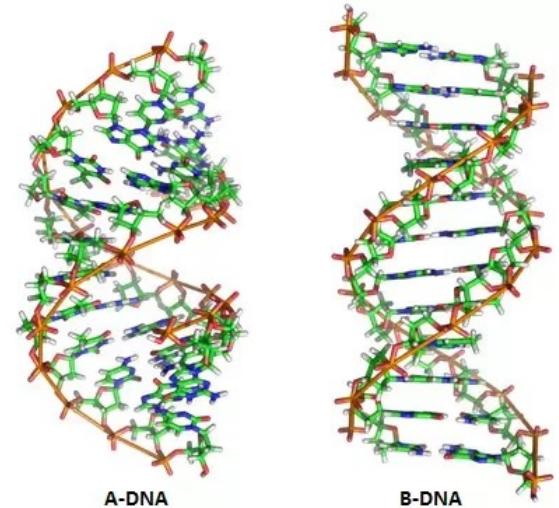
# Primeiros passos da análise de dados

**A I.A. foi, inicialmente projetada, para prever valores e situações.**

**A previsão só é possível pois algo é recorrente.**

- padrão

**Para prever algo com eficiência  
é preciso aprender como  
o padrão se repete**



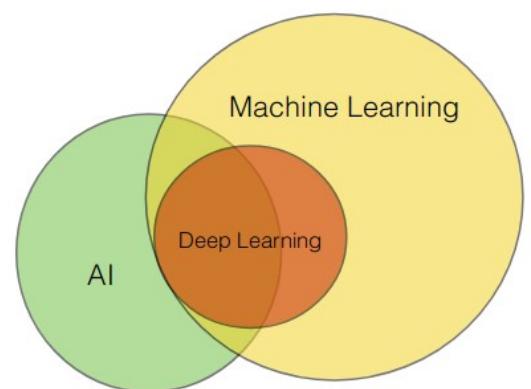
# Aprendizado de máquina

**Surgiu/cresceu em meados dos anos 90 com a busca pela Inteligência Artificial;**

**Para um sistema ser inteligente, deve aprender.**

Focar no processo de aprendizagem pode ser o diferencial do algoritmo;

O processo de KDD envolve toda a machine learning



# Aprendizado de máquina

## Definição:

Um programa de computador aprende a partir de uma experiência **E**, com respeito a algumas classes de tarefas **T**, com performance **P**, se sua performance nas tarefas em **T** melhora com a experiência **E**.

Por exemplo:

T: Reconhecer letras escritas a mão, via imagens;

P: % de instâncias corretamente reconhecidas;

E: Base de dados de imagens de letras com o resultado esperado.

# O que é machine learning

**É uma evolução da identificação de padrões**

Em 1959, Arthur Samuel definiu aprendizado de máquina como o

"campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados"

**Algoritmos voltados para fazer previsões**

aprendizado indutivo

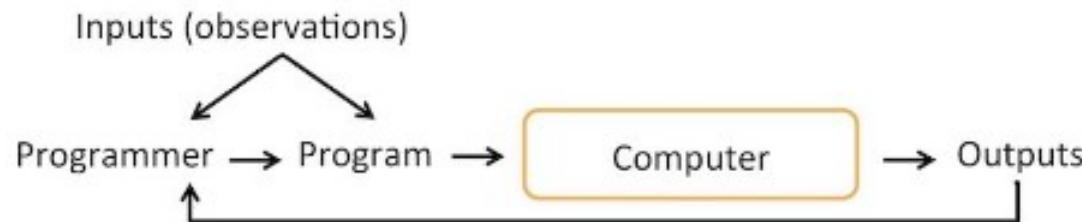
# O que é machine learning

**Não existe um algoritmo para ensinar um computador aprender, como uma criança.**

**Entretanto para diversas situações, existem estratégias que lidam com dados de forma a extrair o máximo de conhecimento possível.**

# O que é machine learning

## The Traditional Programming Paradigm



*Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed*  
– Arthur Samuel (1959)

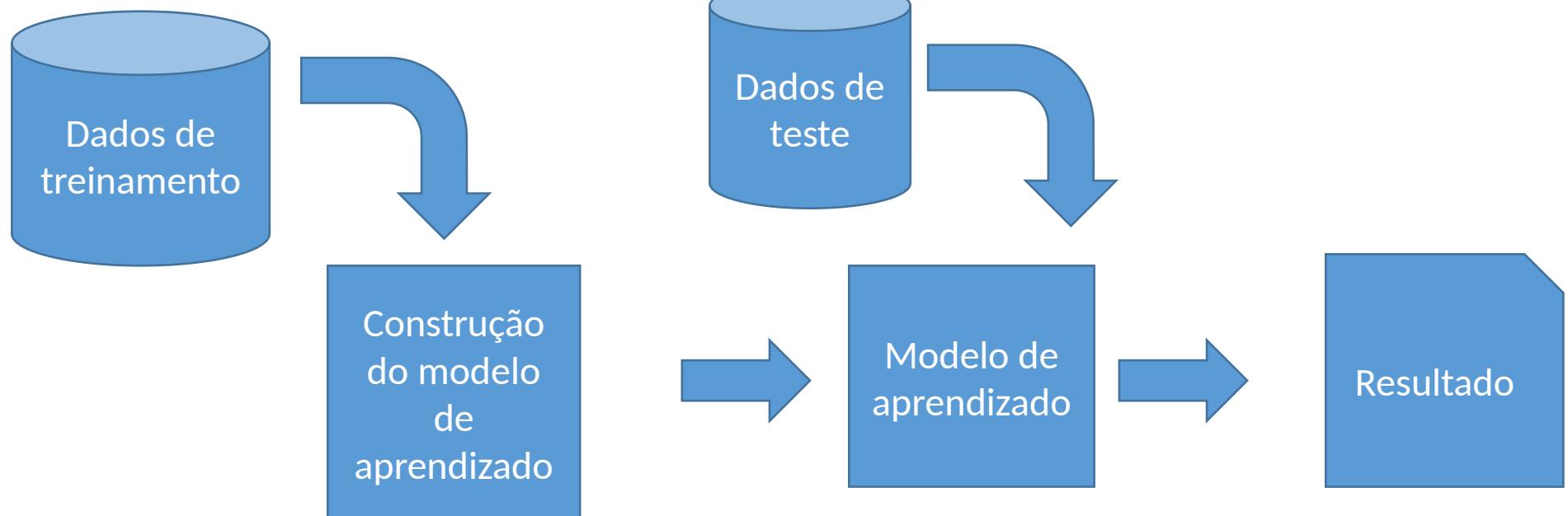
## Machine Learning



Sebastian Raschka, 2016

# O que é machine learning

- Etapa de aprendizagem
- Etapa de execução (teste)



Fonte: Elaborada pelo autor

# Aplicações de Machine Learning

## Detecção de Fraudes

- Bancos e operadoras de cartões, pioneiros na aprendizagem de máquina. Identificar possíveis transações fraudulentas e telefonar para o cliente confirmar compras suspeitas.

## Sistemas de recomendação

- Há quanto tempo um humano não te filmes na Netflix?

## Manuscrito

- Qual a vantagem dos Correios tirarem fotos das nossas assinaturas, uma vez que imagens são mais pesadas (em MB)?

## Autonomia

- Uber e Tesla já têm carros rodando sozinhos nos EUA.

# Aplicações de Machine Learning

## Sistemas de busca

- Os melhores resultados estão na primeira página do Google, SIM!

## Linguagem natural

- Entender o que o humano fala e interagir - bots de vendas.

## Mas nem tudo são flores - obstáculos:

- Dados mal armazenados;
- Falta de metadados;
- Falta de integração de dados;
- ... o problema da mineração de dados são os dados!

# Aplicações de Machine Learning



# Aprendizado

**O que é isso?**



**Por que tem tanta certeza?**

# Aprendizado

**Que tipo de peixe é esse?**



# Aprendizado

**Que tipo de peixe é esse?**



# Aprendizado

E agora?



# Aprendizado

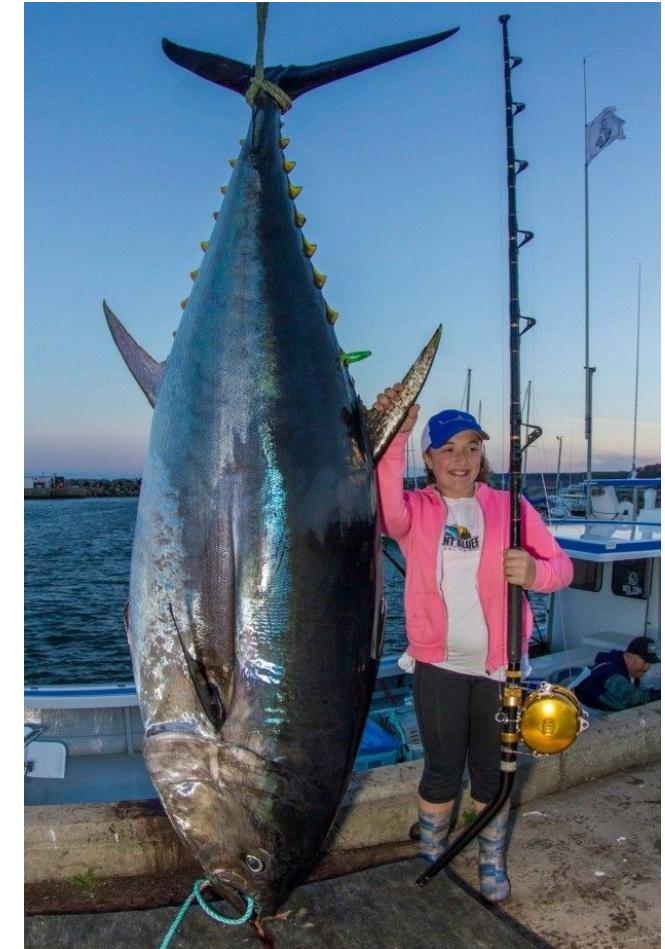
**Qual é o atum?**



# Aprendizado

## Qual a diferença entre atum e lambari?

Quais características definem cada peixe?



# **Seleção de características**

**etapa conhecida como “Seleção de características”...**

Uma das principais etapas do aprendizado

**Quais características são as mais importantes para definir a CLASSE que o algoritmo deve observar?**

**tamanho, peso, comprimento, cor da carne, cor da cauda, ...**

# Aprendizado

**Após selecionar as melhores características para o contexto e treinar o algoritmo (nossa cérebro, no caso) é fácil afirmar que esse peixe é um?**



# Aprendizado

**Após selecionar as melhores características para o contexto e treinar o algoritmo (nossa cérebro, no caso) é fácil afirmar que esse peixe é um?**



...um algoritmo que APRENDE deve estar apto, de alguma forma, a receber e aprender com novas informações.

# Aprendizado

...um algoritmo que APRENDE deve estar apto, de alguma forma, a receber e aprender com novas informações.

**Esse é o centro do estudo dos algoritmos que aprendem → proporcionar uma boa fonte de conhecimento.**

# Aprendizado de máquina

**Hora de aprender como se aprende algo:**

**<https://youtu.be/R9OHn5ZF4Uo>**

# Nem sempre é intuitivo

**Quais as melhores características para serem utilizadas no algoritmo inteligente?**

# **Seleção de características**

**A interpretação ou entendimento de uma cena demanda o reconhecimento de seus objetos.**

**Quando se tem conhecimento especialista - fácil.**

**Que tipo de características considerar?**

# Seleção de características

**Nominal: valores não numéricos e não ordenados.**

- Exemplo: cor e modelo do carro.

**Ordinal: valores não numéricos e ordenados.**

Exemplo: Faixa Etária: Jovem, Adulto, Idoso.

**Intervalar: valores numéricos.**

Exemplo: Temperatura em Graus Celsius (o zero é relativo)

# Seleção de características

## Variáveis qualitativas

escalas nominais ou ordinais.

## Variáveis quantitativas

escalas intervalares e proporcionais.

## Variáveis categóricas/dicotômicas

sexo: M, F

## Variáveis binárias

0 (ausência) ou 1 (presença).

## Variáveis discretas

idade

## Variáveis contínuas

distância.

# Seleção de características

Quando se tem  
conhecimento  
especialista – fácil.

Quais características  
descreveriam a  
Marge Simpson?

...E a Selma?



# Seleção de características

E em uma base de dados desconhecida????

aD1	aD2	aD3	aD4	aD5	amedia	a	bD1	bD2	t
22,38	19,94	20,28	20,8	20,13	20,706	20,71	1,43	3,26	8
20,64	17,45	19,47	19,39	20,07	19,404	19,4	2,71	5,24	7
17,83	16,71	16,82	16,43	14,36	16,43	16,43	4,12	4,91	4
20,56	20,93	21,71	21,56	21,29	21,21	21,21	2,64	2,5	1
9,67	10,95	13,8	13,46	13,97	12,37	12,37	11,98		11
23,06	22,23	20,05	20,15	9,29	18,956	18,96	0,75	1,23	2
21,88	22,18	21,43	22,45	21,16	21,82	21,82	1,49	1,48	1
18,96	18,27	11,93	18,91	15,65	16,744	16,74	3,43	4,22	8
15,06	14,7	15,34	12,54	15,46	14,62	14,62	7,02	7,56	7
13,73	14,85	16,05	17,83	14,77	15,446	15,45	8,18	7,3	6
23,57	23,49	23,58	23,52	23,66	23,564	23,56	0,3	0,39	0
13,7	10,36	10,78	11,67	10,09	11,32	11,32	8,53	11,5	1
13,81	18,74	17,08	15,05	15,42	16,02	16,02	7,5	3,49	4
22,52	20,84	20,1	18,82	20,12	20,48	20,48	1,29	2,85	2
23,98	23,65	23,81	23,32	22,9	23,532	23,53	0,02	0,28	0
11,33	10,84	12,21	10,6	11,03	11,202	11,2	9,66	10,47	8
18	14,02	14,76	17,09	19,41	16,656	16,66	5,3		9
22,86	22,59	21,28	20,45	17,12	20,86	20,86		1,121	1
11,51	13,78	15,54	13,48	14,33	13,728	13,73	10,82	8,95	7

# Pré-processamento de dados

ETL (*Extraction, Transformation and Loading*) :  
Torno de 70 a 80% do tempo.

Dados no formato do algoritmo: **RARO**

Dados completos: **MUITO RARO**

Volume de dados suficientemente correto: **RARO PRA CARAMBA**

Dados organizados: **NOSSASENHORACOMOÉRARO**

# Pré-processamento de dados

**Etapa na qual:**

- **as características não importantes são descartadas - limpeza;**
- **mais de uma base de dados é utilizada - integração;**
- **valores que fogem muito da normalidade são excluídos - detecção de outliers;**
- **ruídos são filtrados;**
- ...

# Pré-processamento de dados

- disparidade de escala;

- variância;

- atr. faltantes;

- muitos atributos;

- etc...

Precipitacao	TempMaxima	TempMinima	Insolacao	Evaporacao	Temp Media	Umidade Relativa Media	Velocidade do Vento Media	Ocorrencia
42.8	22.2	19.3	0	0.4	20.24	95.25	2.80	LANINA
29.4	23.1	19.3	0	0.5	20.06	95.5	3.17	EI NINO
26.2	23.5	18.9	0	0.4	20.64	96.5	2.27	LANINA
14.2	24.7	18.9	0	0.5	21.2	92.75	2.10	EI NINO
17.6	21	18.8	0	0.7	20.08	98.5	3.23	EL NINO
51.6	20.4	19.1	0	0.4	19.98	97.75	3.33	EL NINO
26	27	19.1	0	0.3	22	93.5	1.73	LANINA
1.4	29.9	19.6	0	1.4	23.4	80.5	2.13	LANINA
8.8	28.9	19.9	0	2.3	24.26	80.75	2.53	EL NINO
0.6	28.4	19.9	0	2.2	24.28	83	1.00	LANINA
7.6	20.7	31.1	0	2	24.82	70.25	2.13	EL NINO
0	29.7	18.7	0	4.1	23.88	86.5	3.03	LANINA
							2.13	LANINA
							1.60	LANINA

Dados faltantes: preencher ou excluir?

Boa variância, ou não?

# Pré-processamento de dados

**Extração de atributos:**

Atributos que possuem muitos dados faltantes

**Filtro de baixa variância:**

Atributos com pouca variância não trazem grandes informações

**Filtro de alta correlação:**

Atributos fortemente correlacionados (temp max, temp min e temp média)

**Eliminação de características**

Crescente e decrescente.

# Seleção de características

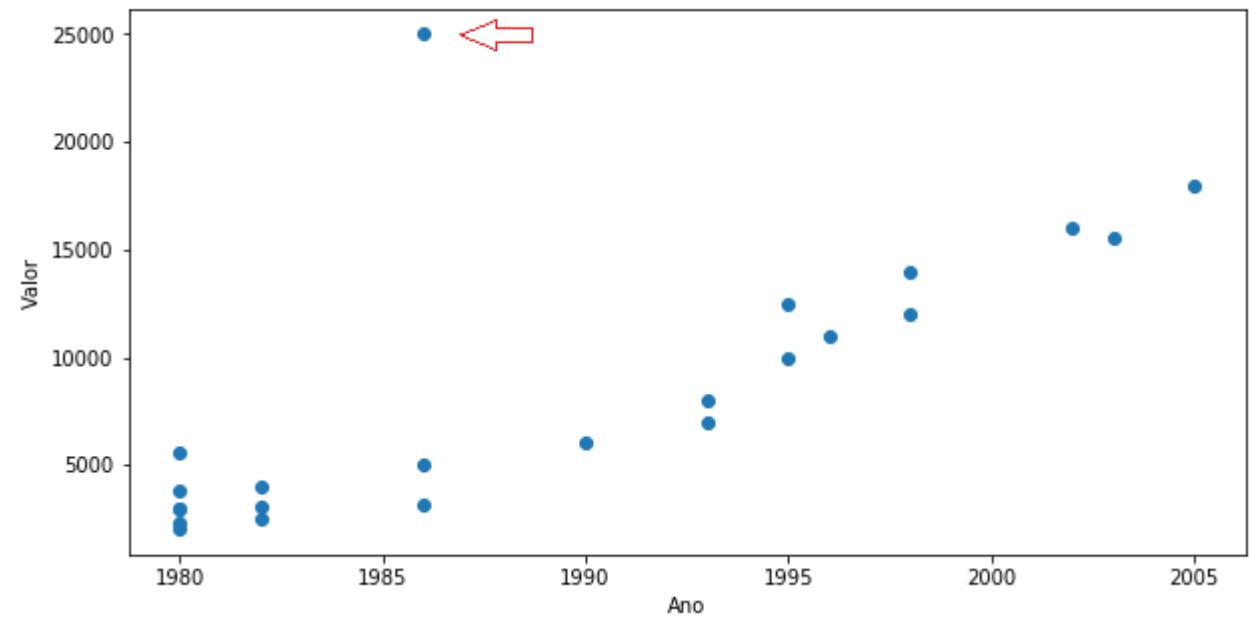
## Poda de nós (node-pruning)

Inicialmente, a rede neural é treinada, sendo posteriormente realizada a eliminação de nós seguida de um re-treinamento da rede, repetindo-se o processo até que seja alcançada a dimensão desejada.

A vantagem do método node-pruning é que ele simultaneamente determina o melhor subconjunto de características

# Pré-processamento - detecção de outliers

**Outliers são pontos que fogem muito do comportamento padrão**



# Pré-processamento - detecção de outliers

**Outliers são pontos que fogem muito do comportamento padrão**

**Sem outlier:**

Média: 41.6

Mediana: 44

Desvio padrão: 11.9

**Com outlier:**

Média: 126

Mediana: 44

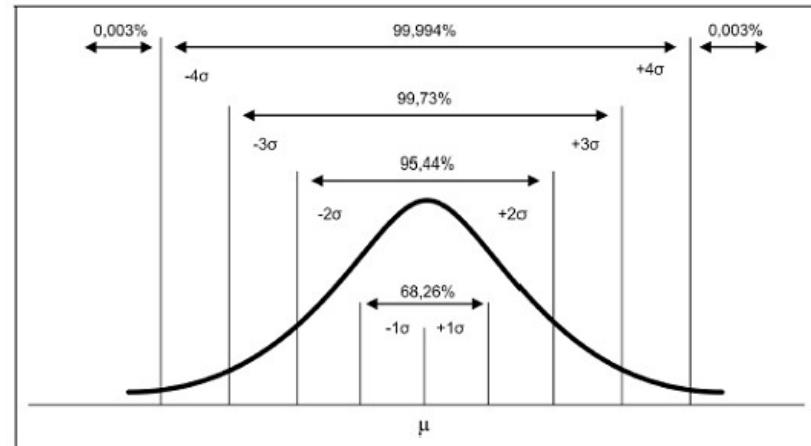
Desvio padrão: 192.5

	Nome	Idade		Nome	Idade
0	José Carlos	33	0	José Carlos	33
1	Manuel da Silva	57	1	Manuel da Silva	57
2	Maria Leite	27	2	Maria Leite	27
3	Antônio Siveira	47	3	Antônio Siveira	470
4	Pedro Lemos	44	4	Pedro Lemos	44

# Pré-processamento - distribuição normal

## Distribuição normal de probabilidade

Na estatística, se uma variável tem seus valores distribuídos de acordo com a distribuição normal, 68% dos dados se concentram em apenas 1 desvio padrão da média e 95% em dois desvios padrões.



**Variáveis aleatórias independentes, quando assumem MUITOS valores, tendem a seguir a distribuição normal em seus valores.**

# Pré-processamento - distribuição normal

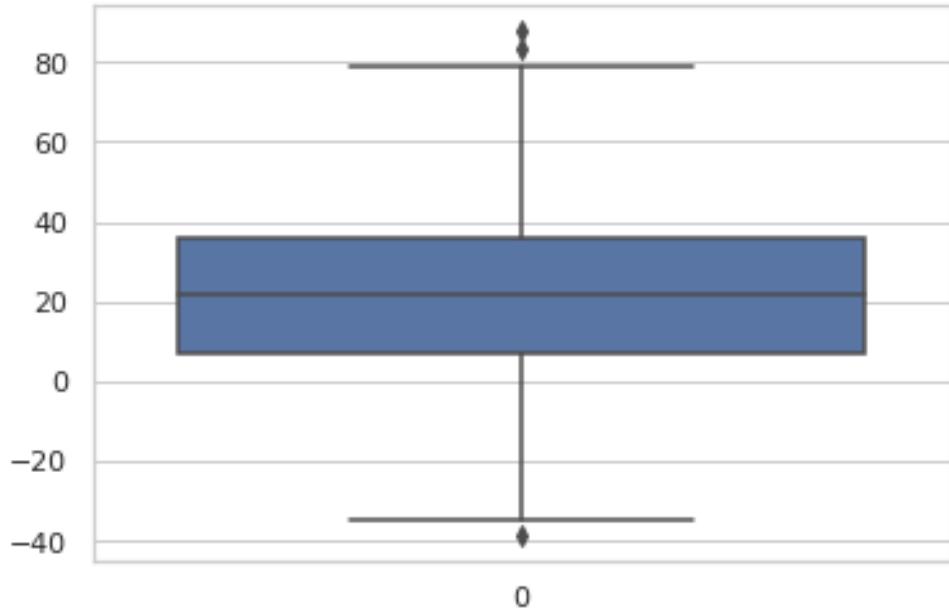
**Entende-se que um valor está fora do comportamento padrão SE seu valor ultrapassa 3 desvios padrões.**

```
import numpy as np
data = np.random.randint(1,100, size=100)
np.mean(data)
np.std(data)

for v in data:
    if v > np.std(data)*3:
        print("fora")
```

# Pré-processamento - outliers

```
import seaborn as sns  
import numpy as np  
  
data = np.random.randn(1000) * 20 + 20  
  
sns.boxplot(data=data)
```



# Pré-processamento - Normalização

Padronizar os valores de um atributo (característica/feature) dentro de um intervalo – geralmente [0,1]

```
from sklearn import preprocessing

entradaBruta = [[1.0, 12.1, 14.5, 2.1, 1.8]]
entradaNormalizada = preprocessing.normalize(entradaBruta)

print(entradaNormalizada)

>>>[[0.05232018 0.63307416 0.75864259 0.10987238 0.09417632]]
```

# Pré-processamento - Binarização

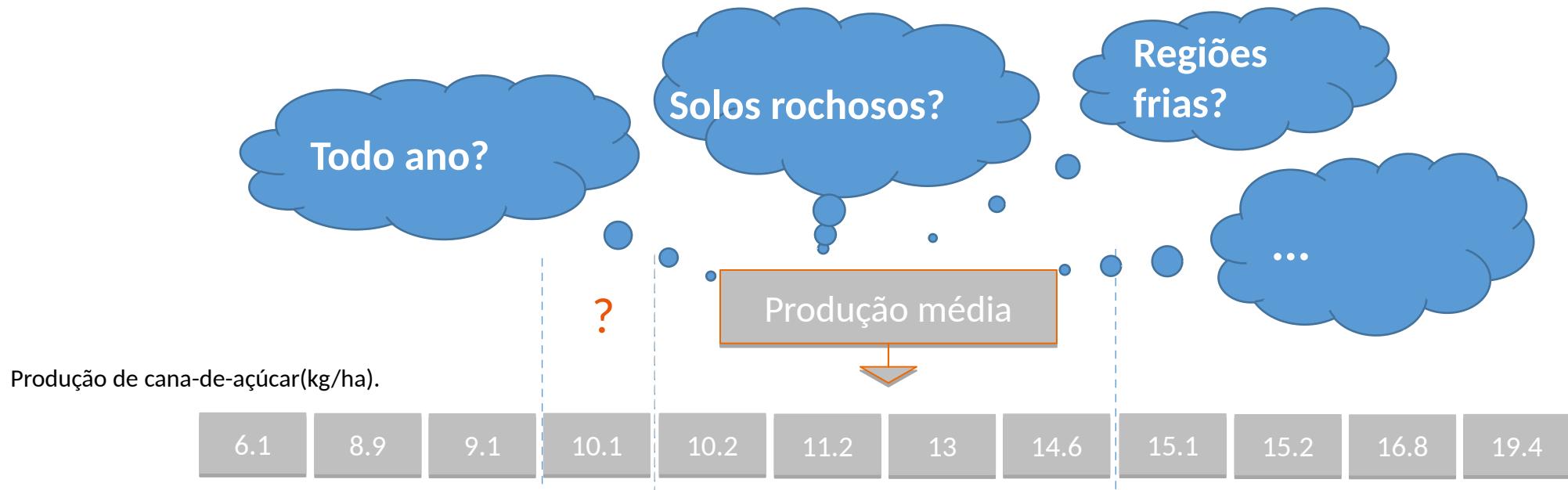
Converte valores de um atributo (característica/feature) em valores binários. 1 se ultrapassar o threshold, 0 cc.

```
from sklearn import preprocessing  
  
entradaBruta = [[1.0, 12.1, 14.5, 2.1, 1.8]]  
binarizer = preprocessing.Binarizer(threshold=1.9)  
entradaBinaria = binarizer.transform(entradaBruta)  
  
print(entradaBinaria)  
  
=>>>[[0. 1. 1. 1. 0.]]
```

# Pré-processamento - Discretização

Discretizar == encapsular informações SEMELHANTES;

- Perda/omissão de informação;
- Sensível ao ponto de vista do especialista;



# Pré-processamento - Discretização

Padronizar os valores de um atributo (característica/feature) em valores que representam cada intervalo.

```
from sklearn import preprocessing

entradaBruta = [[1.0, 12.1, 1.5, 0.1, 1.8],
                [2.0, 6.1, 14.5, 120.1, 1.8],
                [6.0, 2.1, 4.5, 222.1, 1.8]
               ]
discretizer = preprocessing.KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')
discretizer.fit(entradaBruta)
entradaDiscreta = discretizer.transform(entradaBruta)
print(entradaDiscreta)

>>>[[0. 2. 0. 0. 0.]
     [0. 1. 2. 1. 0.]
     [2. 0. 0. 2. 0.]]
```

# Base de dados Simpson

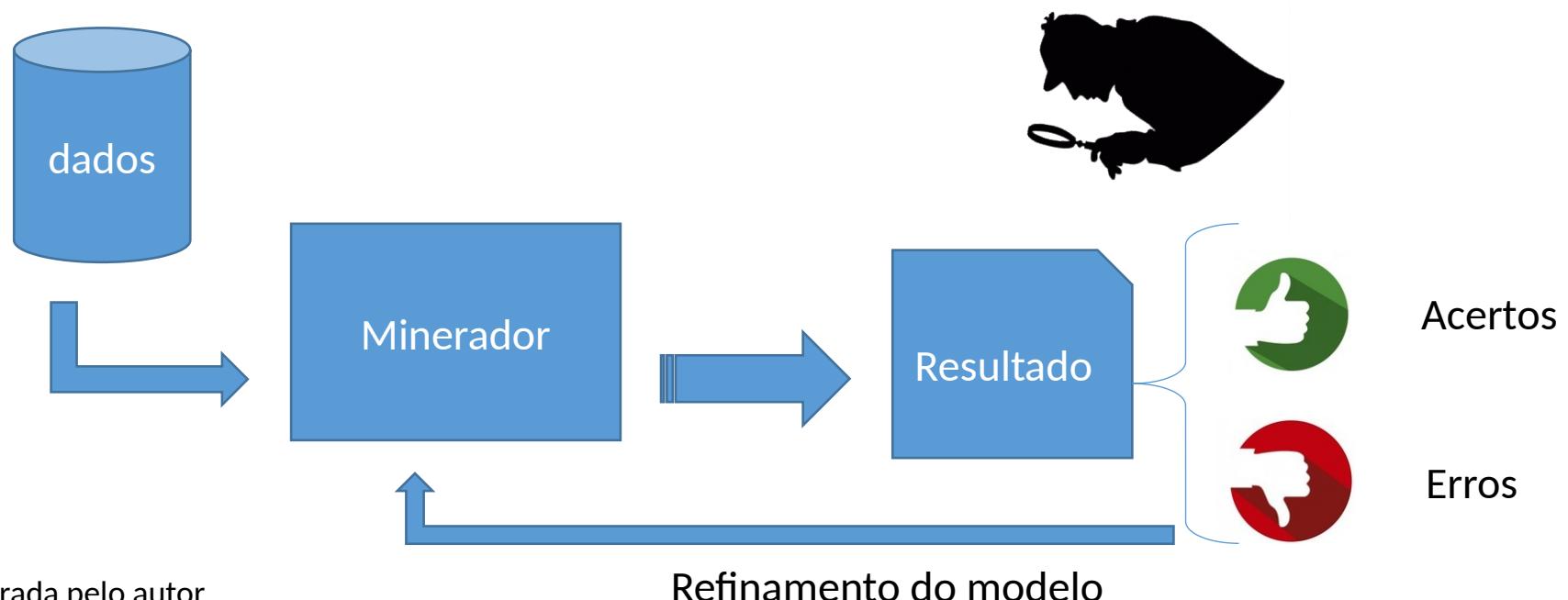
## Como construir uma base de dados Simpson?

E como utilizar  
técnicas de  
pré-processamento  
para melhorar a  
qualidade dos dados?



# Aprendizado supervisionado vs. Aprendizado não supervisionado

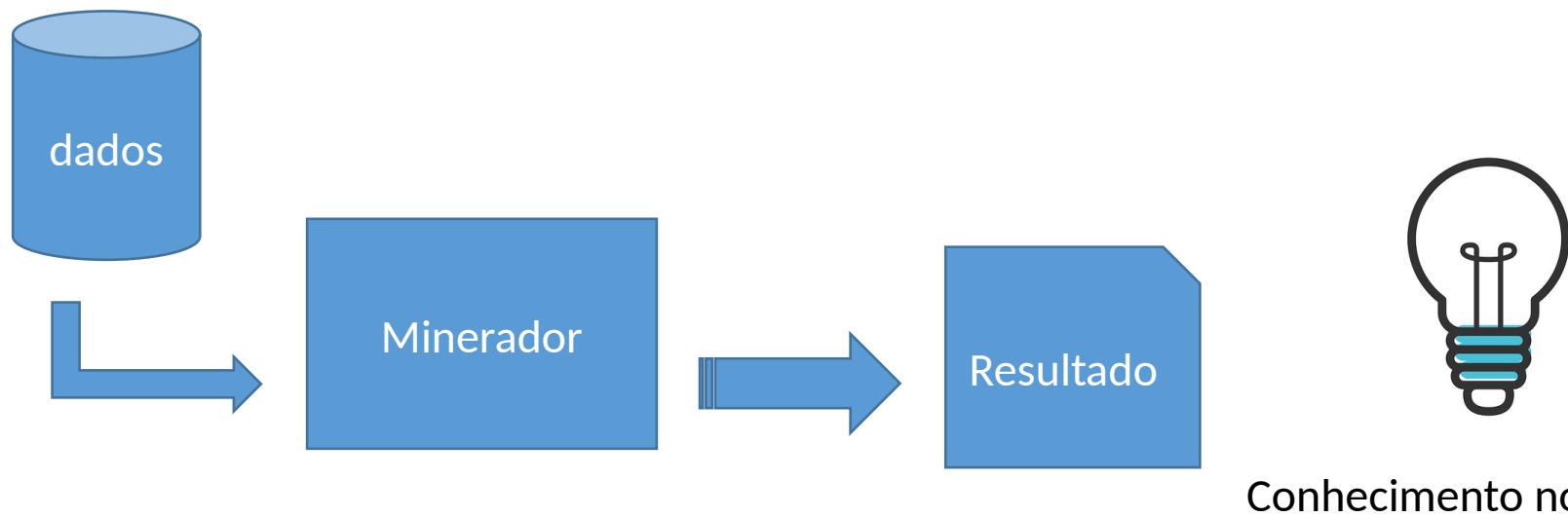
- Supervisionado: Quando se sabe o resultado esperado;



Fonte: Elaborada pelo autor

# Aprendizado supervisionado vs. Aprendizado não supervisionado

- Não supervisionado: Quando não se sabe o que a mineração resultará.



Fonte: Elaborada pelo autor

# Tipos de treinamento

Mesmo conjunto

Utiliza-se o mesmo conjunto para treinamento e para testes (NÃO RECOMENDADO)

Particionamento (%)

Define-se uma % para o conjunto de treinamento e o restante é testado

Validação cruzada:

10-grupos:

Partição do conjunto em 10 subgrupos:

Treinamento com 9 e teste com 1 (PARA TODOS OS GRUPOS)

Inserção independente:

Inclusão de um conjunto de treinamento diferente do conjunto de testes

# Algoritmos de aprendizado de máquina

Classificação: Valores semelhantes que descrevem um grupo de registros.

Regressão: Resultado esperado é um valor numérico (cotação do euro).

Agrupamento: Quais registros tem características em comum? Quais carac.

Descoberta de padrões: Segmentos da base de dados que se repetem;

Regras de associação: Quais itens implicam na ocorrência de outros.

# Ferramenta Weka

**Weka é uma ferramenta que proporciona uma coleção de algoritmos para tarefas de mineração de dados e aprendizado de máquina.**

- Modo Explorer: uso geral (pré-processamento, agrupamento, classificação, ...)
- Modo Experimenter: controle de treinamento (divisão de treino/teste)
- Modo KnowledgeFlow: lidar com fluxo de dados
- Modo Workbench: IDE antiga
- Modo Simple CLI: linha de comandos.

# Ferramenta Weka

**[https://waikato.github.io/weka-wiki/  
downloading\\_weka/](https://waikato.github.io/weka-wiki/downloading_weka/)**

# Arquivo .arff

O Weka é uma ferramenta completa para algoritmos de mineração de dados e aprendizado de máquina.

**É possível criar um arquivo de base de dados específico para ele - .arff**

# Arquivo .arff

@relation nome\_da\_base\_de\_dados

## Duas partes principais:

A primeira - lista de todos os atributos (com os tipos definidos, ou os valores que ele pode assumir entre { } ).

@attribute

A segunda: registros da base de dados (valores separado por vírgula)

@data

# Arquivo .arff

@RELATION iris

```
@ATTRIBUTE sepallength NUMERIC  
@ATTRIBUTE sepalwidth NUMERIC  
@ATTRIBUTE petallength NUMERIC  
@ATTRIBUTE petalwidth NUMERIC  
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

@DATA

```
5.1,3.5,1.4,0.2,Iris-setosa  
4.9,3.0,1.4,0.2,Iris-setosa  
4.7,3.2,1.3,0.2,Iris-setosa  
4.6,3.1,1.5,0.2,Iris-setosa  
5.0,3.6,1.4,0.2,Iris-setosa  
5.4,3.9,1.7,0.4,Iris-setosa  
4.6,3.4,1.4,0.3,Iris-setosa  
5.0,3.4,1.5,0.2,Iris-setosa  
4.4,2.9,1.4,0.2,Iris-setosa  
4.9,3.1,1.5,0.1,Iris-setosa
```

# Arquivo .arff

@RELATION iris

```
@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

@DATA

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

## Atributos podem ser:

- numeric
- string
- date
- relational
- nominal

# Arquivo .arff

## Como construir uma base de dados em .arff para classificar peixes {atum/lambari}

@ATTRIBUTE tamanho_cm	REAL
@ATTRIBUTE dorso_espinhoso	{sim,nao}
@ATTRIBUTE cor_dourada	REAL
@ATTRIBUTE cor_carne	string
@ATTRIBUTE tem_espinhos	NUMERIC
@ATTRIBUTE tem_escamas	{sim,nao}
@ATTRIBUTE qtd_olhos	NUMERIC
@ATTRIBUTE tipo_peixe	{atum,lambari}

# Classificação de dados

Tarefa da mineração de dados e aprendizado de máquina:  
Associa dados a uma Classe por uma estratégia construída;

Novos dados são submetidos à mesma estratégia para afirmar a qual classe pertencem (classificar).

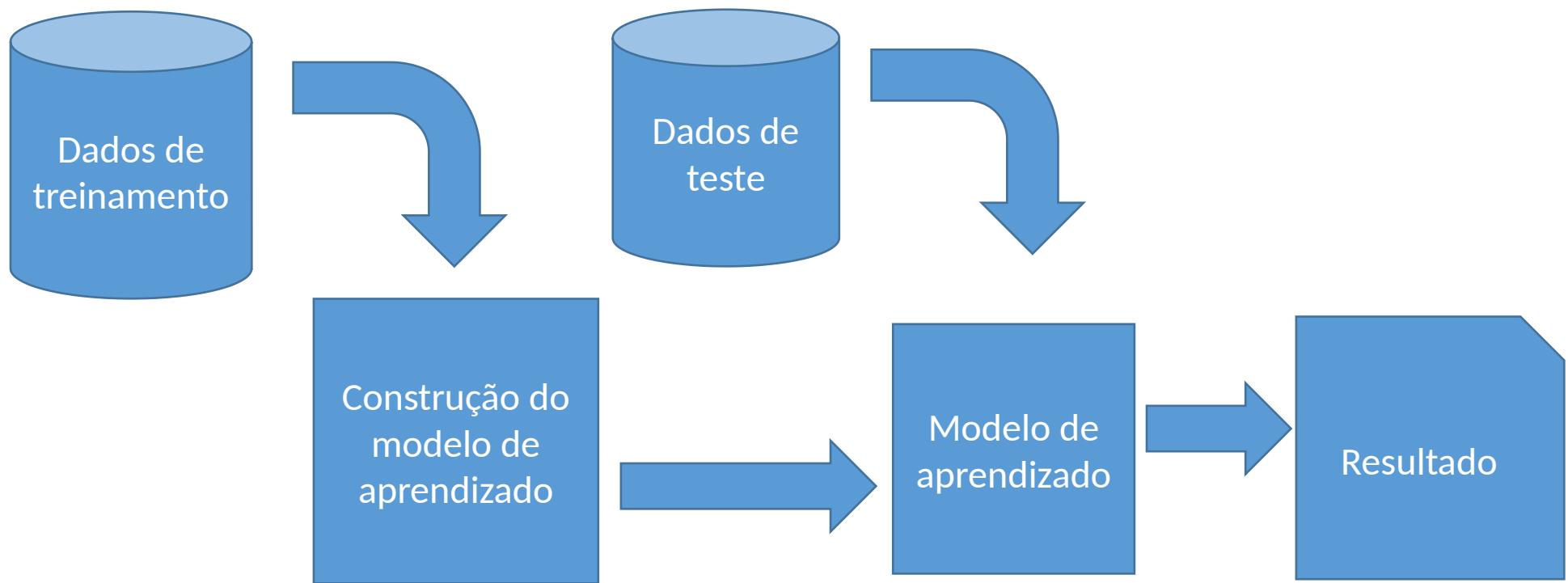


Fonte: Portal Reciclagem

# Classificação de dados

**Geralmente é feita após um treinamento supervisionado.**

Etapa de aprendizagem  
Etapa de execução (teste)



# Classificação de dados - Naive Bayes

Naive = ingênuo:

O nome se dá pelo fato que o algoritmo não supõe relação entre os atributos. - independência das variáveis

- + Simples;
- + Construção e execução Rápidas;
- + Necessita de poucos dados para seu treinamento;
- A grande maioria dos problemas reais tem muita dependência entre variáveis.

# Classificação de dados - Naive Bayes

## Probabilístico

$P(A | B)$ : Probabilidade de A ocorrer sendo que sei a probabilidade de B ocorrer

$P(B | A)$ : Probabilidade de B ocorrer quando A ocorre

$P(A)$ : Probabilidade de A ocorrer

$P(B)$ : Probabilidade de B ocorrer

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

# Classificação de dados - Naive Bayes

aspecto	temperatura	umidade	vento	Bom dia?
sol	quente	alta	fraco	não
sol	quente	alta	forte	não
nublado	quente	alta	fraco	sim
chuva	média	alta	fraco	sim
chuva	frio	normal	fraco	sim
chuva	frio	normal	forte	não
nublado	frio	normal	forte	sim
sol	média	alta	fraco	não
sol	frio	normal	fraco	sim
chuva	média	normal	fraco	sim
sol	média	normal	forte	sim
nublado	média	alta	forte	sim
nublado	quente	normal	franco	sim
chuva	média	alta	forte	não

# Classificação de dados - Naive Bayes

aspecto	temperatura	umidade	vento	Bom dia?
sol	quente	alta	fraco	não
sol	quente	alta	forte	não
nublado	quente	alta	forte	sim
chuva	média	alta	forte	sim
chuva	frio	normal	forte	sim
chuva	frio	normal	fraco	não
nublado	frio	normal	fraco	sim
sol	média	alta	fraco	não
sol	frio	normal	fraco	sim
chuva	média	normal	fraco	sim
sol	média	normal	forte	sim
nublado	média	alta	forte	sim
nublado	quente	normal	franco	sim
chuva	média	alta	forte	não

# Classificação de dados - Naive Bayes

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

aspecto	SIM	NÃO
sol	2	3
nublado		
chuva		
total		

$$5/14 = 0.36$$

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Classificação: Naive Bayes

# Classificação de dados - Naive Bayes

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

aspecto	SIM	NÃO	
sol	2	3	$5/14 = 0.36$
nublado	4	0	$4/14 = 0.29$
chuva	3	2	$5/14 = 0.36$
total	$9/14$	$5/14$	
	0.64	0.36	

Qual a chance de ser um bom dia para jogar, visto que faz sol?

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Classificação: Naive Bayes

# Classificação de dados - Naive Bayes

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

aspecto	SIM	NÃO	
sol	2	3	$5/14 = 0.36$
nublado	4	0	$4/14 = 0.29$
chuva	3	2	$5/14 = 0.36$
total	$9/14$	$5/14$	
	0.64	0.36	

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Qual a chance de ser um bom dia para jogar, visto que faz sol?

$$\begin{aligned} P(\text{SIM} | \text{sol}) &= P(\text{sol} | \text{SIM}) * P(\text{SIM}) / P(\text{sol}) \\ P(\text{SIM} | \text{sol}) &= 2/9 * 9/14 / 5/14 \\ &= 0.22 * 0.64 / 0.36 \\ &= 0.39 \end{aligned}$$

Classificação: Naive Bayes

# Classificação de dados - Naive Bayes

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

aspecto	SIM	NÃO	
sol	2	3	$5/14 = 0.36$
nublado	4	0	$4/14 = 0.29$
chuva	3	2	$5/14 = 0.36$
total	$9/14$	$5/14$	
	0.64	0.36	

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Qual a chance de ser um bom dia para jogar, visto que faz sol?

$$\begin{aligned} P(\text{SIM} | \text{sol}) &= P(\text{sol} | \text{SIM}) * P(\text{SIM}) / P(\text{sol}) \\ P(\text{SIM} | \text{sol}) &= 2/9 * 9/14 / 5/14 \\ &= 0.22 * 0.64 / 0.36 \\ &= 0.39 \end{aligned}$$

39% de probabilidade de ser um bom dia para jogar tênis visto que faz sol

Classificação: Naive Bayes

# Classificação de dados - Naive Bayes

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

temperatura	SIM	NÃO	
quente	2	2	4/14 = 0.28
média			
frio			
total	9	5	= 14

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

$$P(\text{SIM} | \text{frio}) = ????$$

$$P(\text{SIM} | \text{frio}) = P(\text{frio} | \text{SIM}) * P(\text{SIM}) / P(\text{frio})$$

Classificação: Naive Bayes

# Classificação de dados - Naive Bayes

Naive = ingênuo:

O nome se dá pelo fato que o algoritmo não supõe relação entre os atributos. - independência das variáveis

- + Simples;
- + Construção e execução Rápidas;
- + Necessita de poucos dados para seu treinamento;
- A grande maioria dos problemas reais tem muita dependência entre variáveis.

```
import numpy as np
DADOS = np.array([[1.8],[2.1],[1.83],[1.45],[1.52],[1.86]])
CLASSE = np.array(["YES", "YES", "YES", "NO", "NO", "NO"])

from sklearn.naive_bayes import GaussianNB
classificador = GaussianNB()
classificador.fit(DADOS, CLASSE)
print(classificador.predict([[1.42]]))
```

# Classificação de dados - Árvores

- **Raiz:**

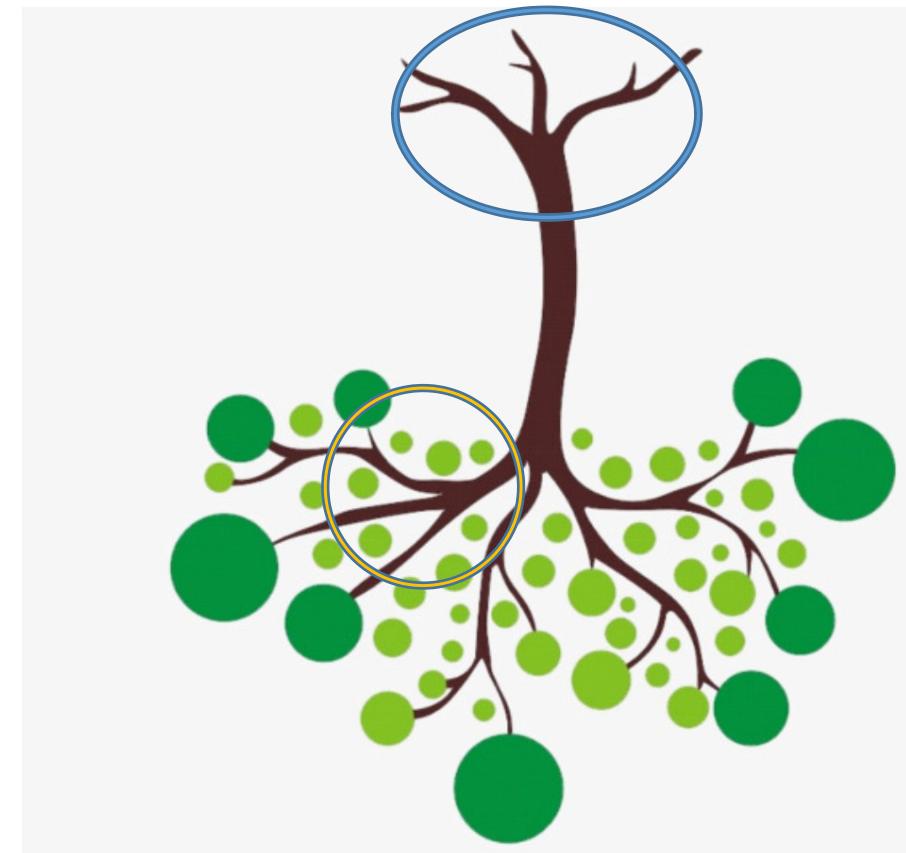
Atributo com maior ganho de informação – maior capacidade de particionar o conjunto;

- **Nós:**

Teste de caminho/decisão;

- **Folhas:**

Classes possíveis.



Fonte: Pixabay free pictures

# Classificação de dados - Árvores

Técnicas mais difundida;

Construção de regras do tipo SE ENTÃO;

- Árvores são os algoritmos mais conhecidos.

**SE** verde **E** amarelo **E** azul **ENTÃO** bandeira do Brasil.

# Classificação de dados - Árvores

Técnicas mais difundida;

Construção de regras do tipo SE ENTÃO;

- Árvores são os algoritmos mais conhecidos.



**SE** verde **E** amarelo **E** azul **ENTÃO** bandeira do Brasil.

Fonte: Soubarato.com.br

# Classificação de dados - Árvores

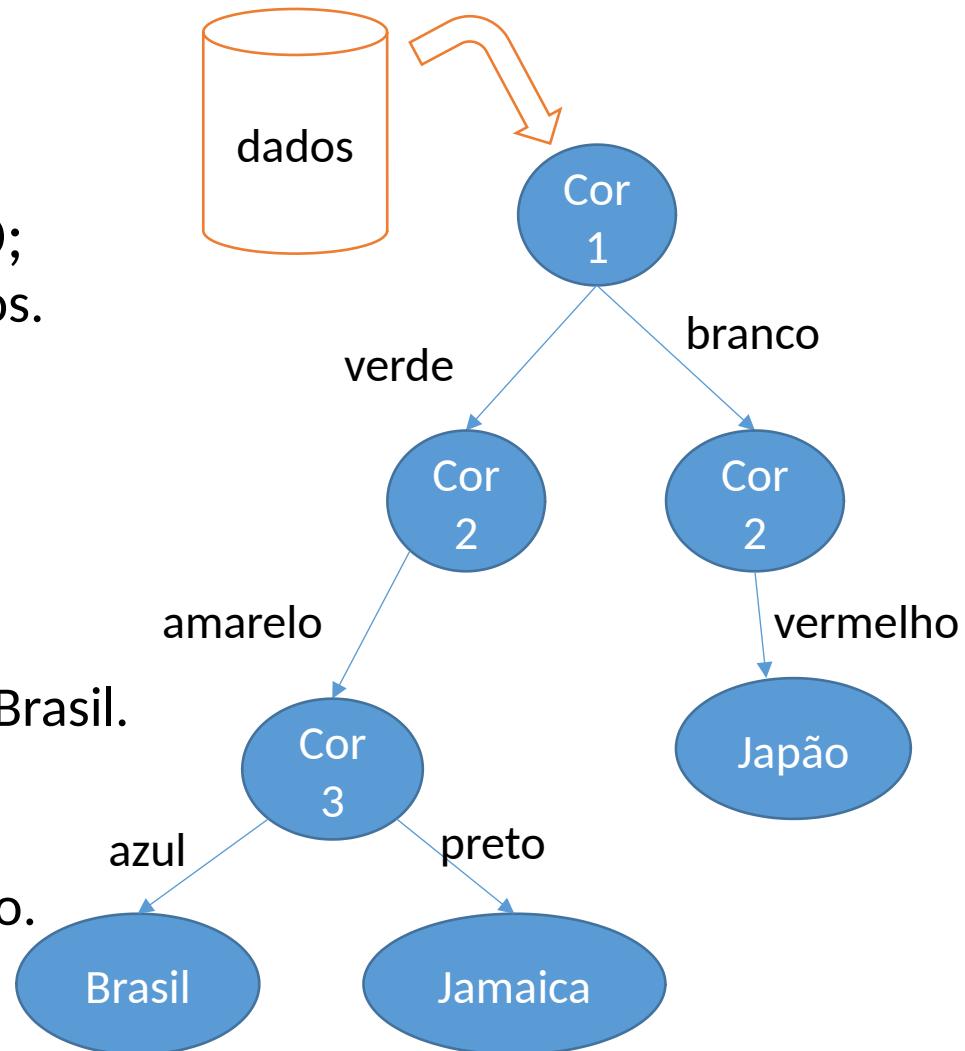
Técnicas mais difundida;

Construção de regras do tipo SE ENTÃO;  
Árvores são os algoritmos mais conhecidos.

**SE** verde **E** amarelo **E** azul **ENTÃO** bandeira do Brasil.

**SE** verde **E** amarelo **E** preto **ENTÃO** bandeira da Jamaica.

**SE** branco **E** vermelho **ENTÃO** bandeira do Japão.



# Classificação de dados - Árvores

Entropia:

O quanto misturados os elementos são;

Proporção de registros para cada classe.

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Ganho de informação:

Qual a força desse atributo para classificar sozinho;

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t)$$

# Classificação de dados - Árvores

aspecto	temperatura	umidade	vento	Bom dia?
sol	quente	alta	fraco	não
sol	quente	alta	forte	não
nublado	quente	alta	fraco	sim
chuva	média	alta	fraco	sim
chuva	frio	normal	fraco	sim
chuva	frio	normal	forte	não
nublado	frio	normal	forte	sim
sol	média	alta	fraco	não
sol	frio	normal	fraco	sim
chuva	média	normal	fraco	sim
sol	média	normal	forte	sim
nublado	média	alta	forte	sim
nublado	quente	normal	franco	sim
chuva	média	alta	forte	não

# Classificação de dados - Árvores

Para cada um dos atributos, calcula a entropia e o ganho de informação;

aspecto	temperatura	umidade	vento	Bom dia?
sol	quente	alta	fraco	não
sol	quente	alta	forte	não
nublado	quente	alta	fraco	sim
chuva	média	alta	fraco	sim
chuva	frio	normal	fraco	sim
chuva	frio	normal	forte	não
nublado	frio	normal	forte	sim
sol	média	alta	fraco	não
sol	frio	normal	fraco	sim
chuva	média	normal	fraco	sim
sol	média	normal	forte	sim
nublado	média	alta	forte	sim
nublado	quente	normal	franco	sim
chuva	média	alta	forte	não

Seleciona o atributo com maior GI para ser a raiz da árvore

# Classificação de dados - Árvores

14 registros no total: 9 sim 5 não

$$\begin{aligned}\text{Entropia inicial} &= - (9/14 * \log_2(9/14) * 9/14) - (5/14 * \log_2(5/14) * 5/14) \\ &= -0.642 * \log_2(0.642) - 0.357 * \log_2(0.357) \\ &= 0.940\end{aligned}$$

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

# Classificação de dados - Árvores

14 registros no total: 9 sim 5 não

$$\begin{aligned}\text{Entropia inicial} &= -(9/14 * \log_2(9/14) * 9/14) - (5/14 * \log_2(5/14) * 5/14) \\ &= -0.642 * \log_2(0.642) - 0.357 * \log_2(0.357) \\ &= 0.940\end{aligned}$$

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Verificar qual o ganho de informação de cada atributo

GI(aspecto) =

GI(temperatura) =

GI(vento) =

GI(umidade) =

# Classificação de dados - Árvores

Entropia inicial = 0.940

Verificar qual o ganho de informação de cada atributo

GI(aspecto) =

GI(temperatura) =

GI(vento) =

GI(umidade) =

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Entropia(Sol):

$$-P(SIM|sol) * \log_2(P(SIM|sol)) - P(NAO|sol) * \log_2(P(NAO|sol))$$

$$-2/5 * \log_2(2/5) - 3/5 * \log_2(3/5) = 0.97$$

aspecto	Bom dia?
sol	não
sol	não
sol	não
sol	sim
sol	sim

# Classificação de dados - Árvores

Entropia inicial = 0.940

Verificar qual o ganho de informação de cada atributo

GI(aspecto) =

GI(temperatura) =

GI(vento) =

GI(umidade) =

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Entropia(Sol) = 0.97

Entropia(nublado):

$$P(\text{SIM}|\text{nublado}) * \log_2(P(\text{SIM}|\text{nublado})) - P(\text{NÃO}|\text{nublado}) * \log_2(P(\text{NÃO}|\text{nublado}))$$

$$-4/4 * \log_2(4/4) - 0/5 * \log_2(0/5) = 0.0$$

aspecto	Bom dia?
nublado	sim

# Classificação de dados - Árvores

Entropia inicial = 0.940

Entropia(Sol) = **0.97**

Entropia(nublado) = **0.0**

Entropia(chuva) = **0.97**

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t)$$

$$GI(\text{aspecto}) = 0.940 - ((5/14) * 0.97) + ((4/14) * 0) + ((5/14) * 0.97)$$

$$GI(\text{aspecto}) = 0.940 - 0.69$$

$$GI(\text{aspecto}) = 0.246$$

aspecto	Bom dia?
sol	não
sol	não
nublado	sim
chuva	sim
chuva	sim
chuva	não
nublado	sim
sol	não
sol	sim
chuva	sim
sol	sim
nublado	sim
nublado	sim
chuva	não

# Classificação de dados - Árvores

GI(vento) = ?

Entropia inicial = 0.940

Entropia(forte) = ?

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Entropia(fraco) = ?

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t)$$

$$GI(\text{vento}) = 0.940 - ( (6/14) * H(\text{forte}) + (8/14) * H(\text{fraco}) )$$

vento	Bom dia?
fraco	não
forte	não
fraco	sim
fraco	sim
fraco	sim
forte	não
forte	sim
fraco	não
fraco	sim
fraco	sim
forte	sim
forte	sim
franco	sim
forte	não

# Classificação de dados - Árvores

**GI(aspecto) = 0.246**

GI(temperatura) = 0.029

GI(vento) = 0.048

GI(vento) = 0.152

Maior ganho de informação implica que

É maior a chance de chegar a uma classificação  
Por meio deste atributo

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t)$$

# Classificação de dados - Árvores

**GI(aspecto) = 0.246**

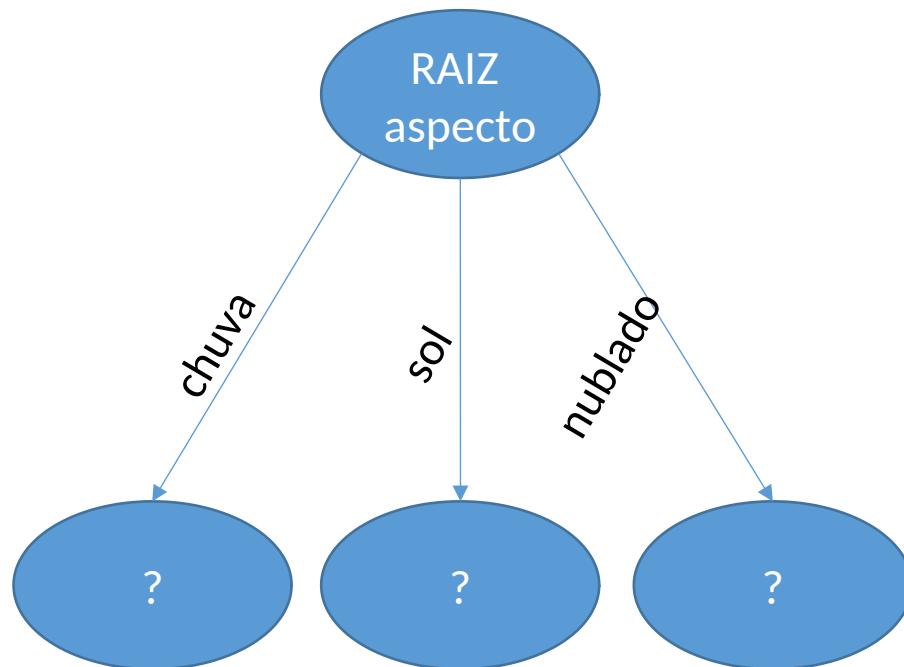
GI(temperatura) = 0.029

GI(vento) = 0.048

GI(vento) = 0.152

Maior ganho de informação implica que

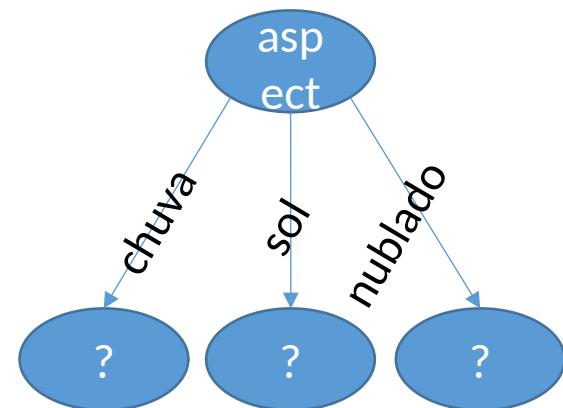
É maior a chance de chegar a uma classificação  
Por meio deste atributo



# Classificação de dados - Árvores

aspecto	temperatura	umidade	vento	Bom dia?
chuva	média	alta	fraco	sim
chuva	frio	normal	fraco	sim
chuva	frio	normal	forte	não
chuva	média	normal	fraco	sim
chuva	média	alta	forte	não

O processo se repete para cada possibilidade do atributo raiz (com maior ganho de informação)

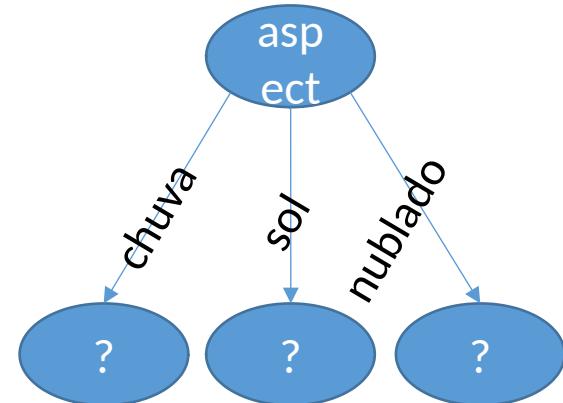


# Classificação de dados - Árvores

aspecto	temperatura	umidade	vento	Bom dia?
chuva	média	alta	fraco	sim
chuva	frio	normal	fraco	sim
chuva	frio	normal	forte	não
chuva	média	normal	fraco	sim
chuva	média	alta	forte	não

O processo se repete para cada possibilidade do atributo raiz (com maior ganho de informação)

**Vamos ver o resultado no WEKA**



# Arquivo .arff

## Como construir uma base de dados em .arff para classificar peixes {atum/lambari}

@ATTRIBUTE tamanho_cm	REAL
@ATTRIBUTE dorso_espinhoso	{sim,nao}
@ATTRIBUTE cor_dourada	REAL
@ATTRIBUTE cor_carne	string
@ATTRIBUTE tem_espinhos	NUMERIC
@ATTRIBUTE tem_escamas	{sim,nao}
@ATTRIBUTE qtd_olhos	NUMERIC
@ATTRIBUTE tipo_peixe	{atum,lambari}

Por que criar um .arff se o Weka aceita .csv?

Algoritmos como C4.5  
não lidam com a

# Atividade de classificação

**Com a base de dados construída para descrever os Simpsons**

**Utilize os algoritmos de classificação para identificar se um indivíduo do desenho é/ou não membro da família Simpson**

Existe uma probabilidade da base de dados precisar de mudanças ainda?

# Questões de projeto

**Qual o intuito disso? Responder questões como:**

- Qual o melhor algoritmo para o meu conjunto de dados?
- Quanto de treinamento é necessário?
- Qual a melhor estratégia para chegar ao resultado com menos treinamento?
- É possível automatizar o treinamento?

# No próximo encontro

## **Outros algoritmos de classificação**

### **Treinamento não supervisionado:**

- algoritmos de agrupamento

# Atividade bônus

## Processar a base de dados dengue