

CURSO BÁSICO DE
INTELIGÊNCIA
ARTIFICIAL E
BATE-PAPO COM
CONVIDADOS
ESPECIAIS

INTELIGÊNCIA ARTIFICIAL PARA TODOS

DE 08/06 A 12/08



COM OS PROFESSORES DO
LABORATÓRIO ARIA/UFPB:
TELMO FILHO, THAÍS
GAUDENCIO E YURI
MALHEIROS



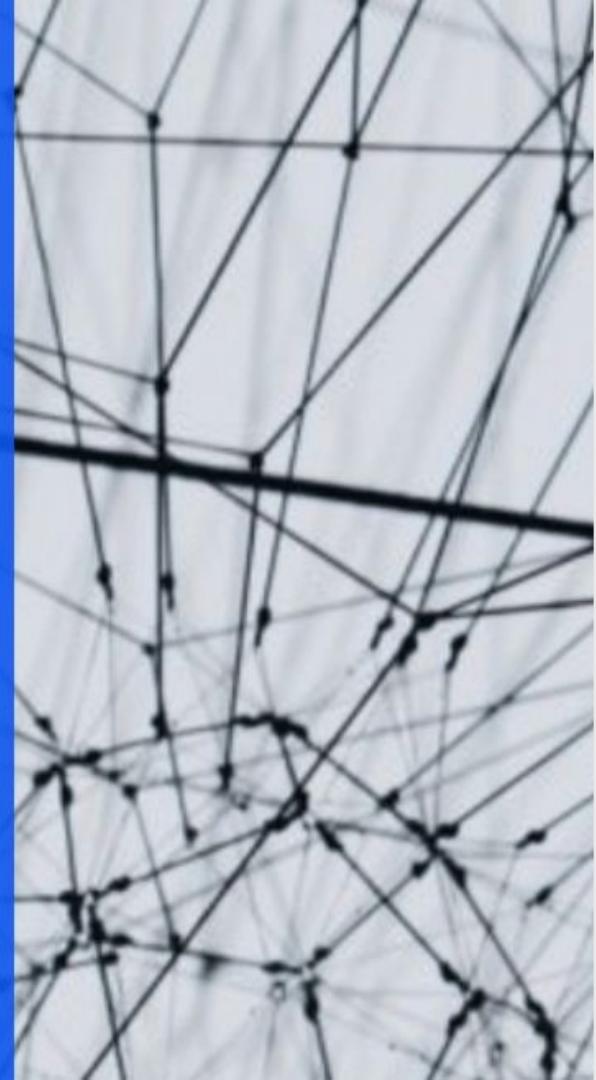
Centro de Informática

UFPB



artificial
intelligence
applications

- CURSO SEM PRÉ-REQUISITOS
- [HTTP://ARIA.CI.UFPB.BR/IAPARATODOS](http://ARIA.CI.UFPB.BR/IAPARATODOS)
- INSCRIÇÃO PARA CERTIFICADO - DE 01/06
A 07/06: [HTTP://BIT.LY/SIGEVENTOS](http://BIT.LY/SIGEVENTOS)
- ENCONTROS: SEGUNDAS E QUARTAS
- HORÁRIO: 19:00 ÀS 20:00





[Início](#) [Sobre](#) [Projetos](#) [Membros](#) [Parceiros](#) [Publicações](#) [Contato](#)

LABORATÓRIO DE APLICAÇÕES EM INTELIGÊNCIA ARTIFICIAL

As experiências definem a aprendizagem. Assim, o ARIA constrói experiências para máquinas e para pessoas, formando especialistas na área de inteligência artificial e ciência de dados, desenvolvendo aplicações e pesquisando seus métodos.

[SAIBA MAIS](#)

SE INSCREVE E JÁ APERTA NO SININHO, QUE VOCÊS PASSAM A RECEBER AS NOTIFICAÇÕES.

NOSSOS ENCONTROS DURARÃO 1 HORA E, ASSIM QUE POSSÍVEL, DEIXAREMOS OS VÍDEOS GRAVADOS NO CANAL.

NÃO PRECISA SE PREOCUPAR EM ESTAR LIGADO ÀS 19:00, MAS ESTANDO, ROLA TIRAR DÚVIDA E PARTICIPAR, O QUE JÁ DEIXA A AULA MAIS ANIMADA.

**SOBRE O MATERIAL DE ACOMPANHAMENTO: O ALUNO PRECISA SE LOGAR EM:
CLASSROOM.GOOGLE.COM**

DEPOIS, CLICAR EM PARTICIPAR DA TURMA (ÍCONE COM UM MAIS +)

POR FIM, ENTRAR COM O CÓDIGO DA TURMA: PXV3ANW

TAMBÉM EM: [HTTPS://ARIA.CI.UFPB.BR/IA-PARA-TODOS-MATERIAL/](https://aria.ci.ufpb.br/ia-para-todos-material/)

ESPERAMOS QUE VOCÊS REALMENTE CURTAM O CURSO E APROVEITEM-NO AO MÁXIMO. NÃO DEIXEM DE INTERAGIR CONOSCO, TAMBÉM, POR E-MAIL OU MENSAGEM NO NOSSO INSTAGRAM (@APRENDIZAGEMDEMAQUINA)



Bases de
dados do
mundo real

IRIS

Hoje é o dia mais
importante da tua
vida



Pré-processamento de Dados



Thaís Gaudencio do Rêgo

Existe uma estimativa de que a cada 20 meses dobrar a quantidade de dados armazenada nos bancos de dados do mundo!

No entanto, tem aumentado também a distância entre a quantidade de dados existente e a porção deles que é analisada e compreendida.

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b bit	0 or 1	1/8 of a byte
B byte	8 bits	1 byte
KB kilobyte	1,000 bytes	1,000 bytes
MB megabyte	1,000 ² bytes	1,000,000 bytes
GB gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

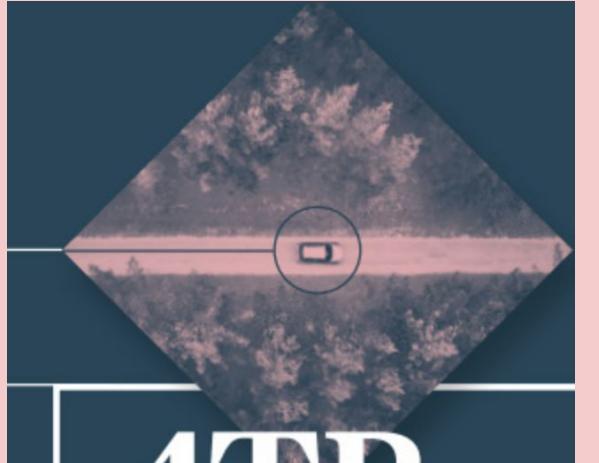
IDC

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

PwC



4TB

of data produced by a connected car

Intel



4PB

of data created by
Facebook, including

350m photos

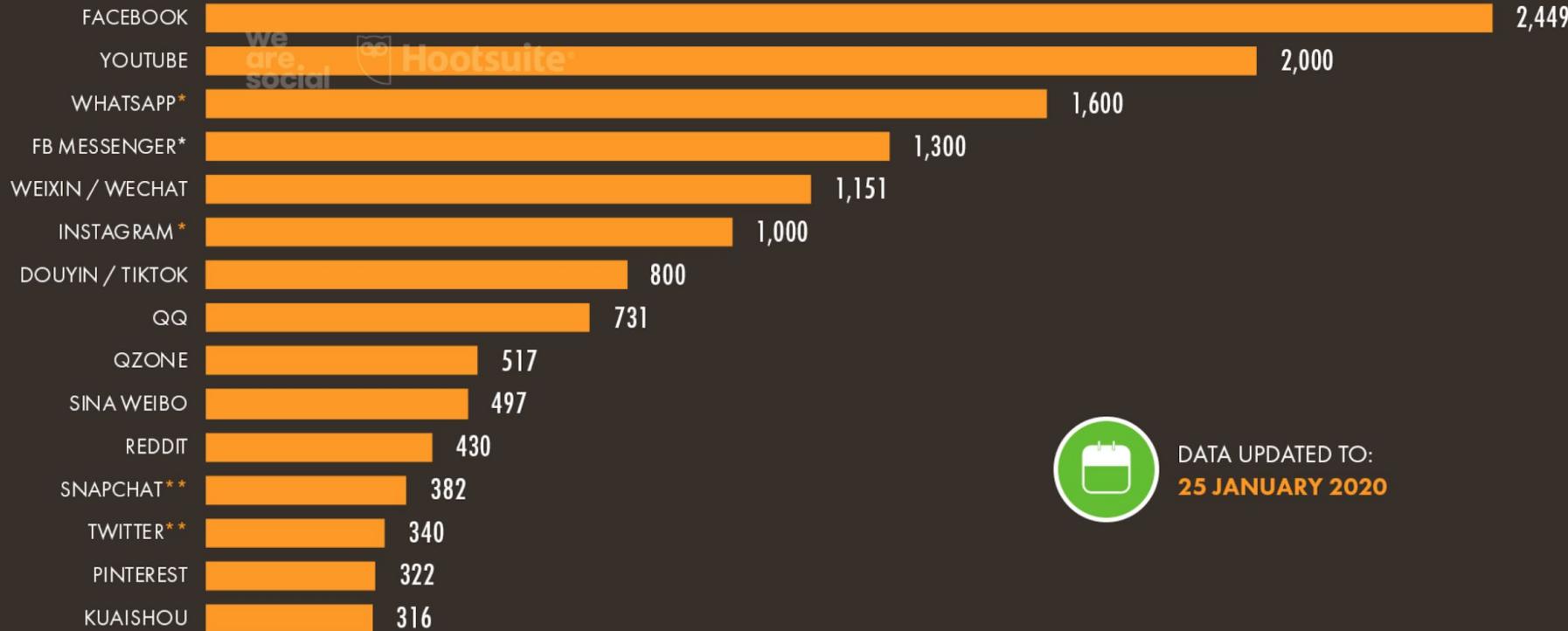
100m hours of video
watch time

Facebook Research

JAN
2020

THE WORLD'S MOST-USED SOCIAL PLATFORMS

BASED ON MONTHLY ACTIVE USERS, ACTIVE USER ACCOUNTS, ADVERTISING AUDIENCES, OR UNIQUE MONTHLY VISITORS (IN MILLIONS)



DATA UPDATED TO:
25 JANUARY 2020

SOURCES: KEPiOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS (ALL LATEST AVAILABLE DATA). **NOTES:** PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER NUMBERS IN THE PAST 12 MONTHS. PLATFORMS IDENTIFIED BY (**) DO NOT PUBLISH MAU DATA. FIGURES FOR TWITTER AND SNAPCHAT USE EACH PLATFORM'S LATEST ADVERTISING AUDIENCE REACH, AS REPORTED IN EACH PLATFORM'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020).

Objetos que representam objetos físicos ou uma noção abstrata (como sintomas);

Cada objeto é descrito por um conjunto de atributos de entrada ou vetor de características;

Cada objeto corresponde a uma ocorrência dos dados;

Cada atributo está associado a uma propriedade do objeto.

```
df = pd.read_csv('hospital1.csv', delimiter=";")
```

```
df
```

	Identificador	Nome	Idade	Sexo	Peso	Manchas	Temperatura	Internacoes	Estado	Diagnostico
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	38.0	3	RJ	Doente
6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Conjunto de Dados

Os dados podem ser representados por uma matriz de objetos $X \in \mathbb{R}^{n \times d}$, em que n é o número de objetos e d é o número de atributos de entrada de cada objeto;

O valor de d define a dimensionalidade dos objetos ou do espaço dos objetos.

Conjuntos de Treinamento e Teste

Identificador	Nome	Idade	Sexo	Peso	Manchas	Temperatura	Internacoes	Estado	Diagnóstico	
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel

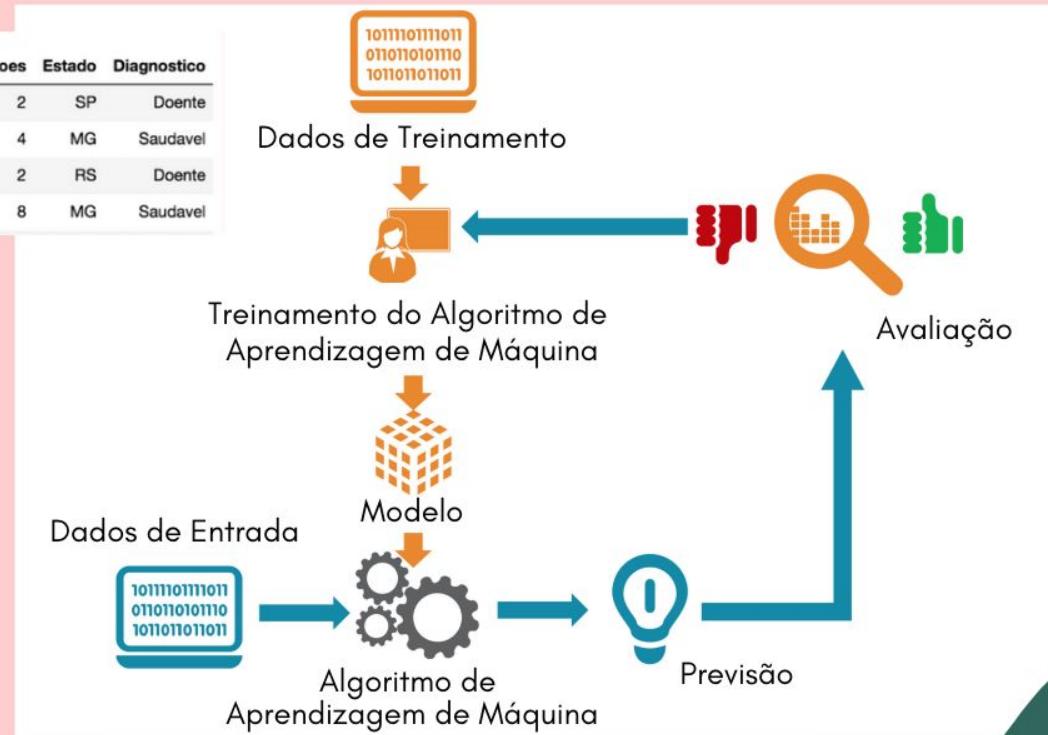
Conjunto Treinamento (supervisionado):

Atributos de entrada e Saída para a construção do modelo

Conjunto Teste (Supervisionado):

Atributos de entradas

Atributos de Saída:
conhecidos, não
apresentados à
máquina, usados para
a avaliação do modelo



Fonte: <https://medium.com/@tekaround/train-validation-test-set-in-machine-learning-how-to-understand-6cdd98d4a764>

Análise de Dados

- Conjunto: hospital
- Objeto: paciente
- Atributos de entrada: características do paciente ou valores de exames

Aprendizagem **SUPERVISIONADA!**

- Atributo de saída (alvo): valor que se deseja prever

Pré-processamento

- Eliminação manual de atributos;
- Integração de dados;
- Amostragem de dados;
- Balanceamento de dados;
- Limpeza de dados;
- Redução de dimensionalidade;
- Transformação de dados.

Exploração de Dados

Estatística Descritiva: resume de forma quantitativa as principais características de um conjunto de dados.

Exemplo:

- Idade média dos pacientes
- Porcentagem de pacientes do sexo masculino

```
df.mean()
```

```
identificador      3045.875
idade              26.125
peso               69.875
temperatura        38.375
internacoes        3.500
dtype: float64
```

```
df['temperatura'].mean()
```

```
38.375
```

```
x = df['sexo'].value_counts()
x/len(df)
```

```
M      0.5
F      0.5
```

Exemplo:

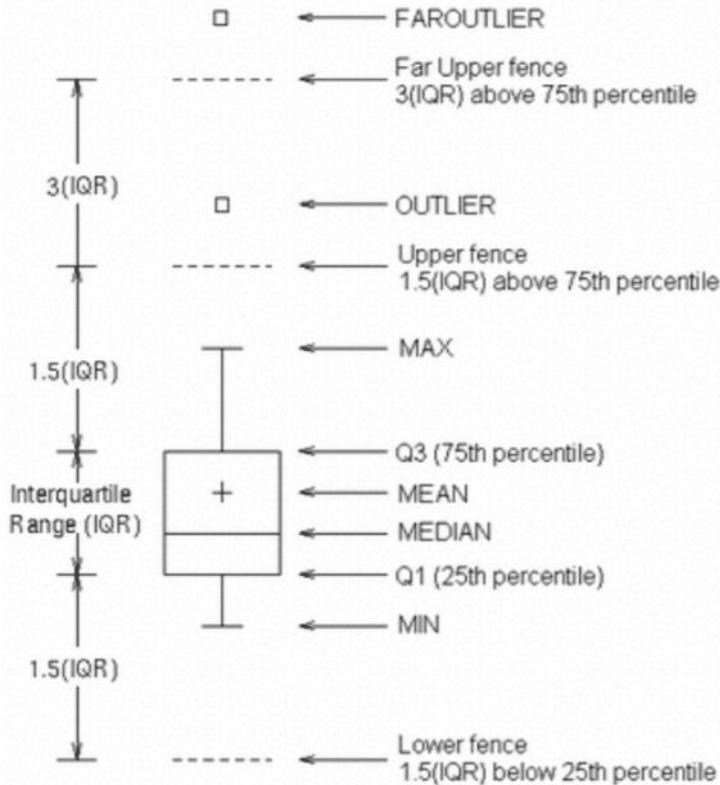
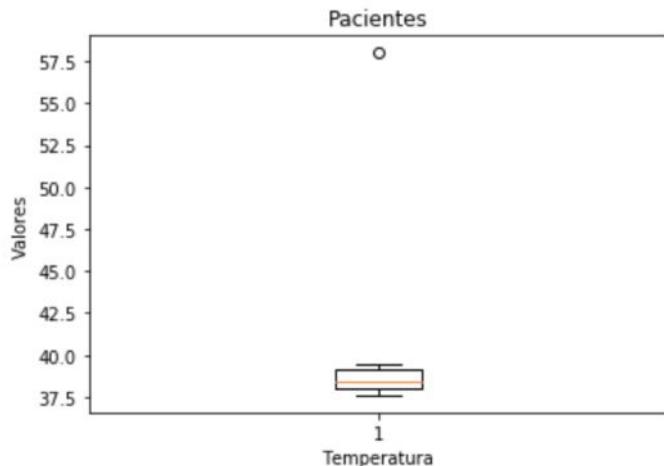
- Idade média dos pacientes
- Porcentagem de pacientes do sexo masculino

De onde pode se calcular a porcentagem!

```

import matplotlib.pyplot as plt
|
plt.boxplot(df['temperatura'])
|
plt.title('Pacientes')
plt.xlabel('Temperatura')
plt.ylabel('Valores')
|
Text(0, 0.5, 'Valores')

```



```
v = df["temperatura"].var() #variância do atributo "temperatura"  
d = df["temperatura"].std() #desvio padrão do atributo "temperatura"  
  
print(v)  
print(d)
```

48.24214285714286
6.9456564021799165

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

N = população do total

n - 1 = amostra da população

Dados Multivariados

Relação entre dois ou mais atributos.

Covariância = mede o grau com que os atributos variam juntos.

$$COV_{x,y} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]$$

$$r_{xy} = \frac{Cov(X, Y)}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

em que \bar{x} e \bar{y} são as médias amostrais de X e Y .

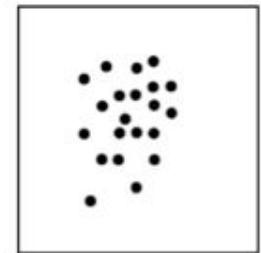
Correlação



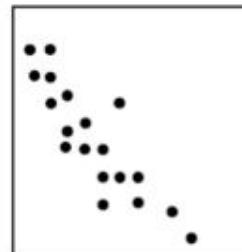
Strong positive correlation



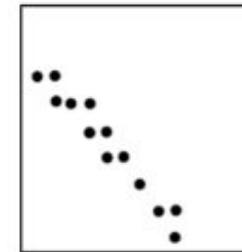
Moderate positive correlation



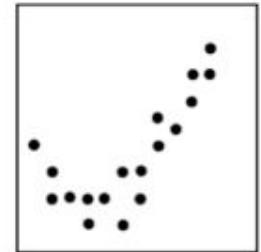
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

Correlação

```
corr = df.corr()  
corr.style.background_gradient(cmap='coolwarm')
```

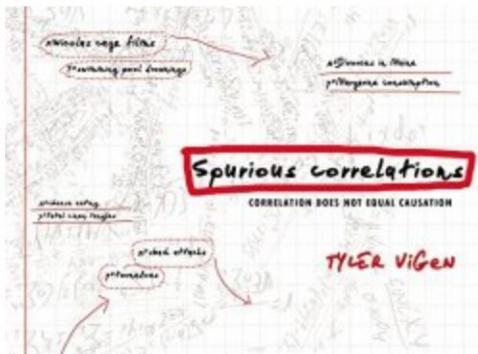
	identificador	idade	peso	temperatura	internacoes
identificador	1	0.48618	0.0620799	-0.319808	-0.817134
idade	0.48618	1	0.560835	-0.191917	-0.512349
peso	0.0620799	0.560835	1	0.0593191	-0.28261
temperatura	-0.319808	-0.191917	0.0593191	1	-0.035277
internacoes	-0.817134	-0.512349	-0.28261	-0.035277	1

<https://www.tylervigen.com/spurious-correlations>

[tylervigen.com](https://www.tylervigen.com)

[about](#) | [twitter](#) | [email](#) | [subscribe](#)

Spurious correlations



Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

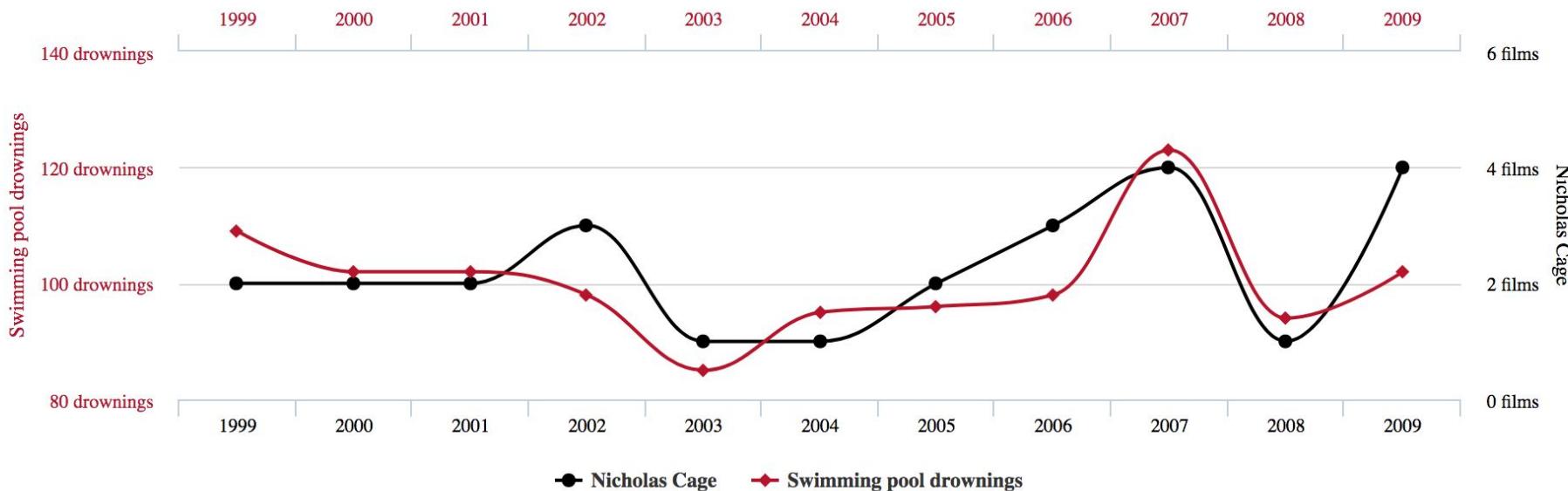
[Amazon](#) | [Barnes & Noble](#) | [Indie Bound](#)

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

tylervigen.com

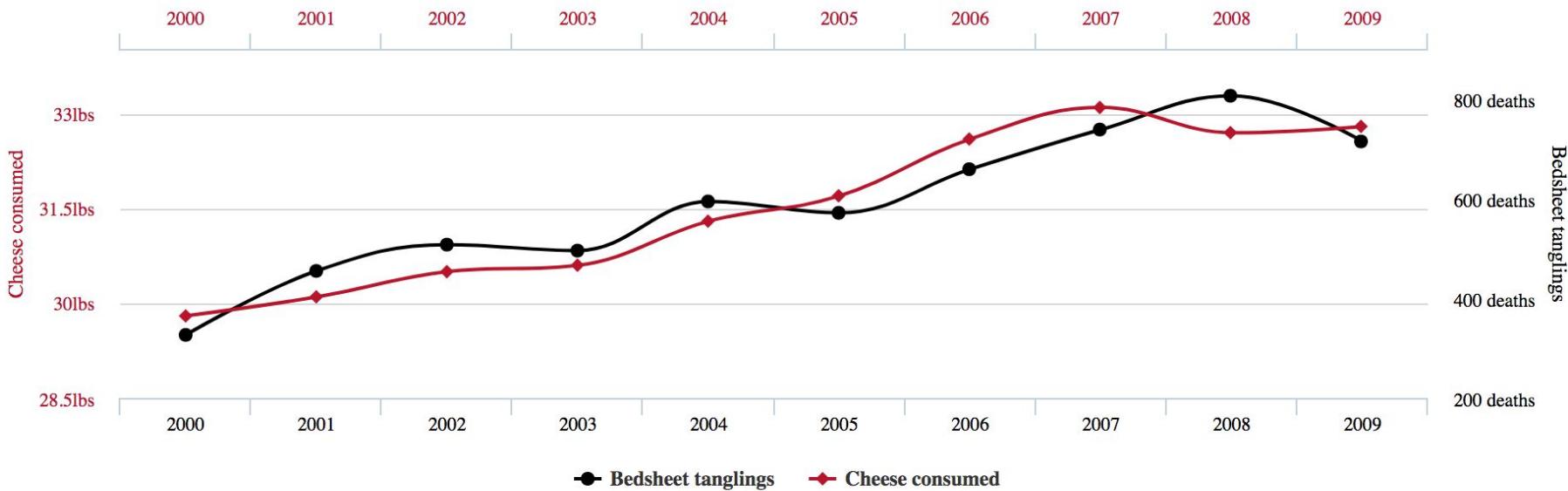


Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% ($r=0.947091$)



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

O GRANDE DIA



ELIMINAÇÃO MANUAL DE ATRIBUTOS

Quando um atributo não contribui para a estimativa do valor do atributo alvo, ele é considerado irrelevante.

O conjunto de dados final deve ser definido de acordo com a experiência de especialistas no domínio dos dados.

PRÉ- PROCESSAMENTO

- **Técnicas utilizadas para melhorar a qualidade dos dados;**
- **Eliminam ou minimizam os problemas como ruídos, outliers, valores incorretos, inconsistentes, duplicados ou ausentes.**
- **Podem ainda tornar os dados mais adequados para sua utilização por um determinado algoritmo.**

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("hospital.csv", sep = ';')

df = df.drop(columns=['identificador', 'nome'])

df

```

	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico								
0	28	M	79	Concentradas	38.0	2	SP	Doente								
1	18	F	67	Inexistentes	39.5	4	MG	Saudavel								
2	49	M	92	Espalhadas	38.0	Identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico	
3	18	M	43	Inexistentes	38.5	0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
4	21	F	52	Uniformes	37.6	1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
5	22	F	72	Inexistentes	58.0	2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
6	19	F	87	Espalhadas	39.0	3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
7	34	M	67	Uniformes	38.4	4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
						5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
						6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
						7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

INTEGRAÇÃO DE DADOS

- **Busca por atributos comuns nos conjuntos a serem combinados;**
- **Atributos utilizados para combinação deve(m) ter um valor único para cada objeto.**
- **CUIDADO: nome do atributo e atualização dos dados.**

AMOSTRAGEM DE DADOS

Algoritmos de AM podem ter dificuldade em lidar com um grande número de objetos.

Balanço entre eficiência computacional e acurácia (taxa de previsões corretas).

- Maior acurácia x menor eficiência computacional.

PROBLEMA: uma amostra pode não representar bem o problema que se deseja modelar.



AMOSTRAGEM DE DADOS

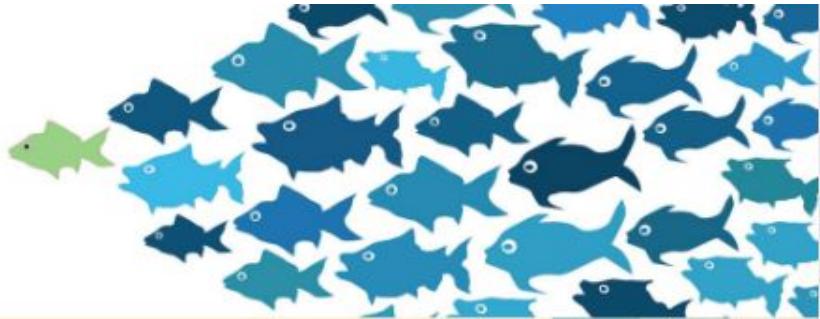
A amostra deve ser representativa do conjunto de dados original.

Diferentes amostras de uma mesma população podem gerar modelos diferentes.

Os dados devem obedecer a mesma distribuição estatística que gerou o conjunto de dados original.

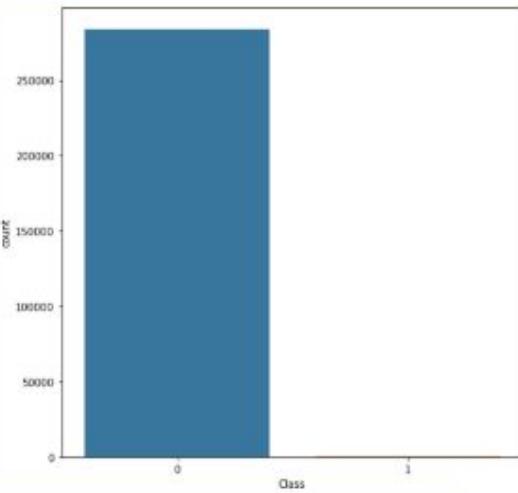


DADOS DESBALANCEADOS

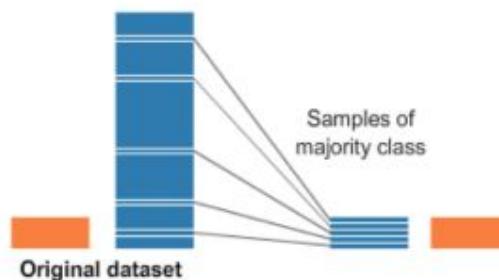


- Técnicas de balanceamento artificial:
 - Redefinir o tamanho do conjunto de dados
 - Utilizar diferentes custos de classificação para as diferentes classes
 - Induzir um modelo para uma classe

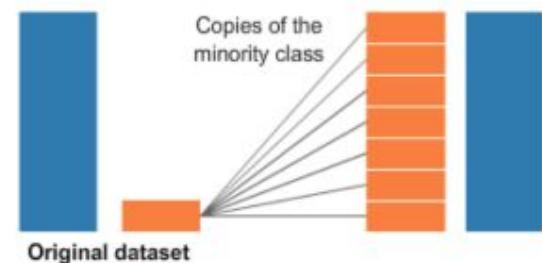
DADOS DESBALANCEADOS



Undersampling



Oversampling



DADOS DESBALANCEADOS

```
#Randomicamente seleciona 4 instâncias a partir da classe 'Doente'  
df.loc[df['diagnostico'] == "Doente"].sample(n=4,random_state=2)
```

identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico	
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente

As instâncias criadas comumente consideram valores de média +/- desvio padrão



Garanta que as instâncias criadas fazem sentido em relação ao esperado. Por exemplo: Uma população de 23,5 pessoas

Lembre sempre de manter a estatística descritiva da base de dados original!!

https://imbalanced-learn.readthedocs.io/en/stable/over_sampling.html#smote-adasyn

LIMPEZA DE DADOS

Problemas relacionados a qualidade dos dados:

- Dados ruidosos: possuem erros ou valores que são diferentes do esperado



DIFERENTE DE OUTLIER, que é um dado real!

- Inconsistente: não combinam ou contradizem valores de outros atributos do mesmo objeto



LIMPEZA DE DADOS

Problemas relacionados a qualidade dos dados:

- Redundantes (quando dois ou mais objetos têm os mesmos valores para todos os atributos ou dois ou mais atributos tem os mesmos valores para dois ou mais objetos)
- Incompletos (com ausência de valores para alguns dos atributos em parte dos dados)





LIMPEZA DE DADOS

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	NaN	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67.0	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92.0	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43.0	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52.0	Uniformes	NaN	1	PE	Doente
5	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
6	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
8	1322	Marta	19	F	87.0	Espalhadas	39.0	6	AM	Saudavel
9	3027	Paulo	34	M	67.0	Uniformes	38.4	2	GO	Saudavel

DADOS INCOMPLETOS

Eliminar os objetos com valores ausentes.

Definir e preencher manualmente valores para os atributos com valores ausentes.

Empregar algoritmos de AM que lidam internamente com valores ausentes (algoritmos indutores de árvores de decisão)

Utilizar algum método ou heurística para automaticamente definir valores para os atributos com valores ausentes.

DADOS INCOMPLETOS

Criar para o atributo um novo valor que indique que o atributo possuía um valor desconhecido.

Utilizar a média, moda ou mediana dos valores conhecidos para esse atributo.

Empregar um indutor para estimar o valor do atributo. É a mais popular!!!

- Utilização do valor utilizado em objetos semelhantes!!

DADOS INCONSISTENTES

Possibilidades!!!

- Problema na anotação dos dados!

- Os atributos de entrada não explicam o atributo alvo!

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	67	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Lulz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Doente
5	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Saudavel
6	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
8	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
9	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
10	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

REDUNDÂNCIA DE DADOS



- **Boosting = duplica-se a quantidade de exemplos difíceis de serem classificados.**
- **Redundância de atributos (idade x data de nascimento).**
- **Alta correlação entre atributos.**
OS ATRIBUTOS TRAZEM A MESMA INFORMAÇÃO EM RELAÇÃO AO ATRIBUTO ALVO - MANTENHA APENAS UM!!!

Dados que contêm objetos que, aparentemente, não pertencem a distribuição que gerou os dados analisados.

- São identificados como observações que diferem de uma distribuição utilizada na modelagem dos dados.
 - São identificados como objetos pertencentes a níveis superficiais.



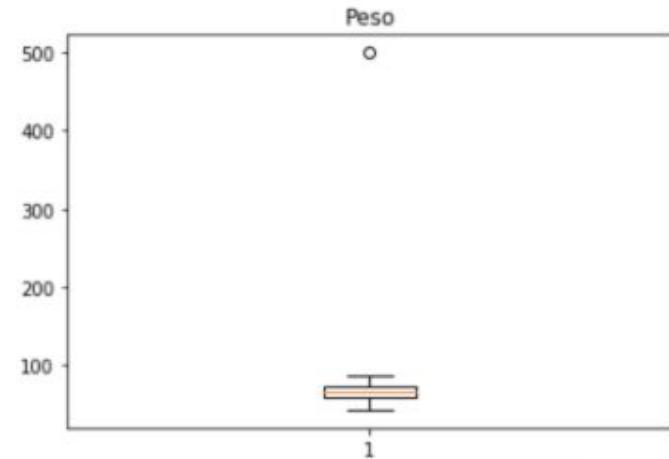
Esses dados, se forem reais, devem ser estudados, sabendo que existe a possibilidade de outros casos semelhantes surgirem



Não sendo reais, devem ser recuperados, eliminados ou estimados!!

```
df = pd.read_csv("hospital.csv", sep = ';')  
plt.boxplot(df['peso'])  
plt.title('Peso')
```

```
Text(0.5, 1.0, 'Peso')
```



TRANSFORMAÇÃO DE DADOS

- Métodos de discretização permitem transformar atributos quantitativos em qualitativos.
- Os valores numéricos são transformados em intervalos ou categorias.

TRANSFORMAÇÃO DE ATRIBUTOS NUMÉRICOS

Quando os limites inferior e superior de valores dos atributos são muito diferentes ou estão em escalas diferentes.



Normalização = evita que um atributo predomine sobre outro, mas não existe garantia.

- Amplitude
- Distribuição

NORMALIZAÇÃO

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Por reescala =

define uma nova escala de valores, limites mínimo e máximo, para todos os atributos.

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Padronização =

define um valor central e um valor de espalhamento comum para todos os atributos.

- Lida melhor com outliers.

