

MODEL LIFECYCLE



CONCEITOS BÁSICOS DE MACHINE LEARNING II

DIEGO RODRIGUES DSC

INFNET

MODEL LIFECYCLE : CONCEITOS BÁSICOS DE MACHINE LEARNING II

PARTE 1 : TEORIA

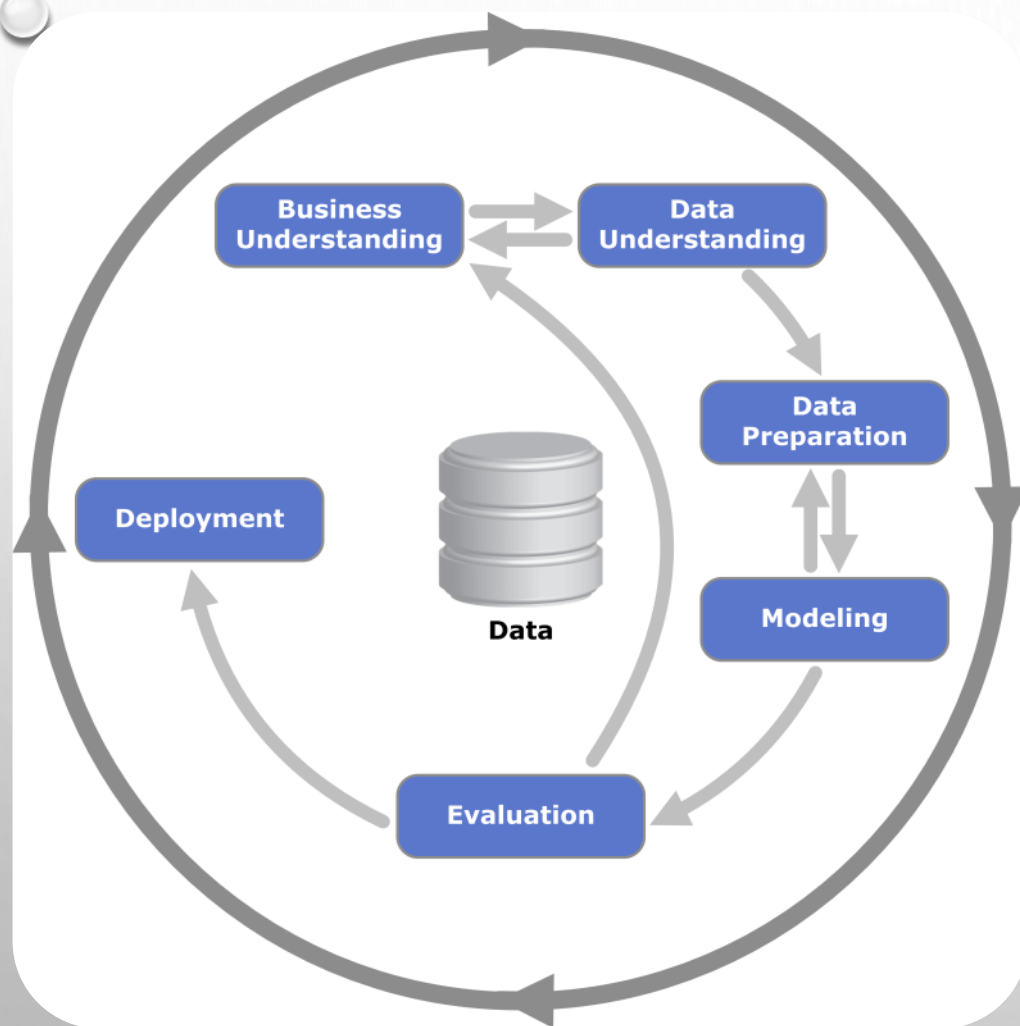
- BUSINESS UNDERSTANDING
- DATA UNDERSTANDING
 - COLETA DE DADOS
 - ANÁLISE EXPLORATÓRIA
- DATA PREPARATION
- MODELING
 - SELEÇÃO DE ATRIBUTOS
 - SELEÇÃO DO MODELO
 - CLASSIFICAÇÃO
 - REGRESSÃO
- EVALUATION

Produzir Ação

CICLO DE VIDA DO MODELO

Baseado em Dados

Cross Industry Standard Process for Data Mining - IBM



1) **Requerimentos e Análise de Negócio**

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) **Preparação dos Dados**

Entendimento das fontes de dados, dos tipos e elaboração da representação.

3) **Modelagem**

Análise Exploratória, Seleção de atributos e treinamento.

4) **Avaliação**

Seleção do melhor modelo.

5) **Liberação**

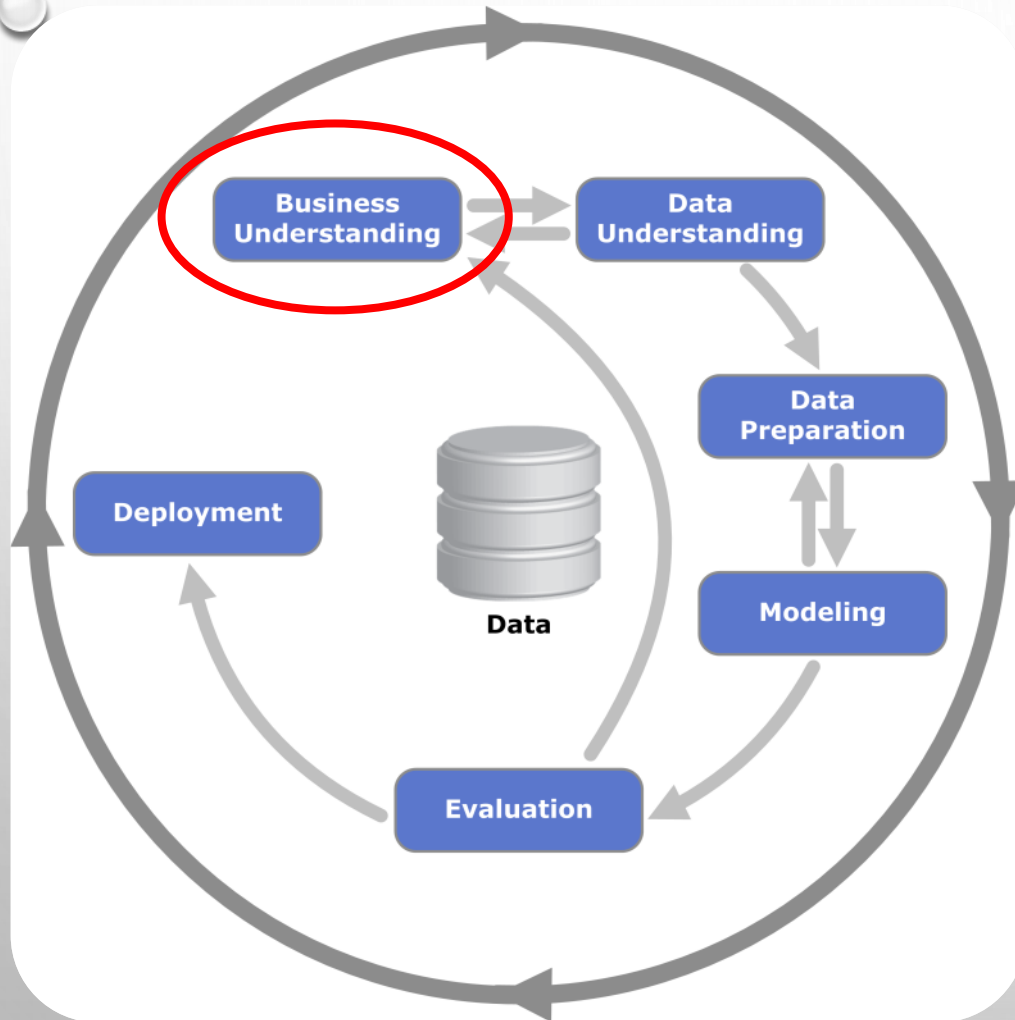
Liberação do modelo no ambiente de produção.

The image features a light gray background with a subtle gradient. In the top-left and bottom-right corners, there are clusters of realistic water droplets of various sizes, some overlapping. A faint, circular, embossed-like pattern is visible in the upper center of the page.

BUSINESS UNDERSTANDING

CRISP-DM

Cross Industry Standard Process for Data Mining - IBM



1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Entendimento das fontes de dados, dos tipos, análise exploratória e representação.

3) Modelagem

Seleção, extração de atributos e treinamento do modelo.

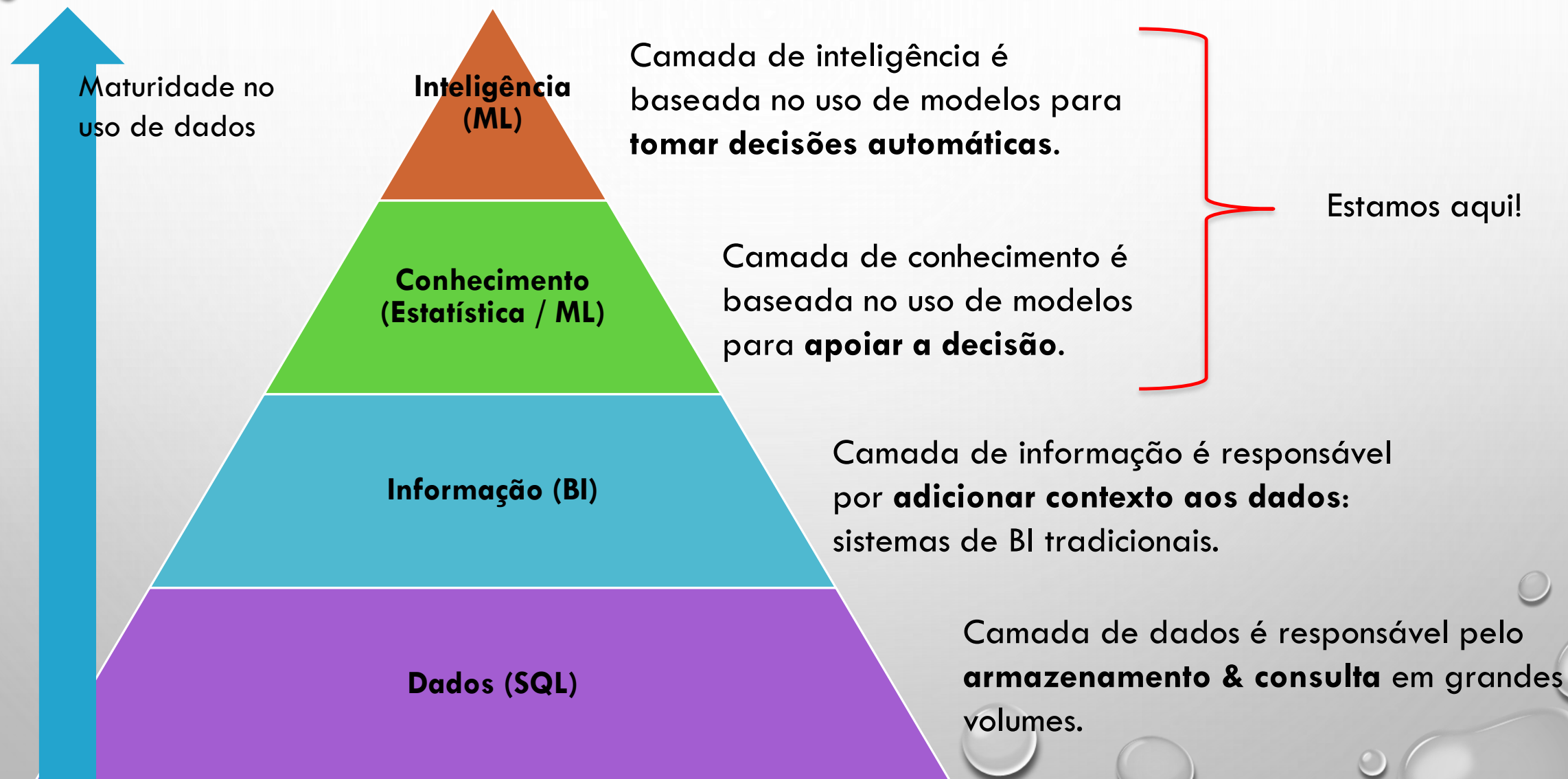
4) Avaliação

Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

Big Data, Business Intelligence, Analytics, Data Science...



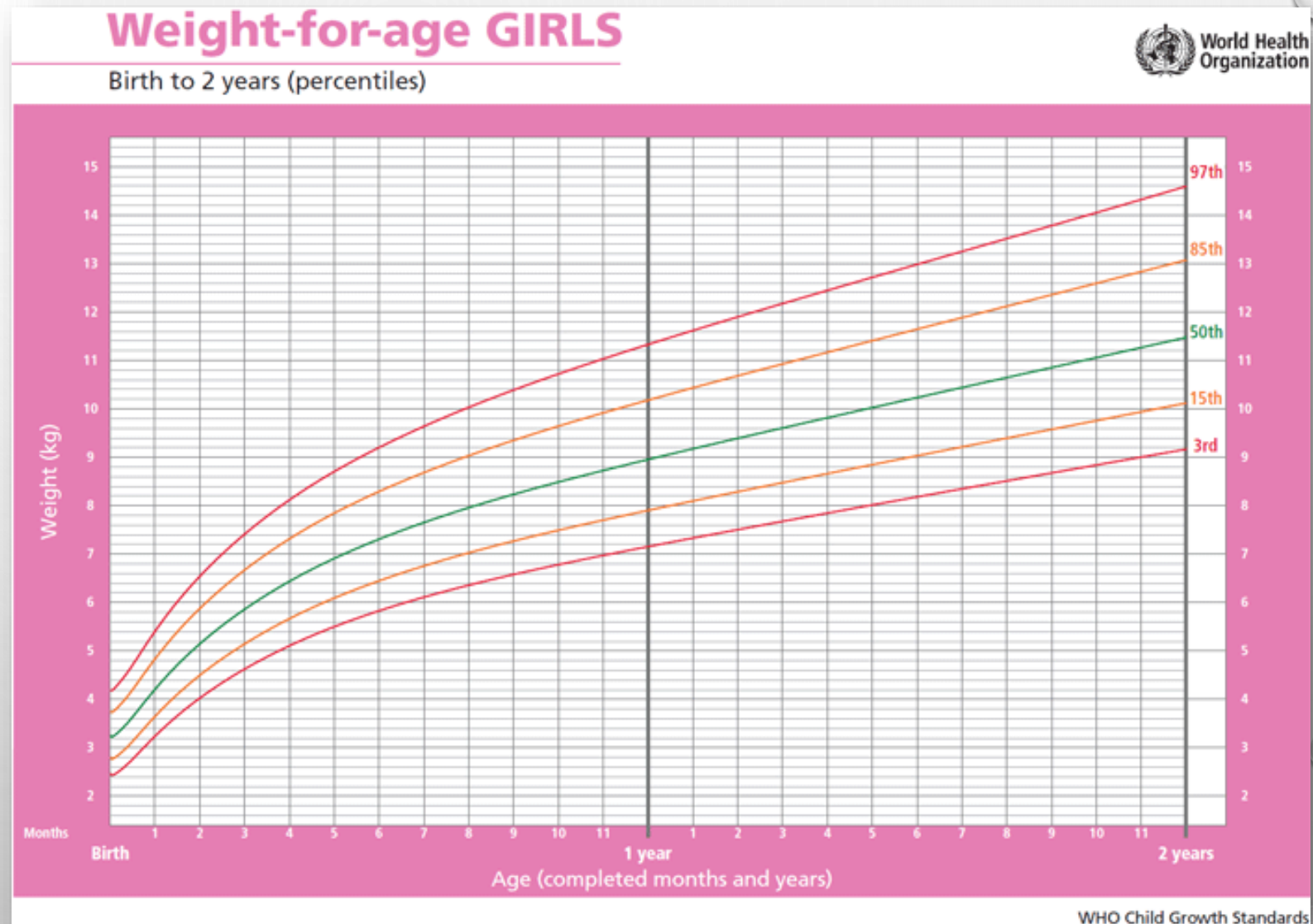
Dado, Informação, Conhecimento, Automação

Dado: 4.300

Informação: Peso em gramas de um neném do sexo feminino com 2 meses de idade.

Conhecimento: Modelo Percentil relacionando milhares de nenéns por idade / Peso.

Automação: Modelo de Previsão de Peso para o próximo mês & se o neném deve suplementar.



FRAMEWORK PARA A ETAPA DE BUSINESS UNDERSTANDING

<Alguém> toma uma decisão que envolve um determinado **<Risco>** utilizando um conjunto de **<Dados>** numa determinada **<Frequência>**.

Descrição do Problema de Negócio

- Um Parágrafo descrevendo a necessidade do usuário, a decisão a ser tomada e os dados disponíveis.

Persona & Usuário

- Quem utilizará o sistema / modelo para tomada de decisão?

Riscos envolvidos na decisão

- Quais riscos envolvidos e quais figuras de mérito para medir o desempenho?

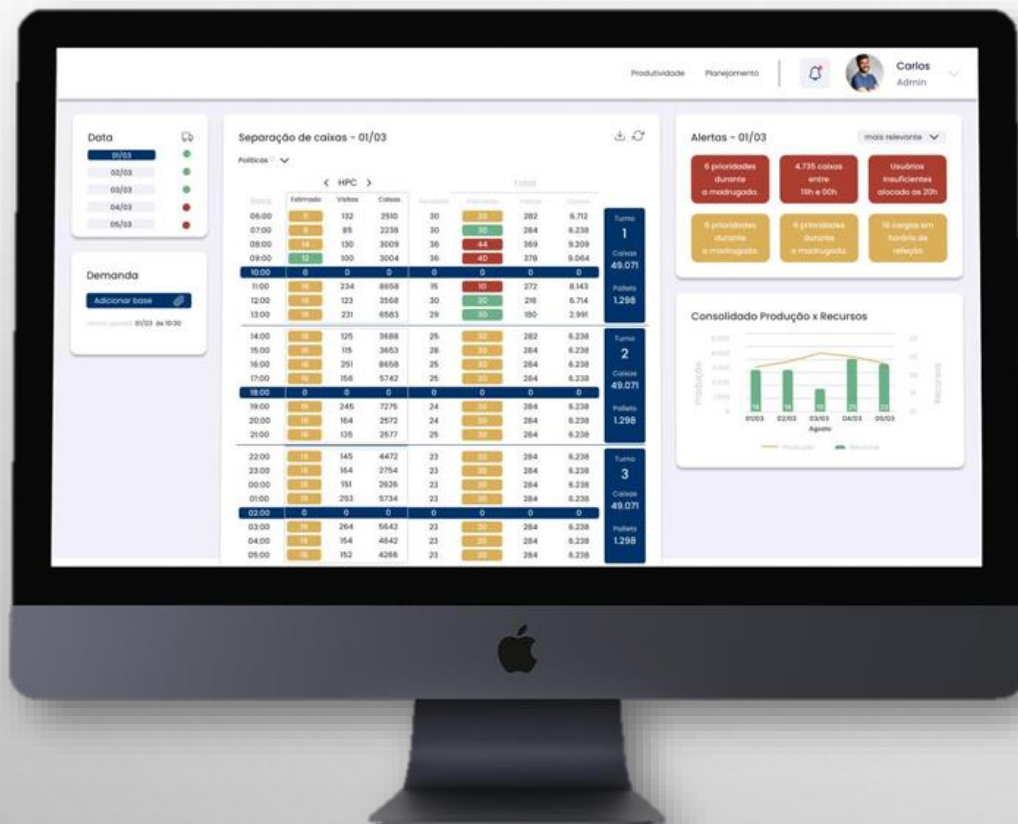
Dados disponíveis

- Listagem dos dados utilizados na tomada de decisão pelo usuário

Frequência da Decisão

- Horária, diária, semanal, etc.

EXEMPLO (1) OTIMIZADOR DE PICKING EM ARMAZÉM



Descrição do Problema de Negócio

- Distribuir por hora, de forma otimizada, a demanda de separação de caixas dos próximos 5 dias, considerando a complexidade da carga, número de separadores escalados, com seus respectivos potenciais produtivos e capacidade de estoque.

Persona & Usuário

- Gestor de Equipe

Riscos envolvidos na decisão

- Atraso nas entregas para os clientes

Dados disponíveis

- Demanda D+5 e produtividade histórica dos separadores.

Frequência da Decisão

- Diária.

EXEMPLO (2) CONSUMO INTELIGENTE DE COMBUSTÍVEL DE NAVIO

Descrição do Problema de Negócio

- Modelo de gestão inteligente de consumo de combustível, utilizando dados meteorológicos para traçar rotas com menor consumo estimado

Persona & Usuário

- 1º Oficial de Náutica

Riscos envolvidos na decisão

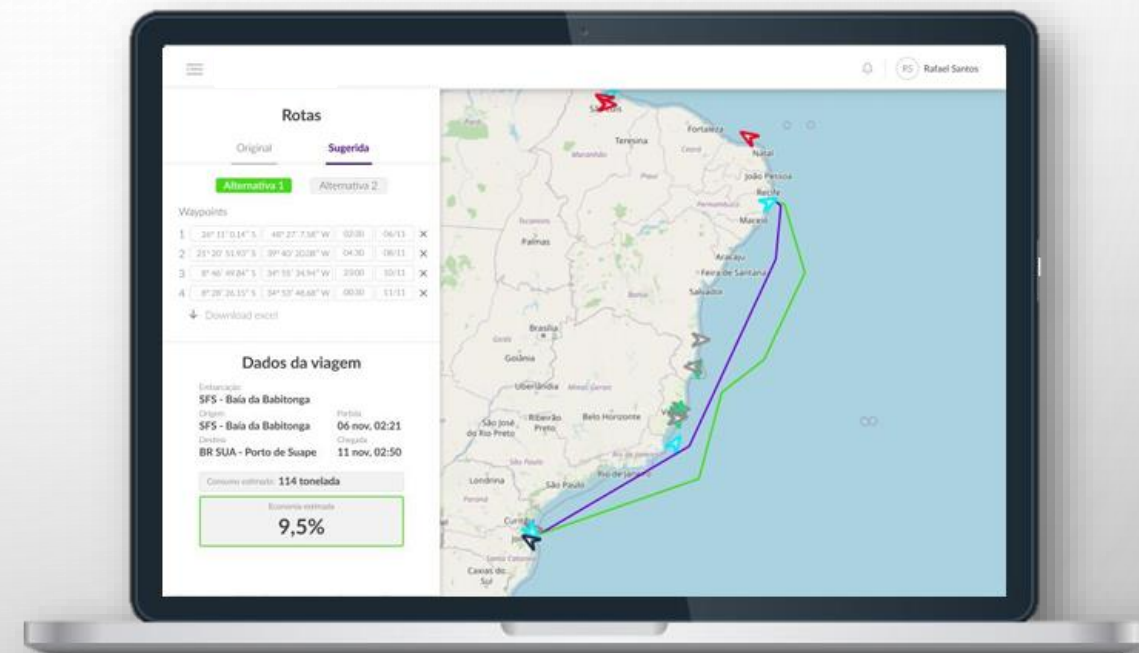
- Consumo excessivo de combustível.

Dados disponíveis

- Telemetria do navio e dados meteorológicos

Frequência da Decisão

- Semanal.



REVISÃO BIBLIOGRÁFICA

Google Scholar



☒ Articles ☐ Case law

New! 2021 Scholar Metrics Released

Articles about COVID-19

CDC NEJM JAMA Lancet Cell BMJ
Nature Science Elsevier Oxford Wiley medRxiv

Stand on the shoulders of giants

SJR

Scimago Journal & Country Rank

Enter Journal Title, ISSN or Publisher Name



WHAT IS SCIMAGOJR FOR?



JOURNAL RANKS

EXPLORE



COUNTRY RANKS

EXPLORE



VIZ TOOLS

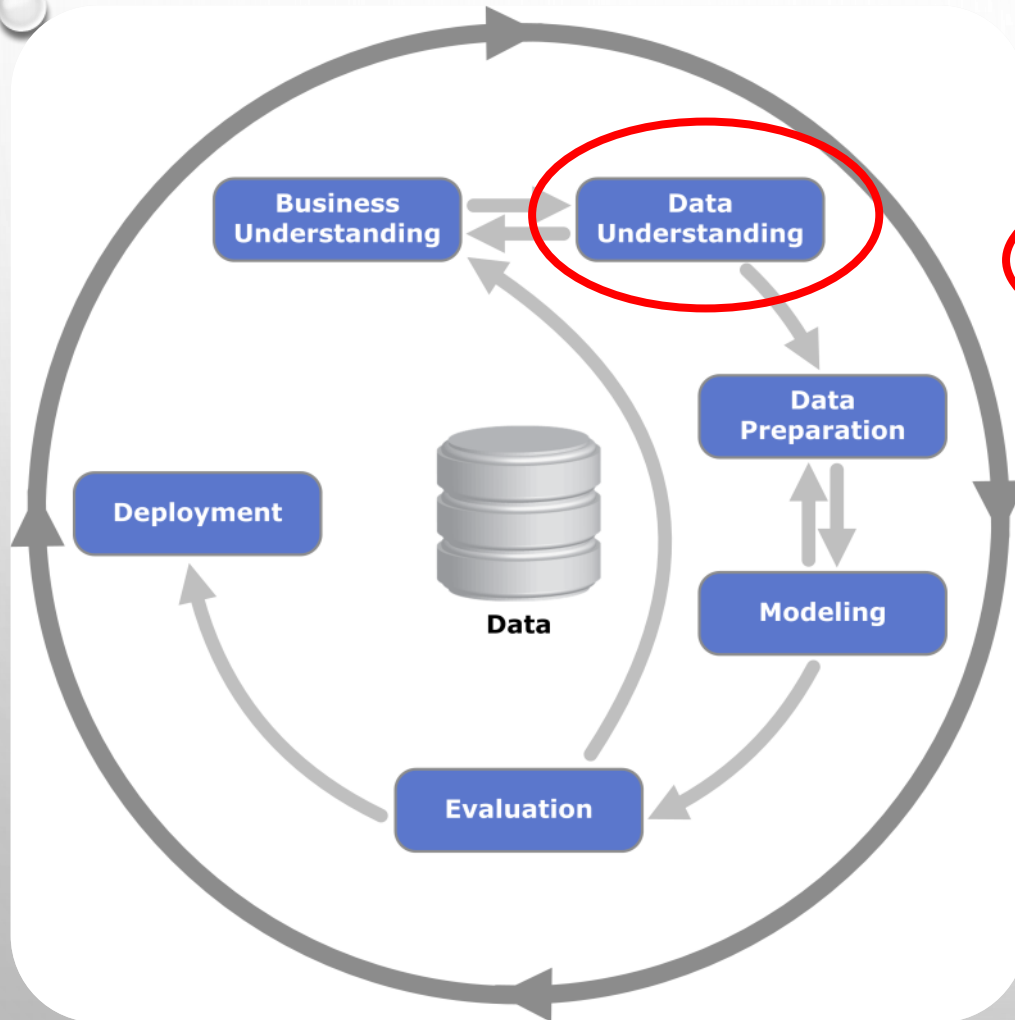
EXPLORE

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of varying sizes, some overlapping. In the center of the top half, there is a faint, circular, embossed-like pattern that resembles a stylized globe or a complex geometric design.

DATA UNDERSTANDING

CRISP-DM

Cross Industry Standard Process for Data Mining - IBM



1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Entendimento das fontes de dados, dos tipos, análise exploratória e representação.

3) Modelagem

Seleção, extração de atributos e treinamento do modelo.

4) Avaliação

Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

DATA UNDERSTANDING

Coleta de Dados

- Obter os dados de arquivos, queries, scrapping ou consulta a API.
- Tratar os dados brutos, combinando diferentes fontes e agregando para a granularidade desejada.

Análise Exploratória

- Tratar cada coluna do problema como uma variável aleatória e avaliar sua distribuição.
- Expurgar variáveis irrelevantes.
- Expurgar outliers (observações problemáticas).
- Avaliar a distribuição conjunta de múltiplas variáveis e eliminar variáveis redundantes.



COLETA DE DADOS

COLETA DE DADOS

Datasets

[+ New Dataset](#)[Filters](#)[Sports](#) ×

 **4,554 Datasets**

[Hotness](#) ▾

EA SPORTS FC 24 FULL PLAYERS DATABASE AND STATS

[Davis Nyagami](#) · Updated 8 days ago

Usability **10.0** · 3 MB

▲ 27

🥉 Bronze ...



NBA Per Game and Advanced Stats (2022-23 Season)

[Jamie Welsh](#) · Updated 9 days ago

Usability **10.0** · 2 Files (CSV) · 81 kB

▲ 28

🥉 Bronze ...



NBA Player Salaries (2022-23 Season)

[Jamie Welsh](#) · Updated 10 days ago

Usability **10.0** · 2 Files (CSV) · 74 kB

▲ 29

🥉 Bronze ...



ODI Men's Cricket Match Data (2002-2023)

[Utkarsh Tomar](#) · Updated 22 days ago

Usability **10.0** · 2 Files (CSV) · 7 MB

▲ 34

🥉 Bronze ...

Formula 1 - who's the best F1 driver?

[Sujay Kapadnis](#) · Updated 7 days ago

Usability **9.4** · 14 Files (other, CSV) · 12 MB

▲ 22

🥉 Bronze ...

COLETA DE DADOS



UTF8
Separador de coluna
Separador de Decimal



Mais seguro salvar
como CSV UTF8 ao
invés de usar XLSX

{ j s o n }

Estrutura de
“Dicionário” Python

Padrão na
comunicação Web



Tensor
Cada frame com 3
canais de cores

ARQUIVO CSV – COMMA SEPARATED VALUES

notebooks > data > players_22.csv > data

```
1  sofifa_id,player_url,short_name,long_name,player_positions,overall,potential,value_eur,wage_eur,age,dob,height_cm,weight_kg,club_team_id,club_name,league_name,league_level,club_positio
2  158023,https://sofifa.com/player/158023/lionel-messi/220002,L. Messi,Lionel Andrés Messi Cuccittini,"RW, ST, CF",93,93,78000000.0,320000.0,34,1987-06-24,170,72,73.0,Paris Saint-Germain
3  188545,https://sofifa.com/player/188545/robert-lewandowski/220002,R. Lewandowski,Robert Lewandowski,ST,92,92,119500000.0,270000.0,32,1988-08-21,185,81,21.0,FC Bayern München,German 1.
4  20801,https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/220002,Cristiano Ronaldo,Cristiano Ronaldo dos Santos Aveiro,"ST, LW",91,91,45000000.0,270000.0,36,1985-02-05,187,83,1
5  190871,https://sofifa.com/player/190871/keymar-da-silva-santos-jr/220002,Neymar Jr,Neymar da Silva Santos Júnior,"RW, CAM",91,91,129000000.0,270000.0,29,1992-02-05,175,68,73.0,Paris Sa
6  192985,https://sofifa.com/player/192985/kevin-de-bruyne/220002,K. De Bruyne,Kevin De Bruyne,"CM, CAM",91,91,125500000.0,350000.0,30,1991-06-28,181,70,10.0,Manchester City,English Premi
7  200389,https://sofifa.com/player/200389/jan-oblak/220002,J. Oblak,Jan Oblak,GK,91,93,112000000.0,130000.0,28,1993-01-07,188,87,240.0,Atlético de Madrid,Spain Primera Division,1,GK,13,,
8  231747,https://sofifa.com/player/231747/kyliau-mbappe/220002,K. Mbappé,Kylian Mbappé Lottin,"ST, LW",91,95,194000000.0,230000.0,22,1998-12-20,182,73,73.0,Paris Saint-Germain,French Lig
9  167495,https://sofifa.com/player/167495/manuel-neuer/220002,M. Neuer,Manuel Peter Neuer,GK,90,90,135000000.0,86000.0,35,1986-03-27,193,93,21.0,FC Bayern München,German 1. Bundesliga,1,G
10 192448,https://sofifa.com/player/192448/marc-andre-ter-stegen/220002,M. ter Stegen,Marc-André ter Stegen,GK,90,92,99000000.0,250000.0,29,1992-04-30,187,85,241.0,FC Barcelona,Spain Prim
11 202126,https://sofifa.com/player/202126/harry-kane/220002,H. Kane,Harry Kane,ST,90,90,129500000.0,240000.0,27,1993-07-28,188,89,18.0,Tottenham Hotspur,English Premier League,1,ST,10,,2
12 215914,https://sofifa.com/player/215914/ngolo-kante/220002,N. Kanté,N'Golo Kanté,"CDM, CM",90,90,100000000.0,230000.0,30,1991-03-29,168,70,5.0,Chelsea,English Premier League,1,RCM,7,,2
13 165153,https://sofifa.com/player/165153/karim-benzema/220002,K. Benzema,Karim Benzema,"CF, ST",89,89,66000000.0,350000.0,33,1987-12-19,185,81,243.0,Real Madrid CF,Spain Primera Divisio
14 192119,https://sofifa.com/player/192119/thibaut-courtois/220002,T. Courtois,Thibaut Courtois,GK,89,91,85500000.0,250000.0,29,1992-05-11,199,96,243.0,Real Madrid CF,Spain Primera Divisi
15 200104,https://sofifa.com/player/200104/heung-min-son/220002,H. Son,손흥민 孙兴慜,"LM, CF, LW",89,89,104000000.0,220000.0,28,1992-07-08,183,78,18.0,Tottenham Hotspur,English Premier Lea
16 200145,https://sofifa.com/player/200145/carlos-henrique-venancio-casimiro/220002,Casemiro,Carlos Henrique Venancio Casimiro,CDM,89,89,88000000.0,310000.0,29,1992-02-23,185,84,243.0,Rea
17 203376,https://sofifa.com/player/203376/virgil-van-dijk/220002,V. van Dijk,Virgil van Dijk,CB,89,89,86000000.0,230000.0,29,1991-07-08,193,92,9.0,Liverpool,English Premier League,1,LCB,
18 208722,https://sofifa.com/player/208722/sadio-mane/220002,S. Mané,Sadio Mané,LW,89,89,101000000.0,270000.0,29,1992-04-10,175,69,9.0,Liverpool,English Premier League,1,LW,10,,2016-07-01
19 209331,https://sofifa.com/player/209331/mohamed-salah/220002,M. Salah,Mohamed Salah Ghalay,RW,89,89,101000000.0,270000.0,29,1992-06-15,175,71,9.0,Liverpool,English Premier League,1,RW,1
20 210257,https://sofifa.com/player/210257/ederson-santana-de-moraes/220002,Ederson,Ederson Santana de Moraes,GK,89,91,94000000.0,200000.0,27,1993-08-17,188,86,10.0,Manchester City,Englis
21 212622,https://sofifa.com/player/212622/joshua-kimmich/220002,J. Kimmich,Joshua Walter Kimmich,"CDM, RB",89,90,108000000.0,160000.0,26,1995-02-08,177,75,21.0,FC Bayern München,German 1
22 212831,https://sofifa.com/player/212831/alisson-ramses-becker/220002,Alisson,Alisson Ramsés Becker,GK,89,90,82000000.0,190000.0,28,1992-10-02,191,91,9.0,Liverpool,English Premier Leagu
23 230621,https://sofifa.com/player/230621/gianluigi-donnarumma/220002,G. Donnarumma,Gianluigi Donnarumma,GK,89,93,119500000.0,110000.0,22,1999-02-25,196,90,73.0,Paris Saint-Germain,Frenc
24 155862,https://sofifa.com/player/155862/sergio-ramos-garcia/220002,Sergio Ramos,Sergio Ramos García,CB,88,88,24000000.0,115000.0,35,1986-03-30,184,82,73.0,Paris Saint-Germain,French Li
25 176580,https://sofifa.com/player/176580/luis-suarez/220002,L. Suárez,Luis Alberto Suárez Díaz,ST,88,88,44500000.0,135000.0,34,1987-01-24,182,83,240.0,Atlético de Madrid,Spain Primera D
26 182521,https://sofifa.com/player/182521/toni-kroos/220002,T. Kroos,Toni Kroos,CM,88,88,75000000.0,310000.0,31,1990-01-04,183,76,243.0,Real Madrid CF,Spain Primera Division,1,LCM,8,,201
27 192505,https://sofifa.com/player/192505/romelu-lukaku/220002,R. Lukaku,Romelu Lukaku Menama,ST,88,88,93500000.0,260000.0,28,1993-05-13,191,94,5.0,Chelsea,English Premier League,1,ST,9,
28 193041,https://sofifa.com/player/193041/keylor-navas/220002,K. Navas,Keylor Navas Gamboa,GK,88,88,15500000.0,130000.0,34,1986-12-15,185,80,73.0,Paris Saint-Germain,French Ligue 1,1,SUB
29 202652,https://sofifa.com/player/202652/raheem-sterling/220002,R. Sterling,Raheem Sterling,"LW, RW",88,89,107500000.0,290000.0,26,1994-12-08,170,69,10.0,Manchester City,English Premier
30 212198,https://sofifa.com/player/212198/bruno-miguel-borges-fernandes/220002,Bruno Fernandes,Bruno Miguel Borges Fernandes,CAM,88,89,107500000.0,250000.0,26,1994-09-08,179,69,11.0,Manc
31 239085,https://sofifa.com/player/239085/erling-haaland/220002,E. Haaland,Erling Braut Haaland,ST,88,93,137500000.0,110000.0,20,2000-07-21,194,94,22.0,Borussia Dortmund,German 1. Bundes
32 153079,https://sofifa.com/player/153079/sergio-aguero/220002,S. Agüero,Sergio Leonel Agüero del Castillo,ST,87,87,51000000.0,260000.0,33,1988-06-02,173,70,241.0,FC Barcelona,Spain Prim
33 167948,https://sofifa.com/player/167948/hugo-lloris/220002,H. Lloris,Hugo Lloris,GK,87,87,13500000.0,125000.0,34,1986-12-26,188,82,18.0,Tottenham Hotspur,English Premier League,1,GK,1,
34 177003,https://sofifa.com/player/177003/luca-modric/220002,L. Modrić,Luka Modrić,CM,87,87,32000000.0,190000.0,35,1985-09-09,172,66,243.0,Real Madrid CF,Spain Primera Division,1,RCM,10,
35 183898,https://sofifa.com/player/183898/angel-di-maria/220002,Á. Di María,Ángel Fabián Di María Hernández,"RW, LW",87,87,49500000.0,160000.0,33,1988-02-14,180,69,73.0,Paris Saint-Germa
```


NOMENCLATURA PARA “OS DADOS”

game	quarter	time	down	distance	field	score	play
1	1	9	1	10	34	0	2
1	1	42	1	10	47	0	1
1	1	83	2	6	49	0	2
1	1	93	3	6	49	0	2
1	1	119	1	10	58	0	2
1	1	163	2	2	66	0	1
1	1	203	3	1	67	0	1
1	1	239	1	10	69	0	2
1	1	270	2	14	65	0	1
1	1	315	3	13	66	0	2
1	1	364	1	10	80	0	1
1	1	397	2	2	88	0	1
1	1	431	3	5	85	0	2
1	1	476	1	9	91	0	1
1	1	514	2	8	92	0	2
1	1	523	3	8	92	0	2
1	1	529	4	8	92	0	3
1	1	852	1	10	34	3	2
1	1	859	2	10	34	3	1
1	1	891	3	8	36	3	2
1	2	0	1	10	59	3	2
1	2	37	1	10	71	3	2
1	2	46	2	10	71	3	2
1	2	53	3	10	71	3	2
1	2	94	1	14	86	3	1

Cada linha é uma observação.

O que define a unicidade da linha é chamado de id, grão ou chave.

Cada coluna é um atributo ou variável independente.

A planilha inteira é uma amostra.

NUMPY

Processamento rápido de matrizes

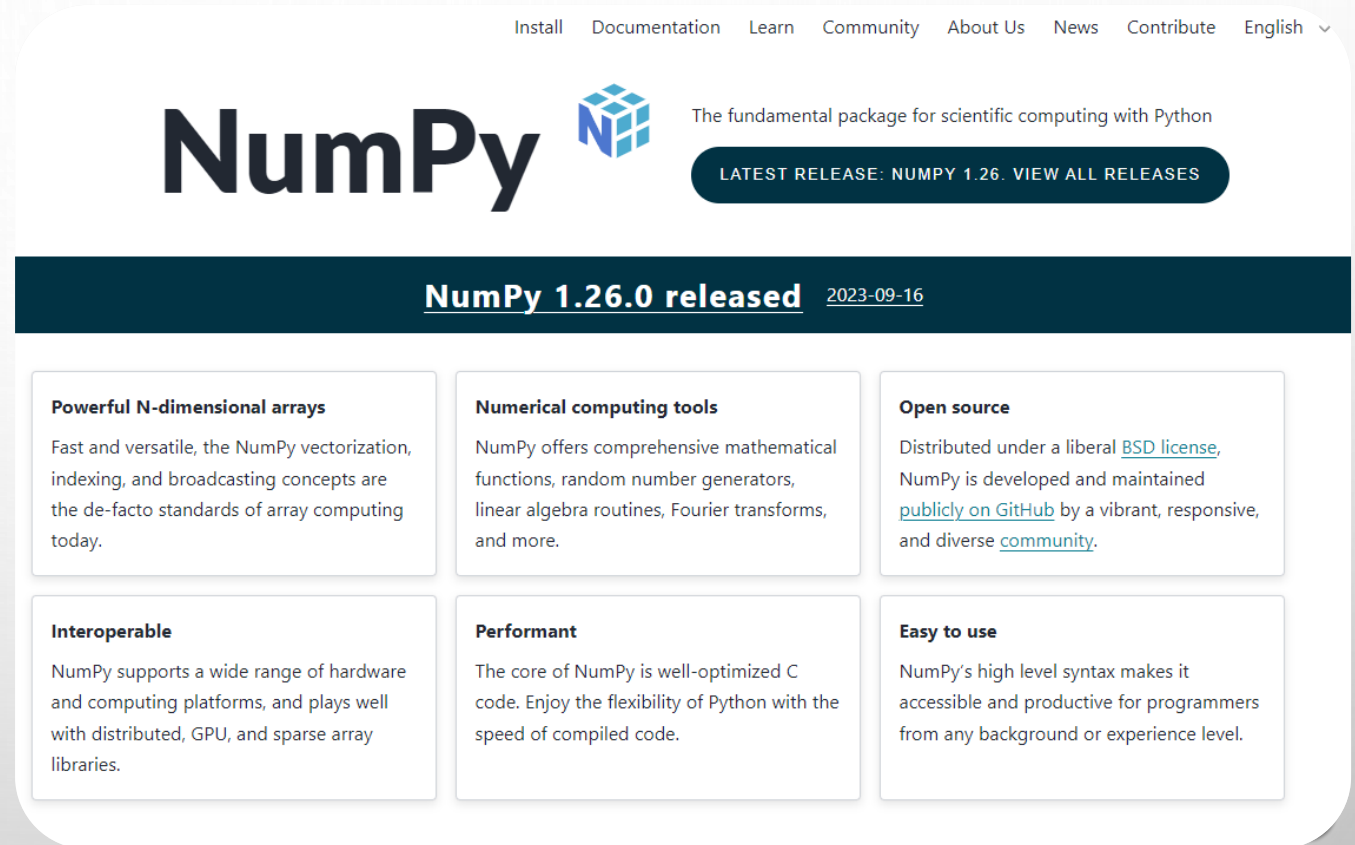
+ Algoritmos

+ Dado “Baixo Nível”

+Open Source

+Performático

+Sintaxe para os Pesquisadores



The screenshot shows the NumPy website homepage. At the top, there is a navigation bar with links: Install, Documentation, Learn, Community, About Us, News, Contribute, and English. The main header features the NumPy logo (a blue cube) and the text "The fundamental package for scientific computing with Python". Below this, a dark blue banner announces "NumPy 1.26.0 released" with the date "2023-09-16". The main content area is divided into six white boxes with dark blue borders, each highlighting a feature of NumPy: "Powerful N-dimensional arrays", "Numerical computing tools", "Open source", "Interoperable", "Performant", and "Easy to use". Each box contains a brief description of the feature.

Install Documentation Learn Community About Us News Contribute English

NumPy

The fundamental package for scientific computing with Python

LATEST RELEASE: NUMPY 1.26. VIEW ALL RELEASES

NumPy 1.26.0 released 2023-09-16

Powerful N-dimensional arrays

Fast and versatile, the NumPy vectorization, indexing, and broadcasting concepts are the de-facto standards of array computing today.

Numerical computing tools

NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

Open source

Distributed under a liberal [BSD license](#), NumPy is developed and maintained [publicly on GitHub](#) by a vibrant, responsive, and diverse [community](#).

Interoperable

NumPy supports a wide range of hardware and computing platforms, and plays well with distributed, GPU, and sparse array libraries.

Performant

The core of NumPy is well-optimized C code. Enjoy the flexibility of Python with the speed of compiled code.

Easy to use

NumPy's high level syntax makes it accessible and productive for programmers from any background or experience level.

NUMPY

ndarray

- + array de qualquer tipo numérico podendo conter vazios
- + criação rápida e manipulação dessas arrays.
 - + Sequência
 - + criar, editar, operações, ordenar, concatenar, indexar, aleatórios, estatísticas
 - + Vetor
 - + Matriz
 - + Tensor

PANDAS

Análise de Séries e Data Frames

+ Compatível com múltiplas fontes

de dados

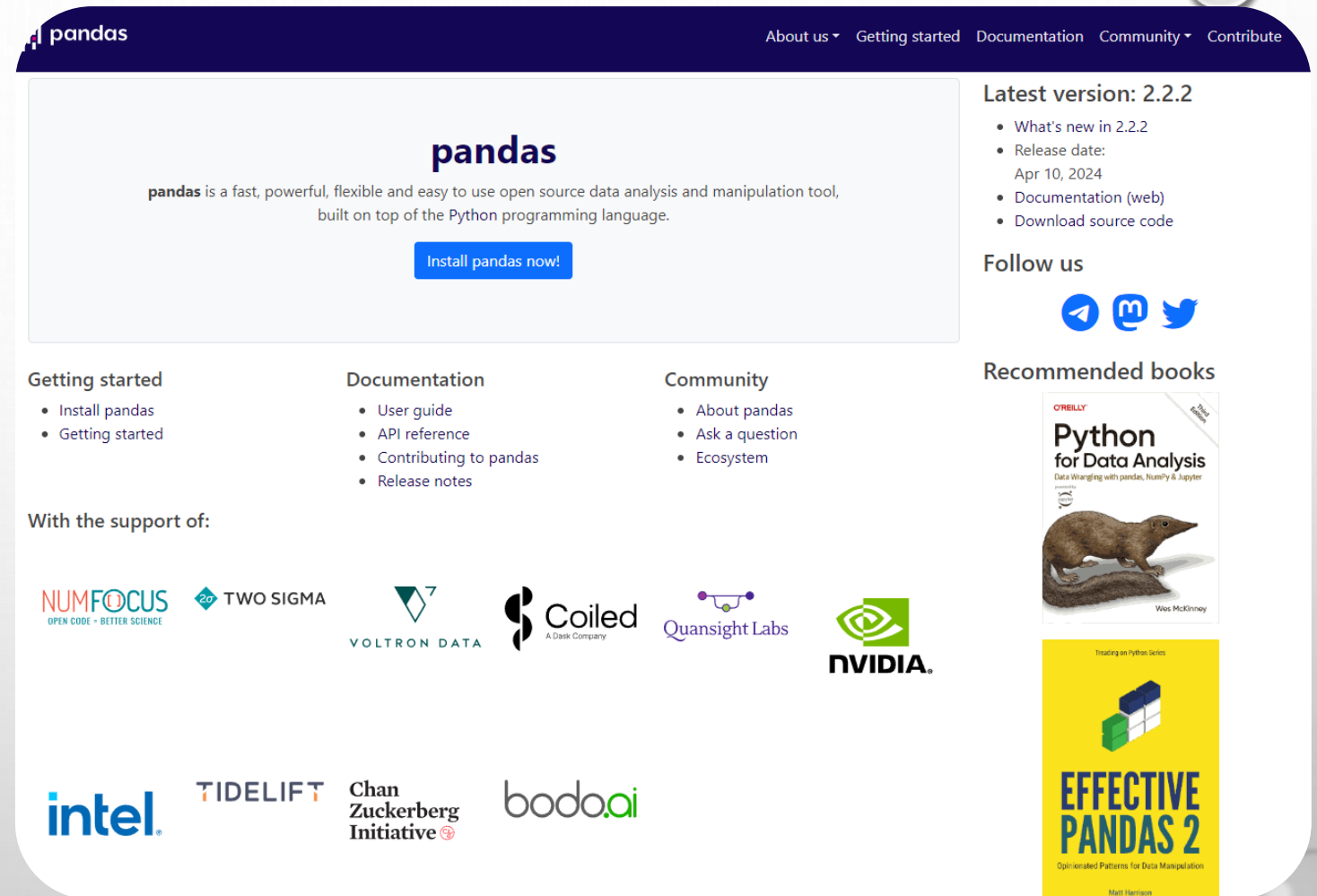
+ Dado mais “Alto Nível”

+ Estatísticas

+ Sumarizações

+ Visualização

+ Operações de Tabelas



The screenshot shows the pandas website homepage. At the top is a dark blue navigation bar with the pandas logo and links for 'About us', 'Getting started', 'Documentation', 'Community', and 'Contribute'. The main content area has a large 'pandas' heading, followed by a description: 'pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.' Below this is a blue button that says 'Install pandas now!'. To the right, under 'Latest version: 2.2.2', there is a list of links: 'What's new in 2.2.2', 'Release date: Apr 10, 2024', 'Documentation (web)', and 'Download source code'. Below this is a 'Follow us' section with icons for Telegram, Mastodon, and Twitter. Further down, there are three columns: 'Getting started' with links to 'Install pandas' and 'Getting started'; 'Documentation' with links to 'User guide', 'API reference', 'Contributing to pandas', and 'Release notes'; and 'Community' with links to 'About pandas', 'Ask a question', and 'Ecosystem'. Below these columns is a section titled 'With the support of:' featuring logos for NUMFOCUS, TWO SIGMA, VOLTRON DATA, Coiled, Quansight Labs, NVIDIA, intel, TIDELIFT, Chan Zuckerberg Initiative, and bodo.ai. On the right side, there are two book covers: 'Python for Data Analysis' by Wes McKinney and 'EFFECTIVE PANDAS 2' by Matt Harrison.

Pandas

pandas.Series

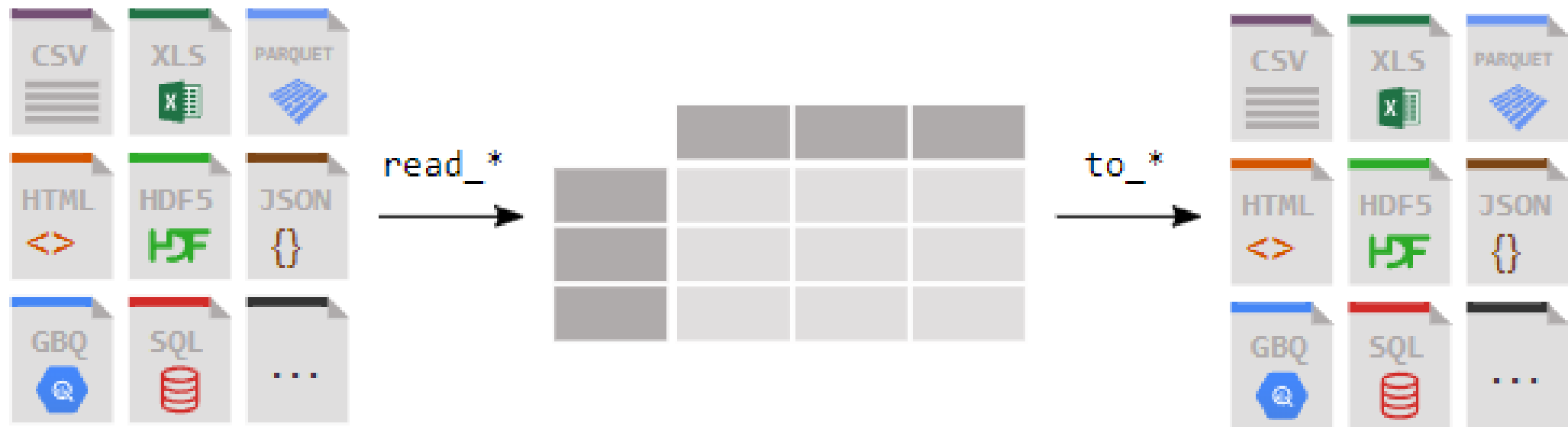
- + ndarray indexada por uma coluna.
- + Métodos para cálculos de estatísticas e séries temporais.
- + “Casca” ao redor do numpy para permitir indexação
chave -> valor para um elemento.
- + Facilita a operação com índices

Pandas

pandas.DataFrame

- + tabela de duas dimensões indexada por linha e por coluna, permitindo colunas com tipos de dados múltiplos
- + Métodos para estatísticas, agrupamento, indexação, visualização de dados.
- + É uma “Planilha Excel” com muito mais ferramentas de análise.
- + Possui operações de “Bancos de Dados” para associar tabelas.

I/O DE DADOS COM PANDAS



QUAIS SÃO OS TIPOS MAIS COMUNS DE ATRIBUTOS?

Nominal ou Categórica

- Conjunto de diferentes valores não ordenados.
- Exemplo: Sexo, cor, palavras, tipo de coisas.

string / bool

Ordinal

- Conjunto ordenado, mas a diferença entre os valores não tem significado.
- Exemplo: scores quantitativos como “excelente”, “bom”, “regular”, “ruim”.

string / int

Intervalo

- Conjunto ordenado, a diferença tem significado mas não as proporções.
- Exemplo: Datas.

datetime

Ratio

- Conjunto ordenado onde diferenças & proporções tem significado.
- Exemplo: Idade, peso, altura, dinheiro, massa, etc.

int / float

Texto

- Sequência de palavras de tamanho finito.
- Exemplo: “Ontem eu fui passear”.

string

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. In the center of the slide, there is a faint, circular watermark. It features a stylized sun or flower-like symbol in the middle, surrounded by text in a circular arrangement, though the text is too light to read clearly.

ANÁLISE EXPLORATÓRIA

EXEMPLO (1): ANÁLISE EXPLORATÓRIA DO DATASET IRIS



Iris Setosa



Iris Versicolor

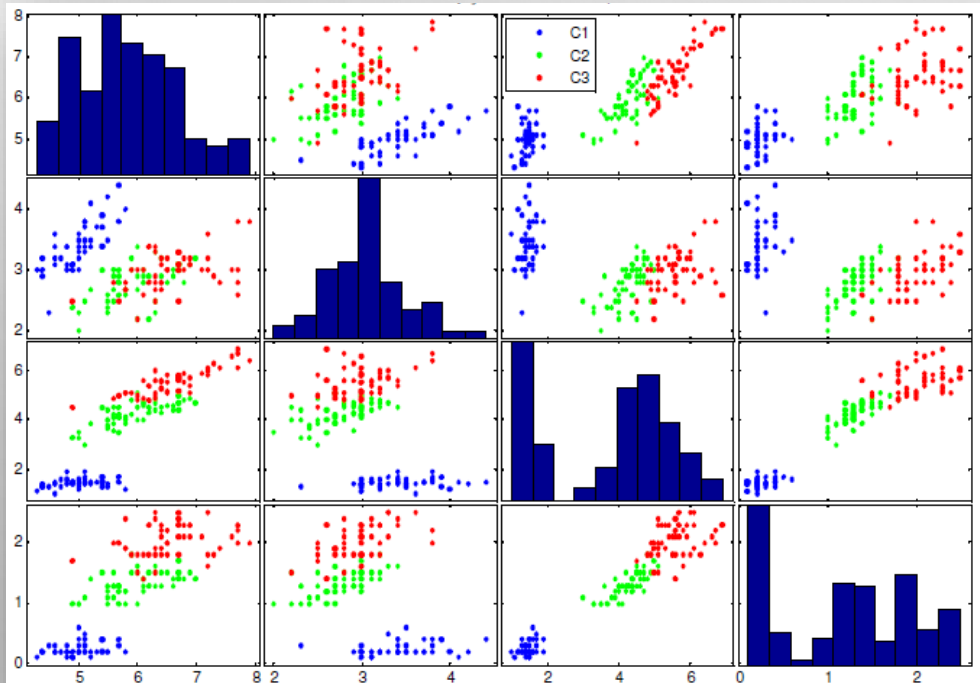


Iris Virginica

Características das flores

Largura & comprimento da pétala

Largura & comprimento da sépala



<http://archive.ics.uci.edu/ml/datasets/Iris>

**Espaço de Atributos
com 4 dimensões!**

PORQUE É NECESSÁRIO MANTER O NÚMERO DE ATRIBUTOS O MENOR POSSÍVEL?

1) “Maldição” da Dimensionalidade

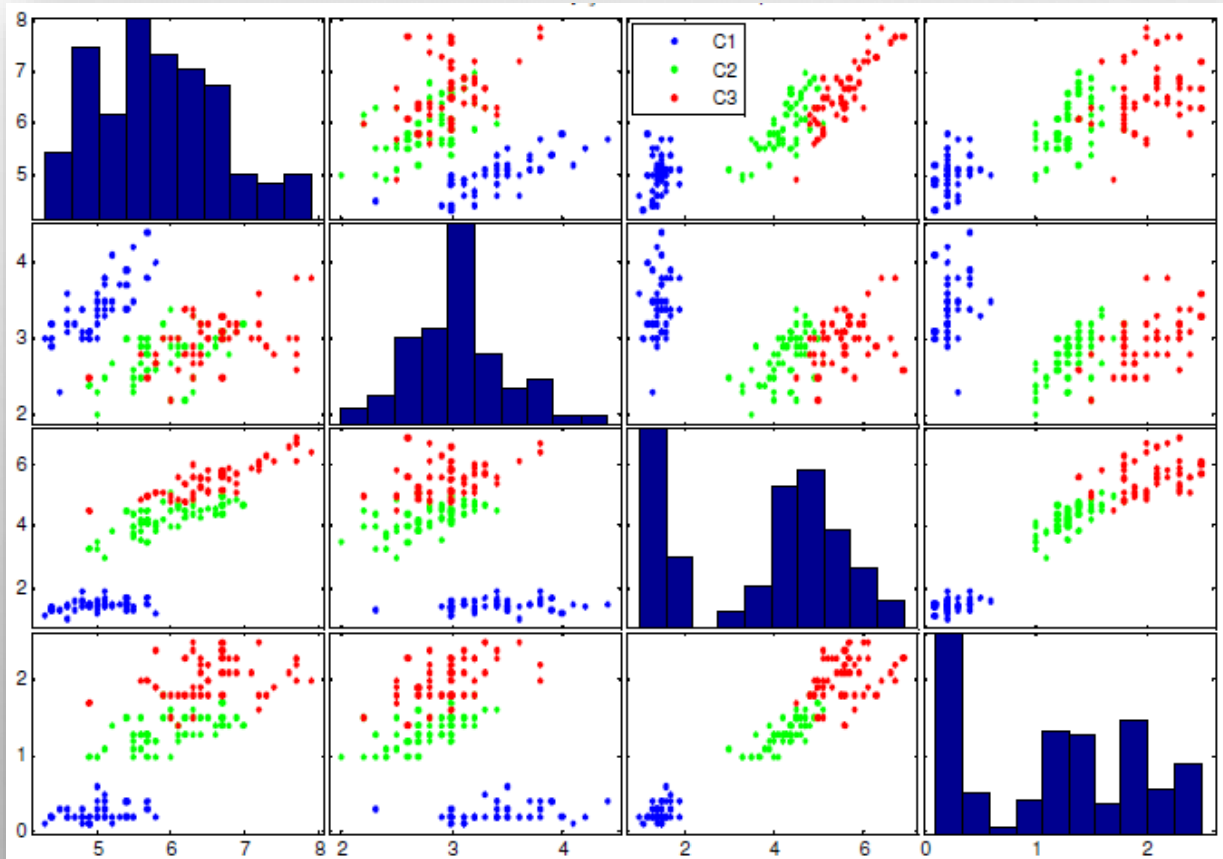
Suponha que 10.000 observações são distribuídas aleatoriamente no intervalo $[0, 1]$. Qual é a distância média entre os pontos? E se as observações são distribuídas no cubo $[0, 1]^3$? Ou em um hipercubo de 100 dimensões?

Esparsidade do espaço de atributos aumenta com o **número de dimensões!**

2) Multi-colinearidade

Dois atributos com uma relação significativa pode sugerir causalidade entre ambos ou relação com uma variável latente desconhecida. De uma maneira ou outra, o atributo “independente” vai ter mais importância para o modelo, já que está representado por mais de um atributo.

QUAIS ATRIBUTOS UTILIZAR?



Para separar a Iris Setosa (azul)?

Para separar Iris Virginica (Vermelha)?

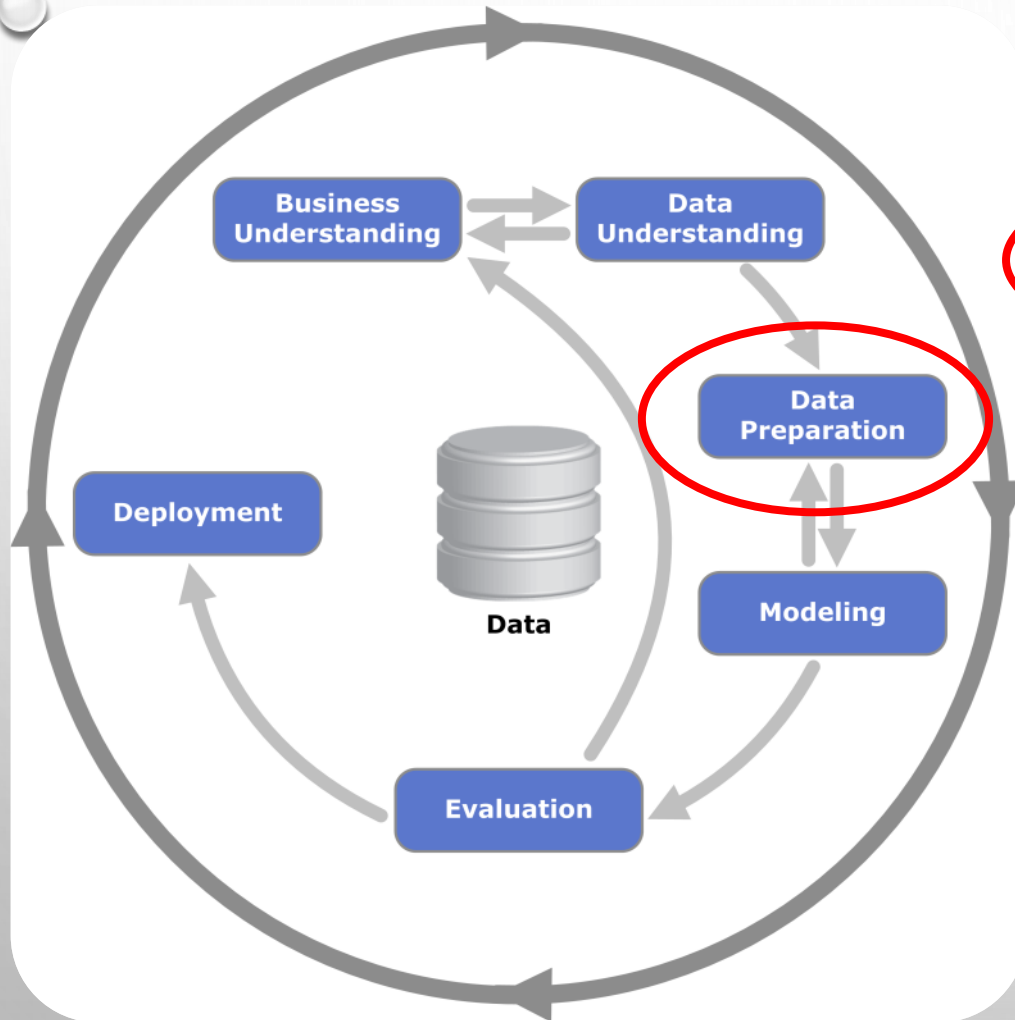
Para encontrar corretamente
3 grupos de flores?

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. In the center of the slide, there is a faint, circular watermark. It features a globe with latitude and longitude lines, and the text "UNIVERSITY OF CALIFORNIA" is visible around the perimeter of the circle.

DATA PREPARATION

CRISP-DM

Cross Industry Standard Process for Data Mining - IBM



1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Entendimento das fontes de dados, dos tipos, análise exploratória e representação.

3) Modelagem

Seleção, extração de atributos e treinamento do modelo.

4) Avaliação

Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

DATA PREPARATION

Quantificação dos Atributos

- Transformar todos os atributos em atributos numéricos.

Escalonamento

- Transformar todos os atributos para a mesma faixa dinâmica, de maneira a assegurar que todos tenham o mesmo “peso numérico” para o treinamento do modelo.

Normalização

- Garantir que os dados tenham uma distribuição de probabilidade gaussiana (Normal).

ATRIBUTOS CATEGÓRICOS

One Hot Encoding

Gender
Female
Male
Male
Female



Gender
1
0
0
1

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

DATAS

Componentes da Data

- Ano
- Mês
- Dia
- Dia do Ano
- Dia da Semana
- Hora
- Minuto
- Segundo

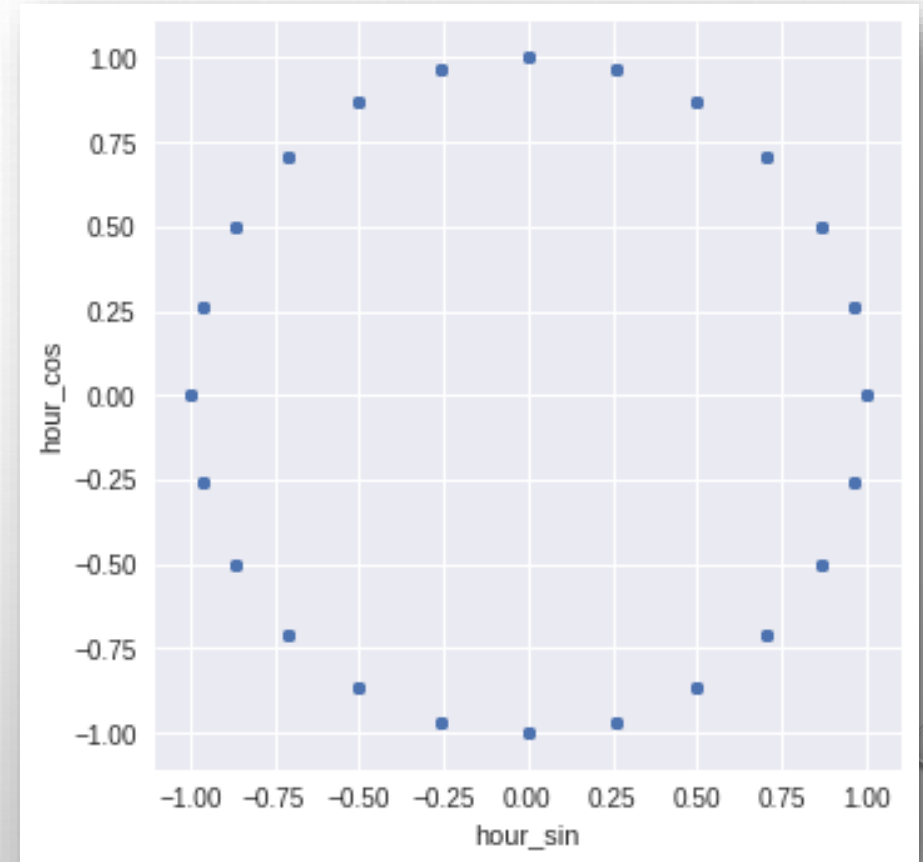
Flags

- É final de semana
- É feriado

Diferença entre Datas

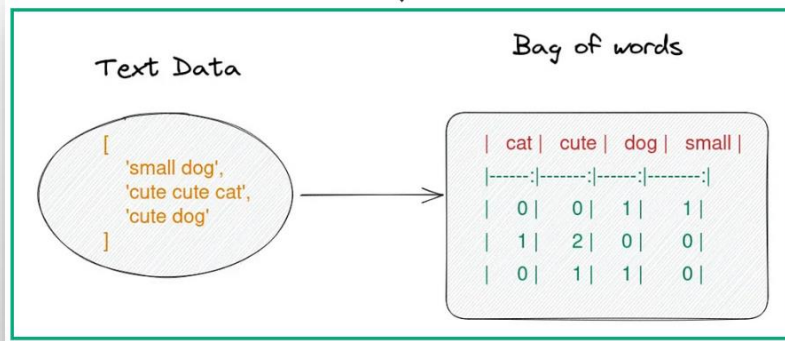
- Diferença em Dias
- Diferença em Horas
- Diferença em Meses

Encoding Cíclico



ATRIBUTOS TEXTUAIS

BAG OF WORDS



Variants of term frequency (tf) weight	
weighting scheme	tf weight
binary	0, 1
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

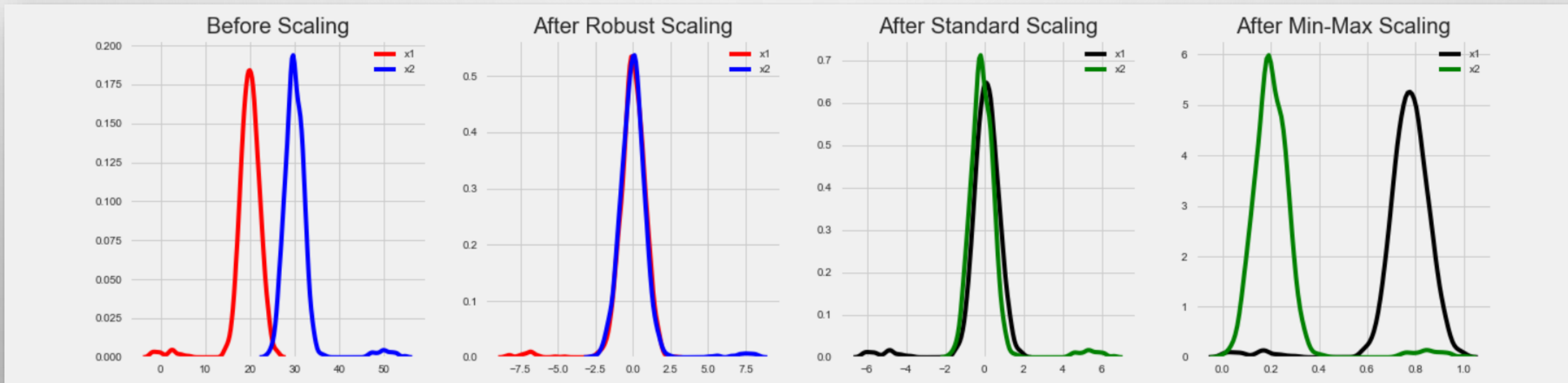
TF-IDF

Variants of inverse document frequency (idf) weight	
weighting scheme	idf weight ($n_t = \{d \in D : t \in d\} $)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log \left(\frac{N}{1 + n_t} \right) + 1$
inverse document frequency max	$\log \left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t} \right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

TF-IDF Calculation Example							
Words	Count		Term Frequency (TF)		Inverse Document Frequency (IDF)	TF * IDF	
	Document 1	Document 2	Document 1	Document 2		Document 1	Document 2
read	1	1	0.17	0.17	0	0	0
svm	1	0	0.17	0	0.3	0.05	0
algorithm	1	1	0.17	0.17	0	0	0
article	1	1	0.17	0.17	0	0	0
dataaspirant	1	1	0.17	0.17	0	0	0
blog	1	1	0.17	0.17	0	0	0
randomforest	0	1	0	0.17	0.3	0	0.05

ESCALONAMENTO E NORMALIZAÇÃO

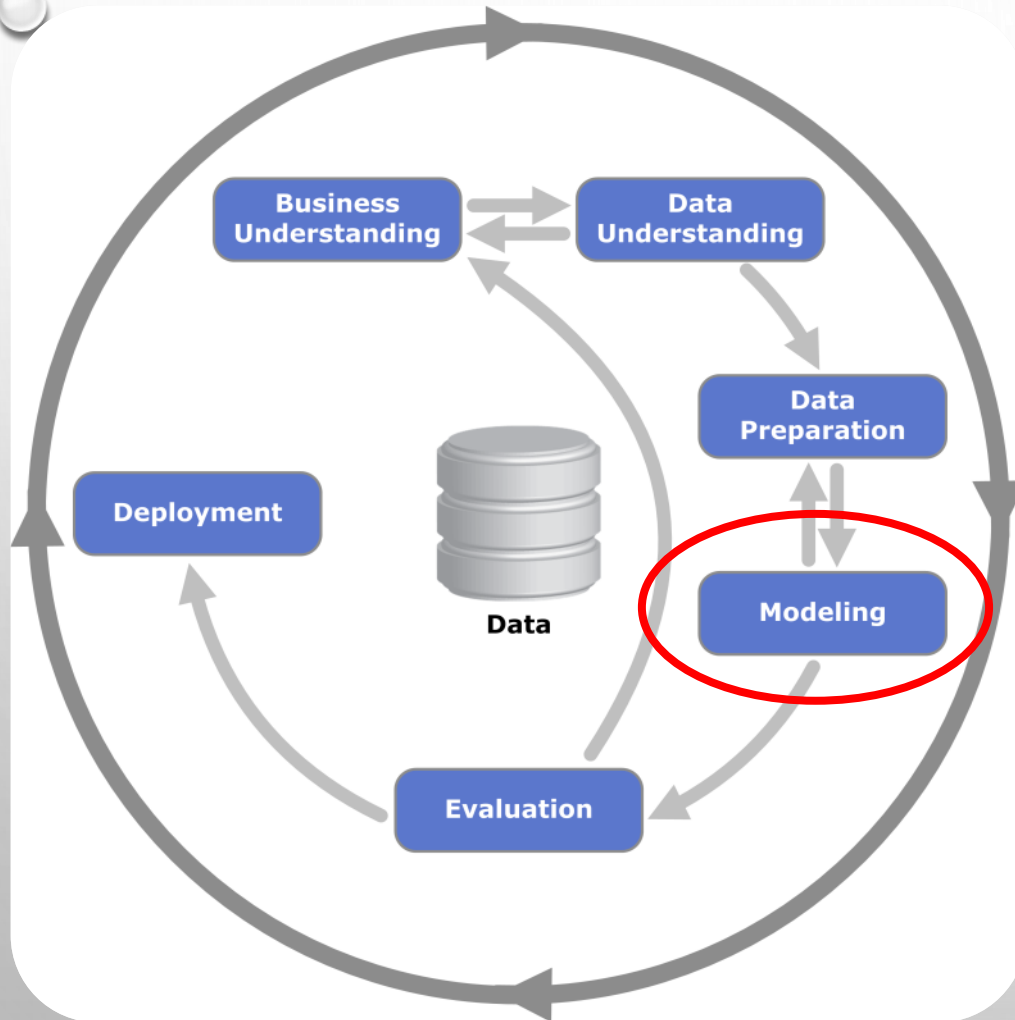
- Garantir que as variáveis possuam a mesma escala
- Mesmo efeito numérico na otimização independente da escala.



MODELING

CRISP-DM

Cross Industry Standard Process for Data Mining - IBM



1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Entendimento das fontes de dados, dos tipos, análise exploratória e representação.

3) Modelagem

Seleção, extração de atributos e treinamento do modelo.

4) Avaliação

Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

MODELING

Seleção de Atributos

- Quantificar e ordenar os atributos por importância para o problema.
- Eliminar atributos irrelevantes.

Extração de Atributos

- Transformar os atributos do espaço original para um espaço que favoreça a modelagem.

Treinamento

- Encontrar os hiper-parâmetros e parâmetros do modelo, para os dados disponíveis, avaliando a figura de mérito selecionada.

The background of the slide is a light gray gradient. In the top-left and bottom-right corners, there are clusters of realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. In the center of the slide, there is a faint, circular watermark. It contains a stylized graphic of a person's head and shoulders, surrounded by text in a circular arrangement, likely representing a university or institutional logo.

SELEÇÃO DE ATRIBUTOS

TÉCNICAS DE SELEÇÃO DE ATRIBUTOS

Filtragem – mede a relação entre atributos ou atributos e classes, utilizando estatísticas, sem depender do modelo.

- **Coeficiente de Correlação de Pearson** – Estatística que mede a relação linear entre duas variáveis aleatórias.
- **Teste T de diferença de médias** – Informa se a média de um determinado atributo muda de acordo com uma categoria binária.
- **ANOVA** – O mesmo que o teste T, mas serve para múltiplas categoria.
- **Informação Mútua** – Estatística que mede relação não-linear entre duas variáveis aleatórias.

Wrapper – mede a relação entre atributos e classes, utilizando um modelo treinado.

- **Gini** – Estatística que representa a importância de um atributo na divisão da base de dados por uma árvore de decisão.
- **Relevância** – Estatística que representa a variação causada na saída do modelo quando um atributo é substituído por sua média.

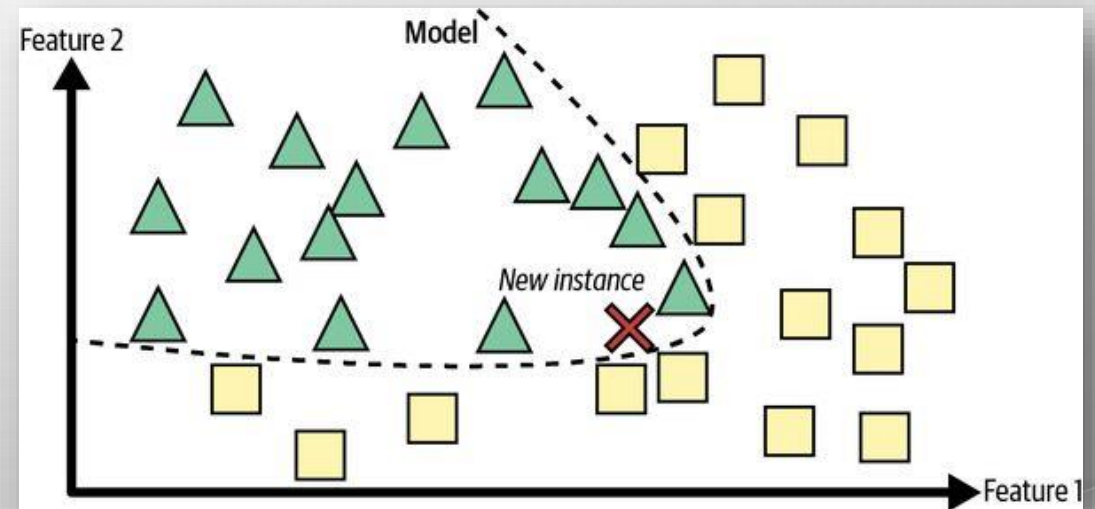
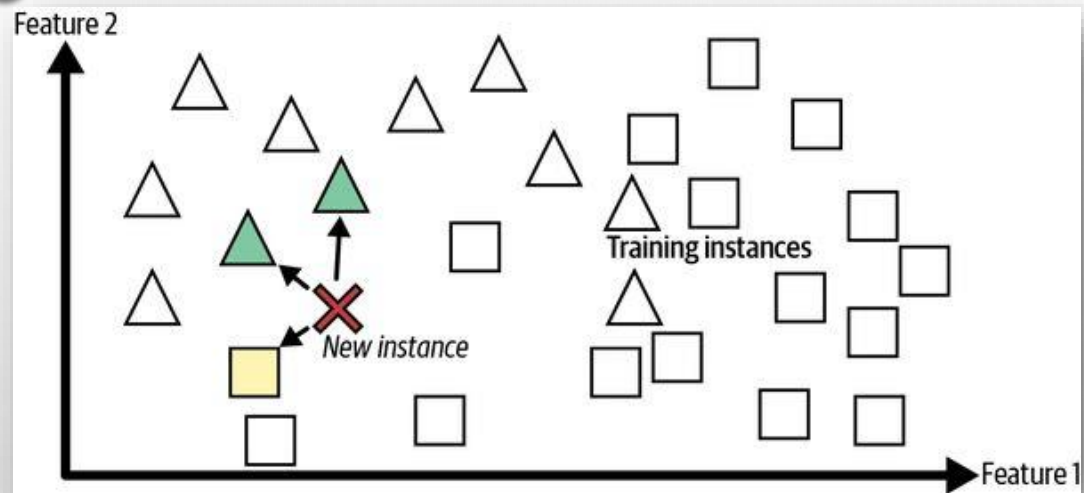
The slide features a light gray background with a subtle gradient. In the top-left and bottom-right corners, there are clusters of realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. Faintly visible in the upper center is a circular watermark containing a stylized sun or flower-like emblem.

SELEÇÃO DO MODELO

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. In the center of the slide, there is a faint, circular watermark. It features a globe in the background with a play button icon in the center, surrounded by the text 'UNIVERSIDADE FEDERAL DO RIO DE JANEIRO' and 'FACULDADE DE ENGENHARIA' at the bottom.

CLASSIFICAÇÃO

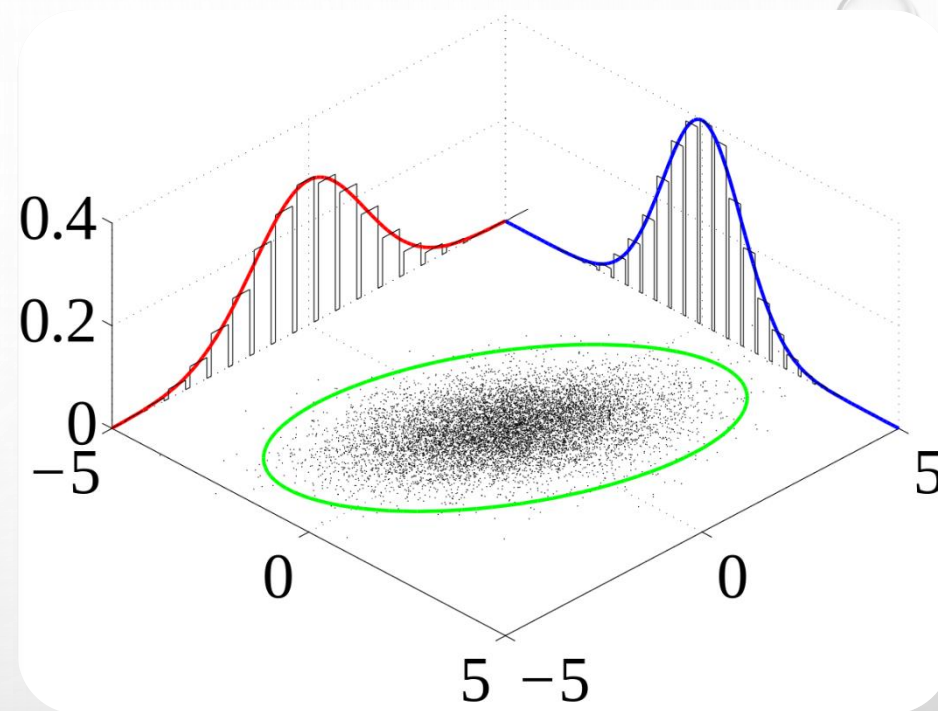
ALGORITMOS BASEADOS EM INSTÂNCIA VS MODELO



ALGORITMOS BASEADOS EM DENSIDADE

Algoritmos que dependem da **função densidade de probabilidade** dos dados, ou aproximações locais, para determinar a classe de observações fora da amostra de treino.

- 1) Classificador Bayesiano
- 2) Classificador Bayesiano “Naïve”
- 3) K-Vizinhos mais próximos



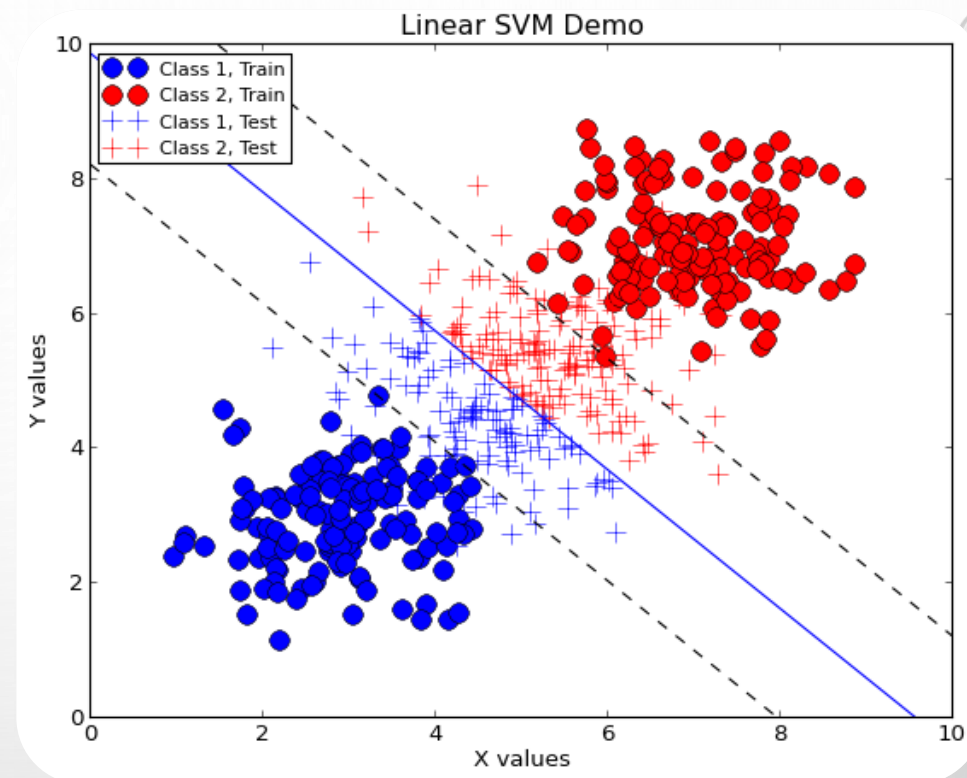
Algoritmos baseados em densidade dependem da **DENSIDADE (!!!)**. Consequentemente, se beneficiam de um **conjunto grande de observações e de baixa esparsidade do espaço de atributos**. O Classificador Bayesiano é considerado o classificador “ótimo”, mas é raramente utilizado, dada a dificuldade de estimar a função densidade de probabilidade dos dados. É normalmente utilizado como benchmark para comparação teórica entre os algoritmos de classificação.

MODELOS FUNCIONAIS

Algoritmos que dependem da **estimação dos parâmetros de uma função** que é utilizada como **superfície de separação** entre as classes.

- 1) Funções Polinomiais
- 2) Regressão Logística
- 3) Máquina de Vetores Suporte
- 4) Neurônio Sigmoidal / Tangente Hiperbólica
- 5) Árvores de Decisão

Algoritmos baseados em funções são **mais simples**, usualmente tem um **número menor de parâmetros** e não dependem em armazenar muitos dados para manter uma “memória”, como por exemplo K-vizinhos mais próximos.



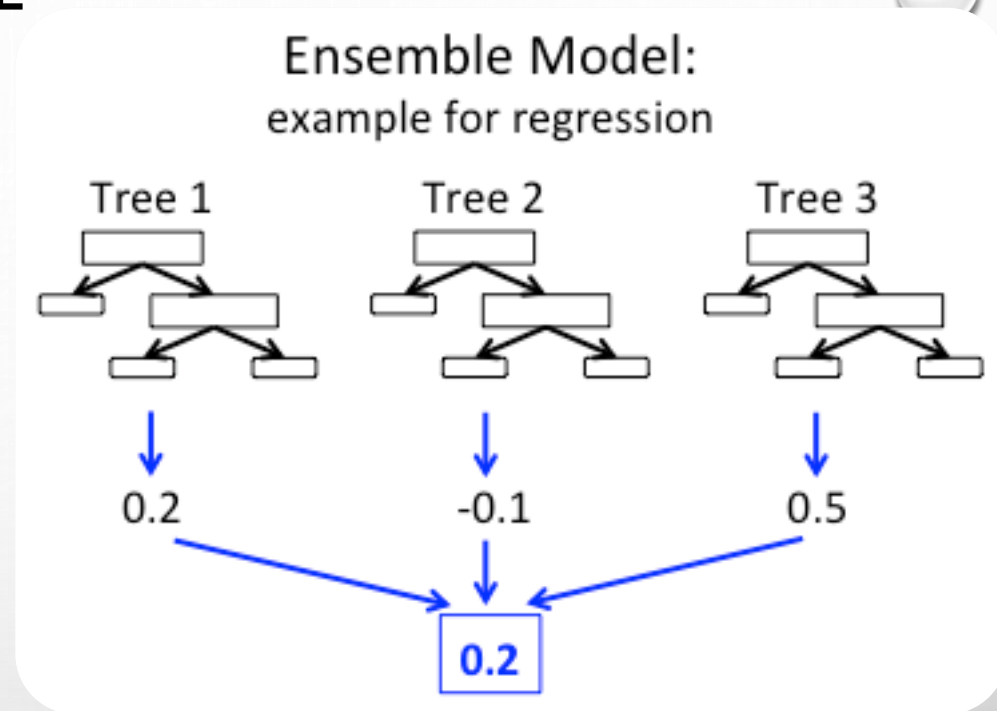
ALGORITMOS BASEADOS EM ENSEMBLE

Algoritmos que **combinam modelos simples**,
usualmente através de **votação ou ponderação**, para
atingir maiores taxas de classificação.

1) Random Forest

2) Boosting

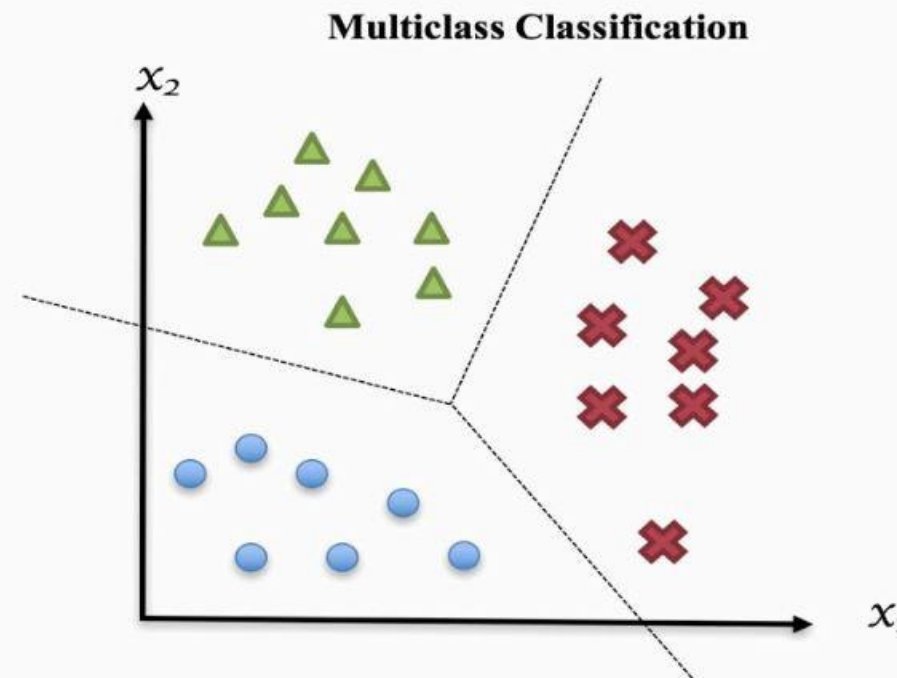
Boa **capacidade de generalização** gerado através de **arranjos complexos** de múltiplos modelos simples de machine learning.



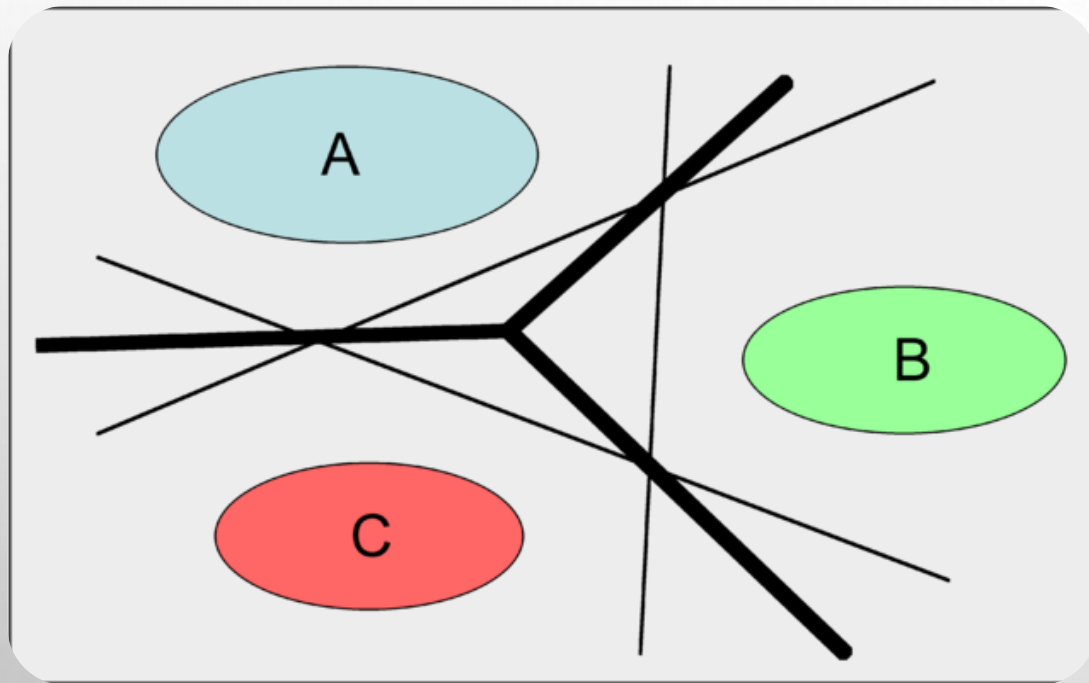
ALGORITMOS BASEADOS EM ENSEMBLE

- **Modelos Multiclasse**

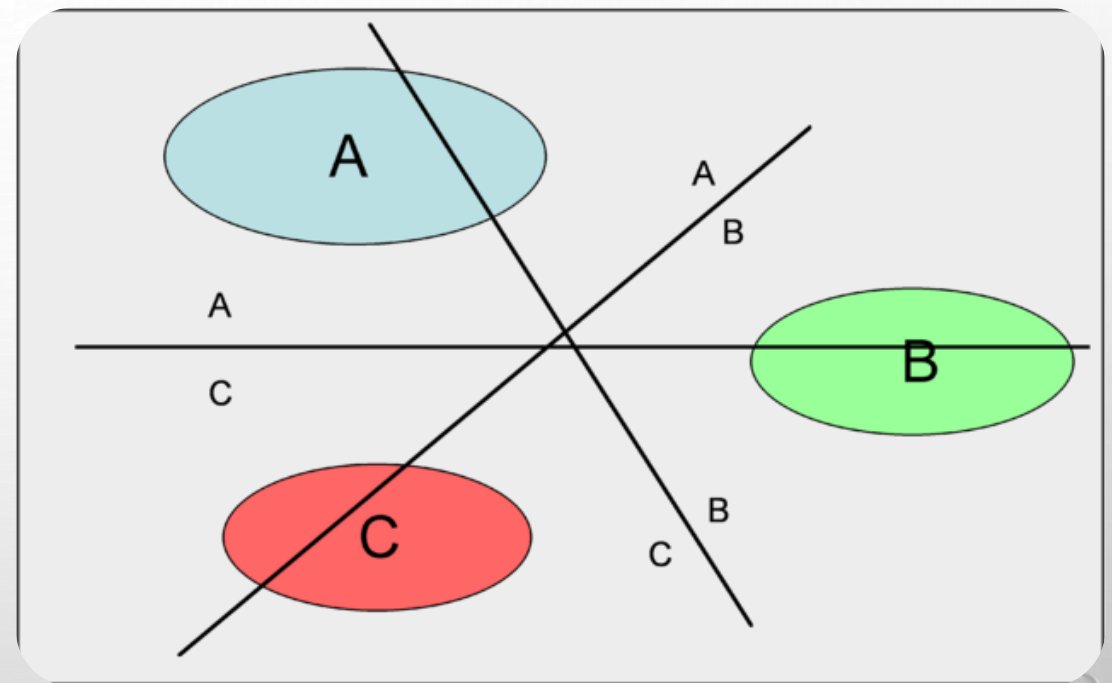
- Discriminar múltiplos objetos em paralelo.
- Ensembles podem ser utilizados para especializar modelos.
- Alguns modelos são naturalmente multiclasse, como redes neurais.



ENSEMBLES BÁSICOS



ONE AGAINST ALL



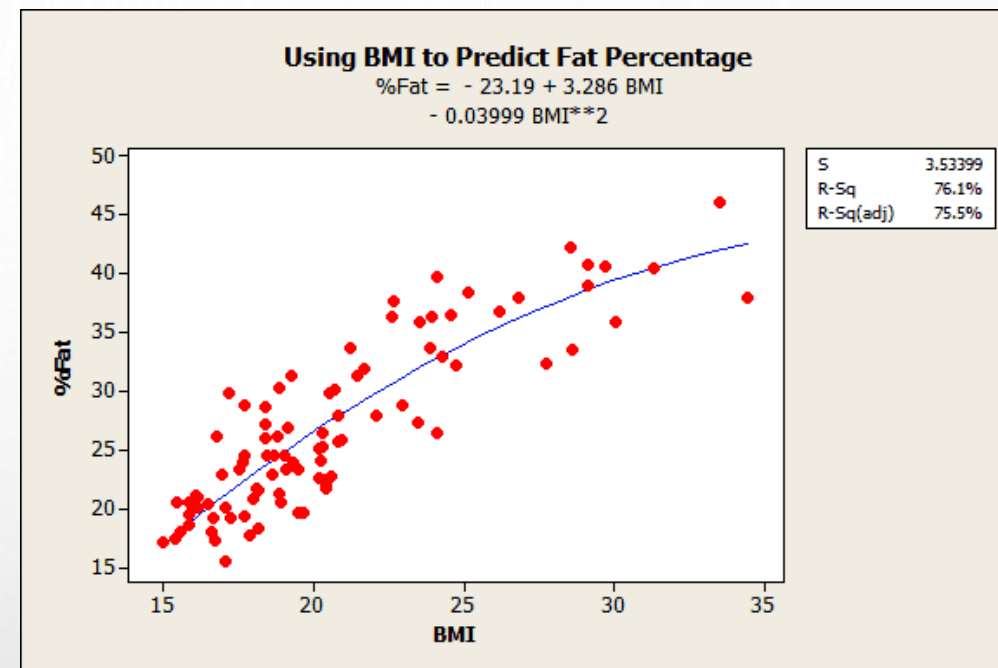
ONE AGAINST ONE

The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, rendered with soft shadows and highlights to give them a three-dimensional appearance. In the center of the image, the word "REGRESSÃO" is written in a bold, black, sans-serif font.

REGRESSÃO

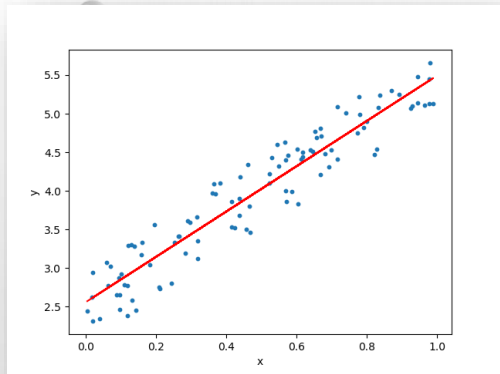
MODELOS DE REGRESSÃO

- 1) Regressão Linear
- 2) K Vizinhos mais Próximos
- 3) Regressão Não-Linear
- 4) Processos Gaussianos
- 5) Máquina de Vetores Suporte
- 6) Redes Neurais

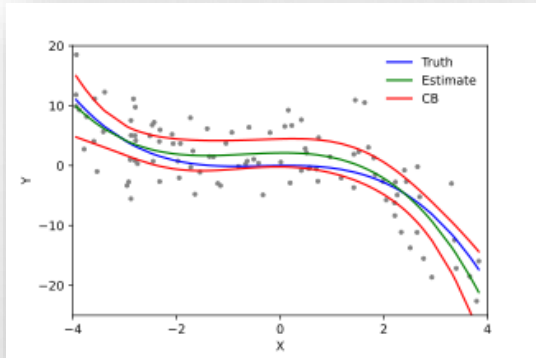


Algoritmos de regressão geralmente são modelados combinando uma **parte determinística e uma parte aleatória**. Os parâmetros correspondente à parte determinística são encontrados utilizando estimadores como máxima verossimilhança ou máximo a posteriori (MAP).

MODELOS DE REGRESSÃO



$$Y = \alpha^T x + \varepsilon$$

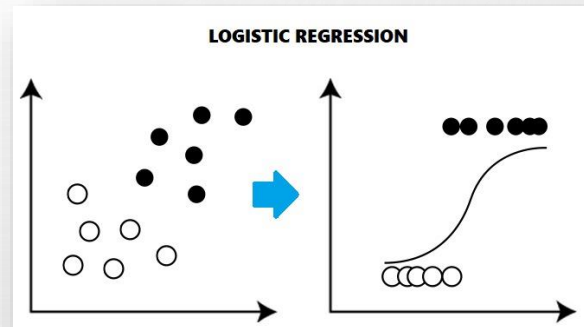


$$Y = X\alpha + \varepsilon$$

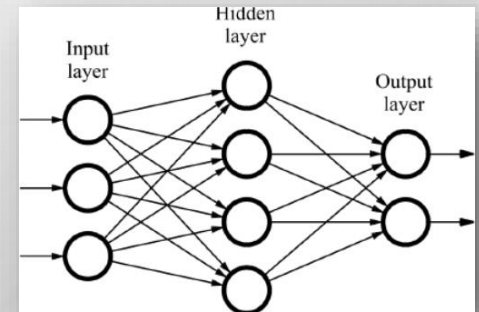
$$Y = \underbrace{F(X)}_{\text{Parte Determinística}} + \underbrace{\varepsilon}_{\text{Parte Estocástica}}$$

Parte Determinística

Parte Estocástica



$$Y = \frac{1}{1 + e^{\alpha^T x + \varepsilon}}$$

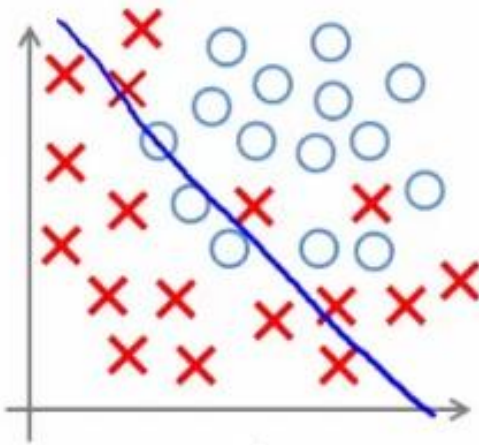


$$Y = \varphi(x) + \varepsilon$$



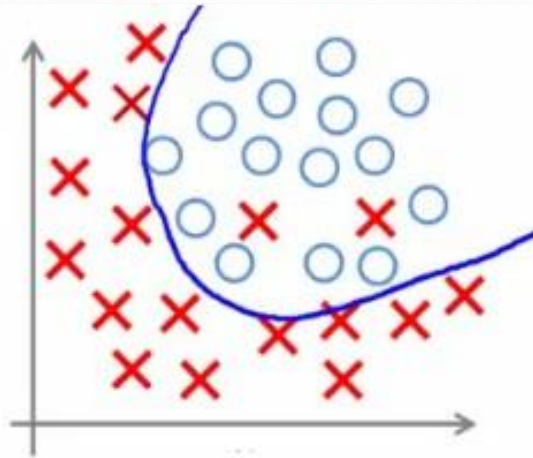
EVALUATION

CAPACIDADE E GENERALIZAÇÃO

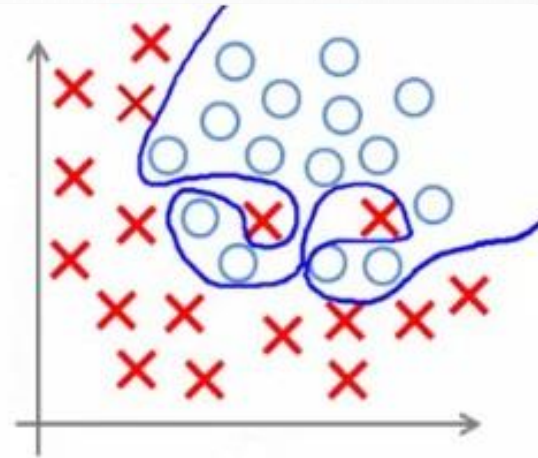


Under-fitting

(too simple to
explain the
variance)



Appropriate-fitting



Over-fitting

(forcefitting -- too
good to be true)

ESTIMANDO O ERRO DE GENERALIZAÇÃO

SINGLE SPLIT (GRUPO DE CONTROLE)

- Amostra é dividida entre treino e teste, mantendo um percentual das observações como grupo de teste externo ao treinamento.

LEAVE ONE OUT

- Uma única observação é deixada de fora a cada treinamento. N treinamentos são realizados para calcular a estatística de erro.

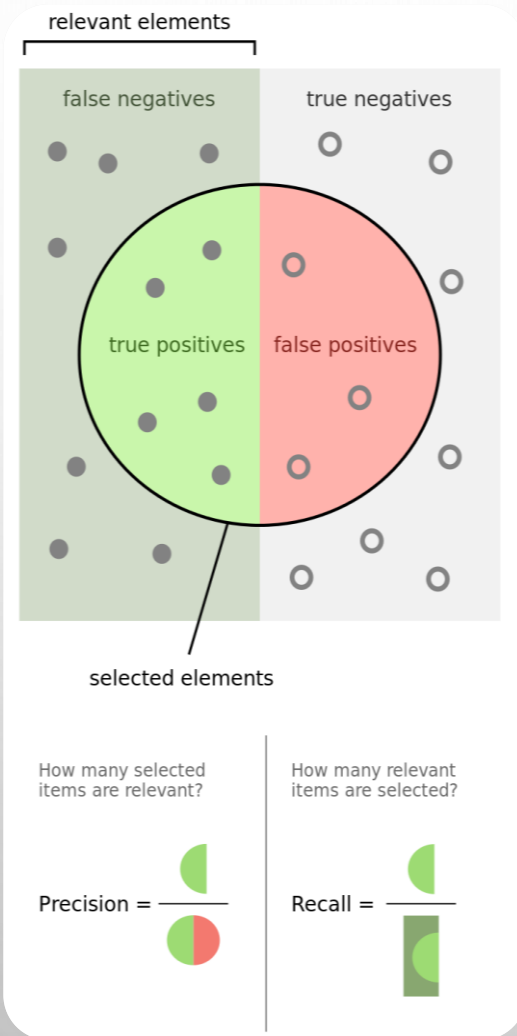
K FOLDS

- Amostra é dividida em K conjuntos. K treinamentos são realizados, mantendo um conjunto como fora-da-amostra.

BOOTSTRAPPING

- O algoritmo itera, amostrando aleatoriamente M observações, para a quantidade Q desejada de treinamentos.

FIGURAS DE MÉRITO CLASSIFICAÇÃO



Acurácia

- $(TP+TN)/(P+N)$

Taxa de Erro

- $1 - \text{Acurácia}$

Sensibilidade (Recall)

- $TP/(TP+FN)$

Especificidade

- $TN/(TN+FP)$

Precisão

- $TP/(TP+FP)$

Produto Sp

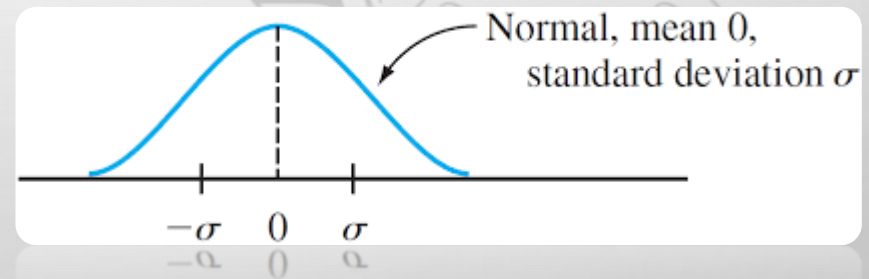
- $\text{SQRT}[\text{SQRT}(R1 * R2) * (R1 + R2)/2]$

FIGURAS DE MÉRITO - REGRESSÃO

- R QUADRADO

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- RESÍDUO NORMAL DE MÉDIA
ZERO E VARIÂNCIA CONSTANTE



The background is a light gray gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, some overlapping. The text is centered in the middle of the slide.

CRIANDO MODELOS SIMPLES DE MACHINE LEARNING I