Ciência de Dados
(Big Data Processing and Analytics)

Big Data Analytics – Mineração e Análise de Dados

Professor curador
Prof. Dr. Rogério de Oliveira

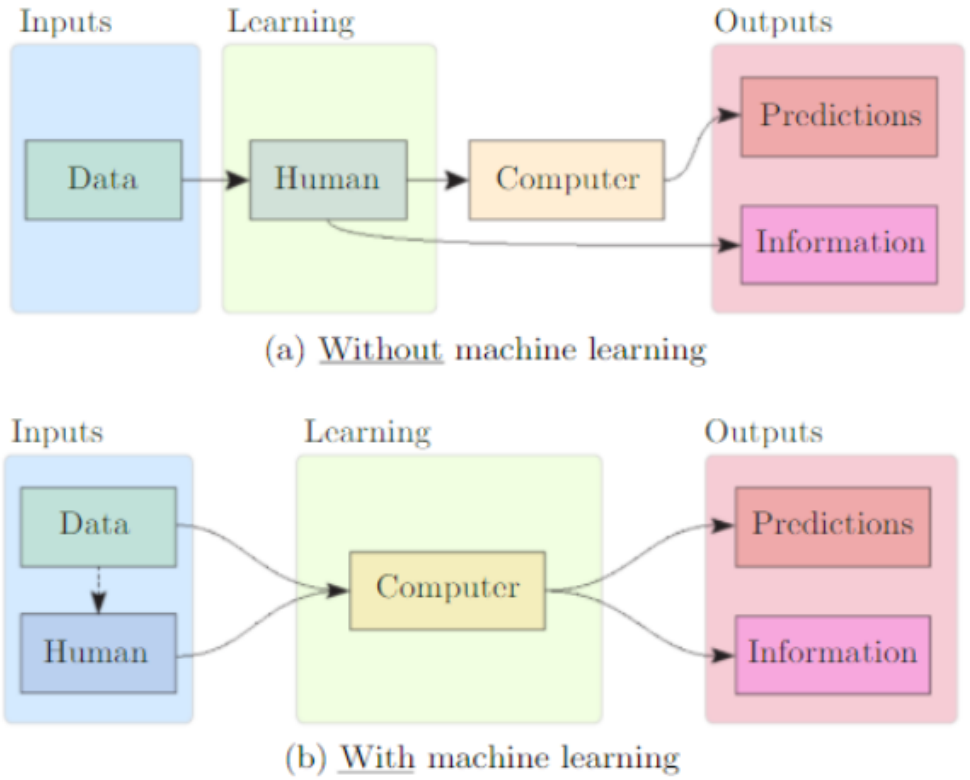
TRILHA 1
Introdução: Mineração, Ciência de Dados e o Aprendizado de Máquina

Parte A

Ciência de Dados e Análise de Dados Tradicional

Tipo	Ciência de Dados	Análise Tradicional
Tipo de Dados	Não estruturados	Estruturados (linhas e colunas)
Volume de Dados	Big Data (centenas de Terabytes)	Dezenas de Terabytes ou menos
Fluxo de Dados	Contínuo, Big Data	Estático
Métodos de Análise	Machine Learning	Visual, Hypothesis-Based
Propósito	Data-base Products	Internal Support Decision

Aprendizado de Máquina – Um novo Paradigma



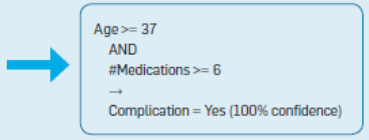
Goulet, J.-A. (2020), Probabilistic Machine Learning for Civil Engineers, The MIT press</h8>

Tipos de Padrões e Aplicações

Tarefa	Exemplos
Classificação	Fraud/not Fraud, Churn/Not Churn
Regressão	Preços de Imóveis, Aluguéis, Veículos
Clusterização	Segmentação de Clientes, Produtos, Documentos
Regras de Associação	Pacotes de produtos
Anomaly Detection	Mal funcionamento de dispositivos, Ataques cibernéticos
Matching	Recomendação de produtos ou amigos de uma rede social

Figure 3. Extracting interesting patterns in health outcomes from health-care system use.

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes



Exemplos

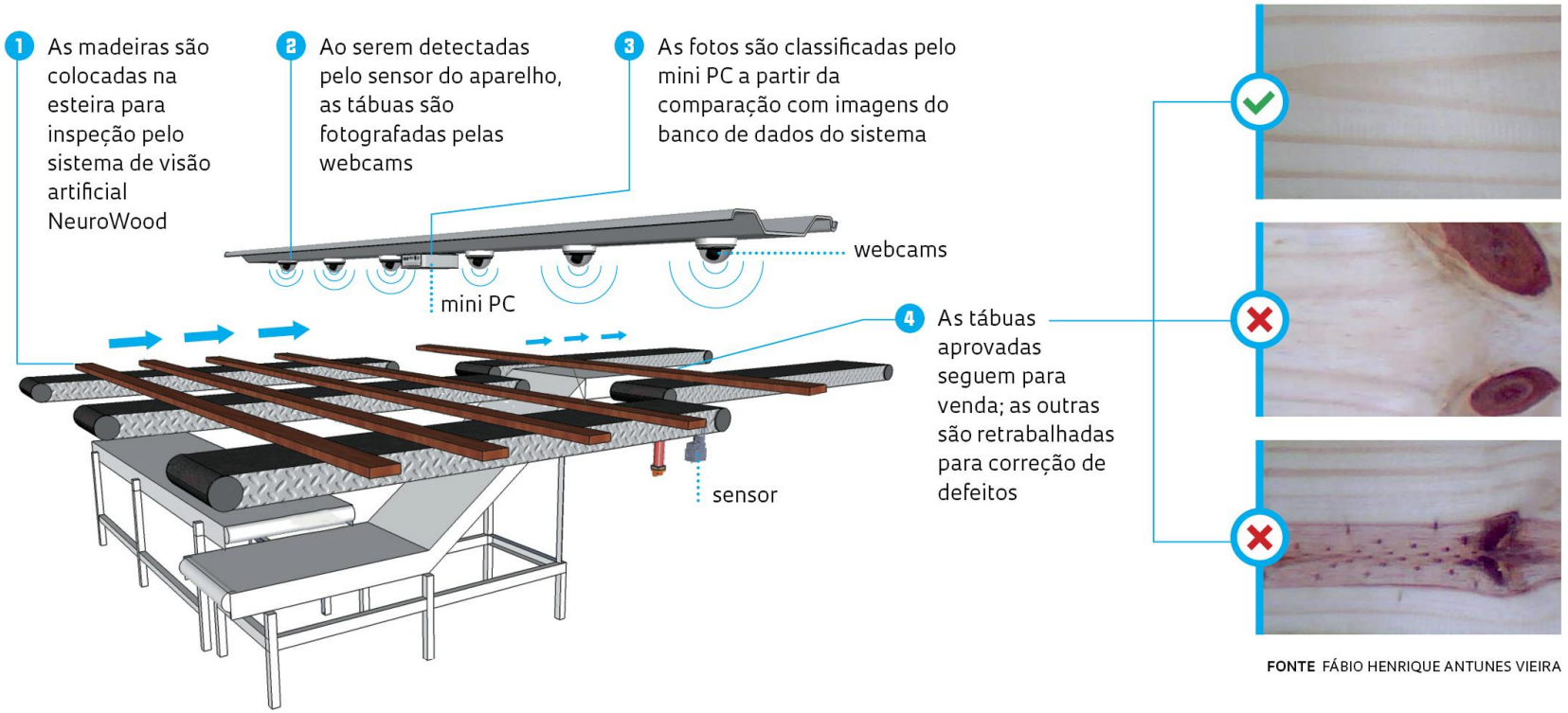
Machine Learning and Data Science Applications in Industry | https://github.com/ashishpatel26/Real-time-ML-Project

Accommodation & Food	Agriculture	Banking & Insurance
Biotechnological & Life Sciences	Construction & Engineering	Education & Research
Emergency & Relief	Finance	Manufacturing
Government and Public Works	Healthcare	Media & Publishing
Justice, Law and Regulations	Miscellaneous	Accounting
Real Estate, Rental & Leasing	Utilities	Wholesale & Retail

Exemplos

De olho na madeira

Saiba como funciona o aparelho que analisa e classifica as tábuas conforme a qualidade



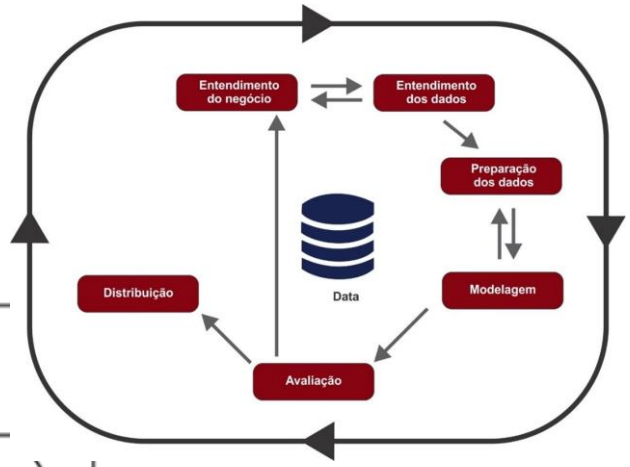
FONTE FÁBIO HENRIQUE ANTUNES VIEIRA

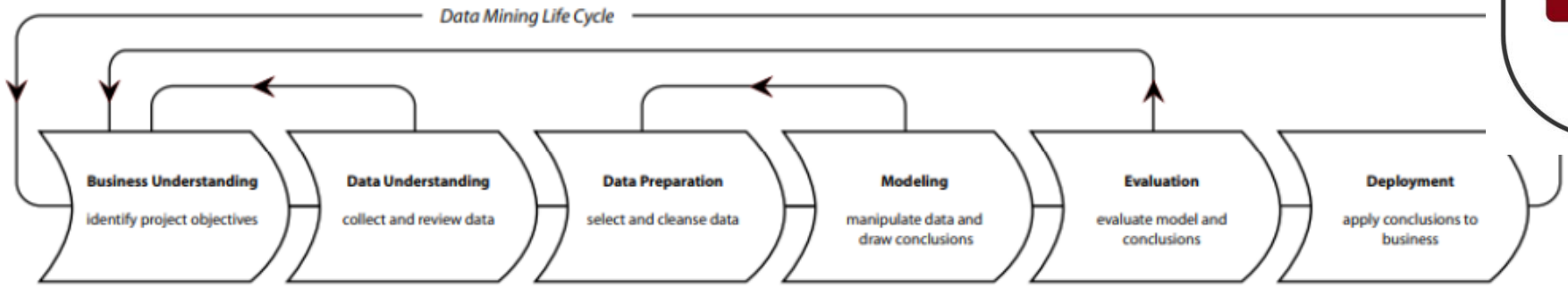
INFOGRÁFICO ANA PAULA CAMPOS **ILUSTRAÇÃO** ANDRÉ LUIS DEBIASO ROSSI **FOTOS** FÁBIO HENRIQUE ANTUNES VIEIRA

TRILHA 1
Introdução: Mineração, Ciência de Dados e o Aprendizado de Máquina

Parte B

Fases do CRISP-DM





Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

Aprendizado Supervisionado

Tarefas de Aprendizado Supervisionado

Breast Cancer Data

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
M	15.340	14.26	102.50	704.4
B	12.880	28.92	82.50	514.3
M	17.080	27.15	111.20	930.9
B	16.140	14.86	104.30	800.0
M	13.480	20.82	88.40	559.2
B	14.470	24.99	95.81	656.4
B	12.490	16.85	79.19	481.6
M	23.210	26.97	153.50	1670.0
B	11.620	18.18	76.38	408.8
B	9.787	19.94	62.11	294.5
M	21.750	20.99	147.30	1491.0
B	10.800	21.98	68.79	359.9
M	25.730	17.46	174.20	2010.0
B	11.870	21.54	76.83	432.0
B	7.691	25.44	48.34	170.4

y

X

y

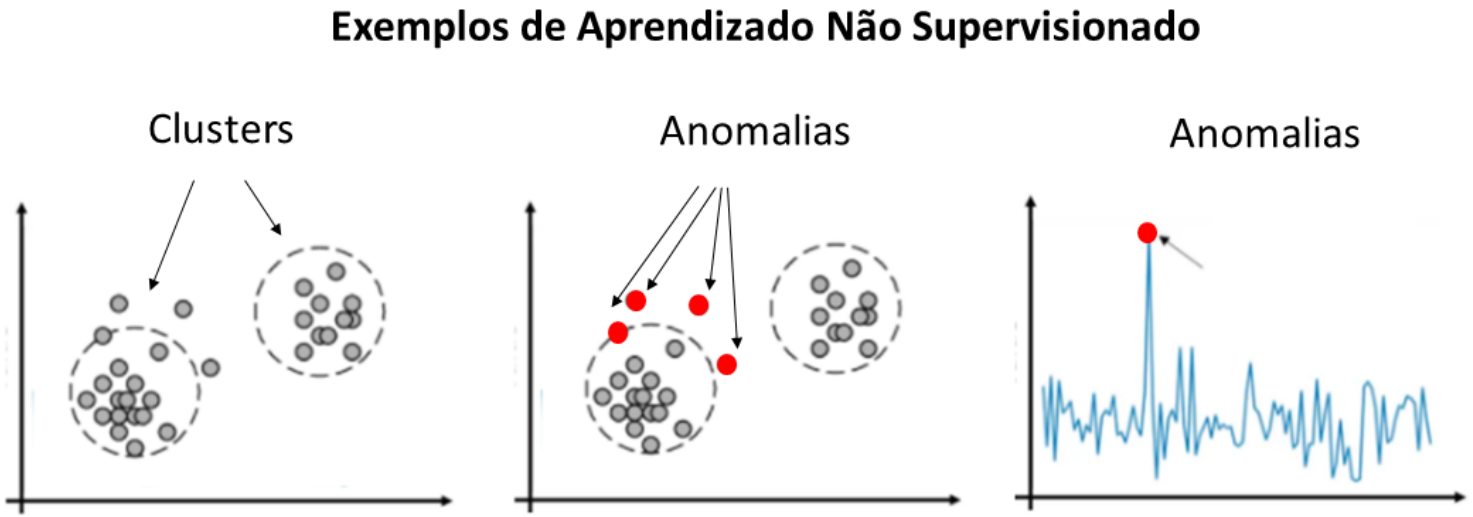
Classificação

Árvores de Decisão
Regressão Logística
K-Vizinhos mais Próximos
Support Vector Machines

Regressão

Regressão Linear
Regressão Polinomial
Modelos Neurais para Regressão

Aprendizado Não Supervisionado



Universidade Presbiteriana
Mackenzie