

Introdução

Mineração de Dados

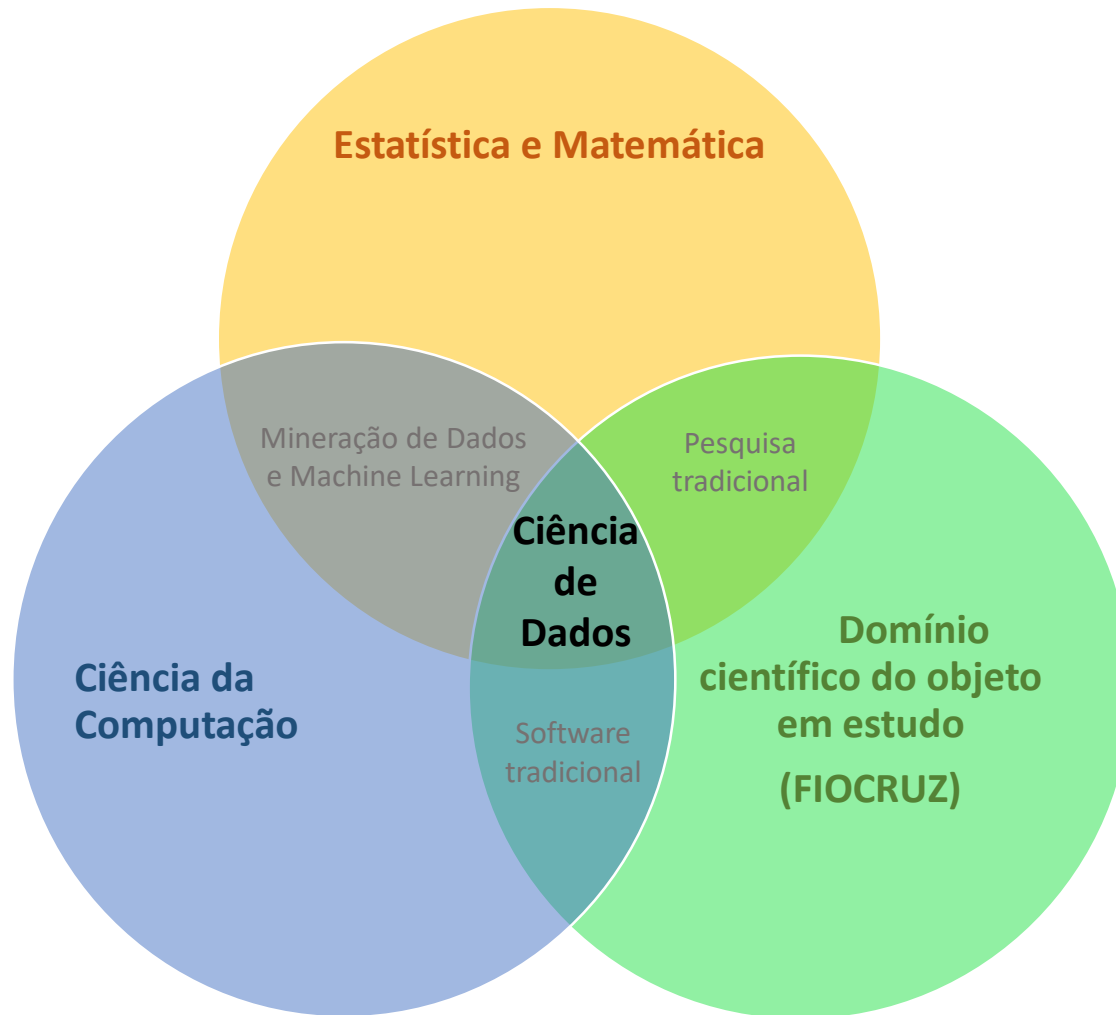
Machine Learning

Prof. Dr. Marcel Pedroso
Pesquisador em Saúde Pública

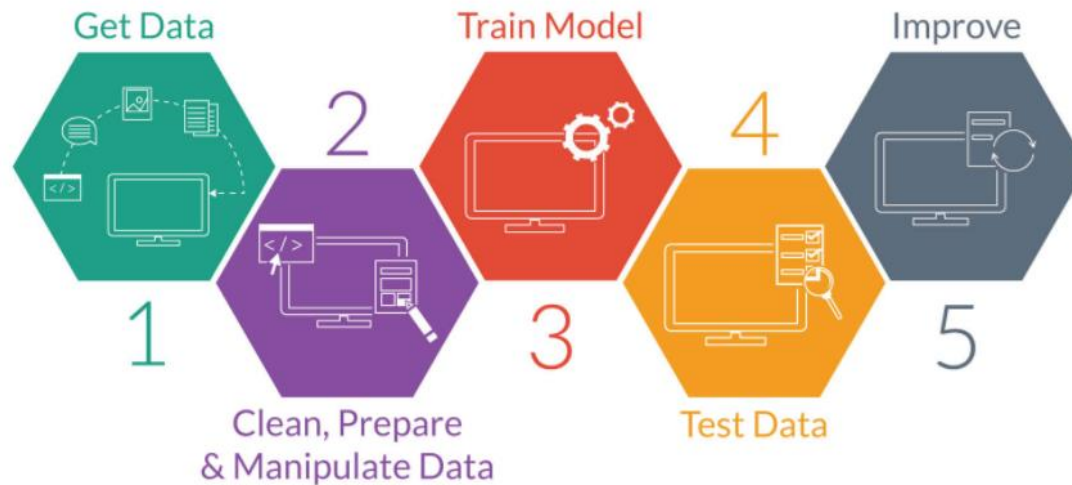
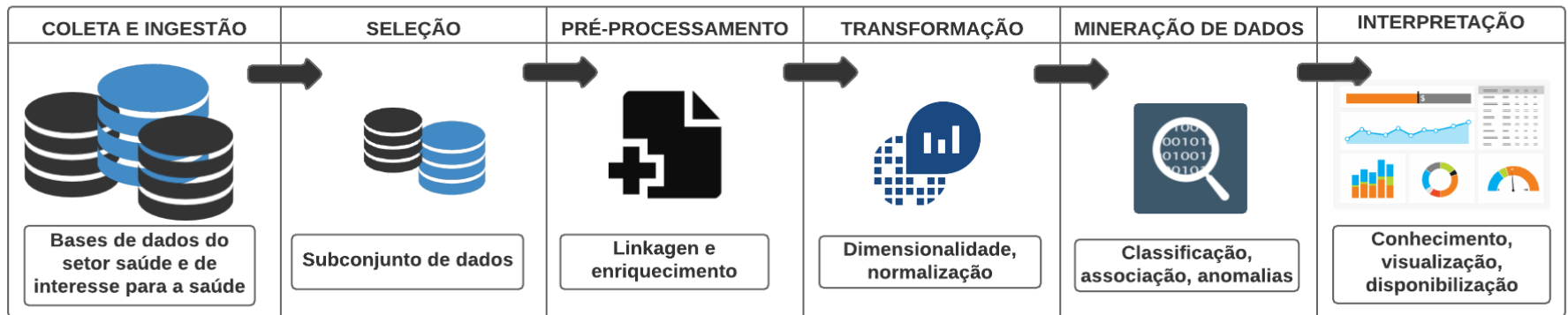
DEFINIÇÃO DE CIÊNCIA DE DADOS

Ciência de Dados é um conjunto de estratégias, ferramentas e técnicas que busca reunir equipes multidisciplinares formadas por pesquisadores com conhecimento substantivo do problema em análise - no nosso caso saúde pública - estatísticos, matemáticos e cientistas da computação. Trata-se de um campo de estudo promissor e destaca-se pela capacidade de auxiliar a descoberta de informação útil a partir de grandes bases de dados e a tomada de decisão orientada por dados

Ciência de Dados - Principais Componentes



Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases – KDD*)



The Periodic Table of Data Science

An overview of key companies, resources and tools in data science (as of 4/12/2017)



Dc DataCamp	Ga General Assembly	Sd Strata Data
Sb SpringBoard	M Metis	Od ODSC
Ex Edx	Di Data Incubator	Tc Tableau Conference
C Coursera	In Insight	U Udacity
Uda Udacity	Dsa NYC Data Science Academy	Pd PyData
Ude Udemy	G Galvanize	Paw Predictive Analytics World
Ps Pluralsight	Dsg Data Science for Social Good	Kdd ACM SIGKDD Conference
Ly Lynda	Dsy Data Society	Tpc Teradata Partners Conference
Ti TeamTreeHouse	Dsj Data Science Dojo	Icd IEEE International Conference on Data Mining
Bdu Big Data University		

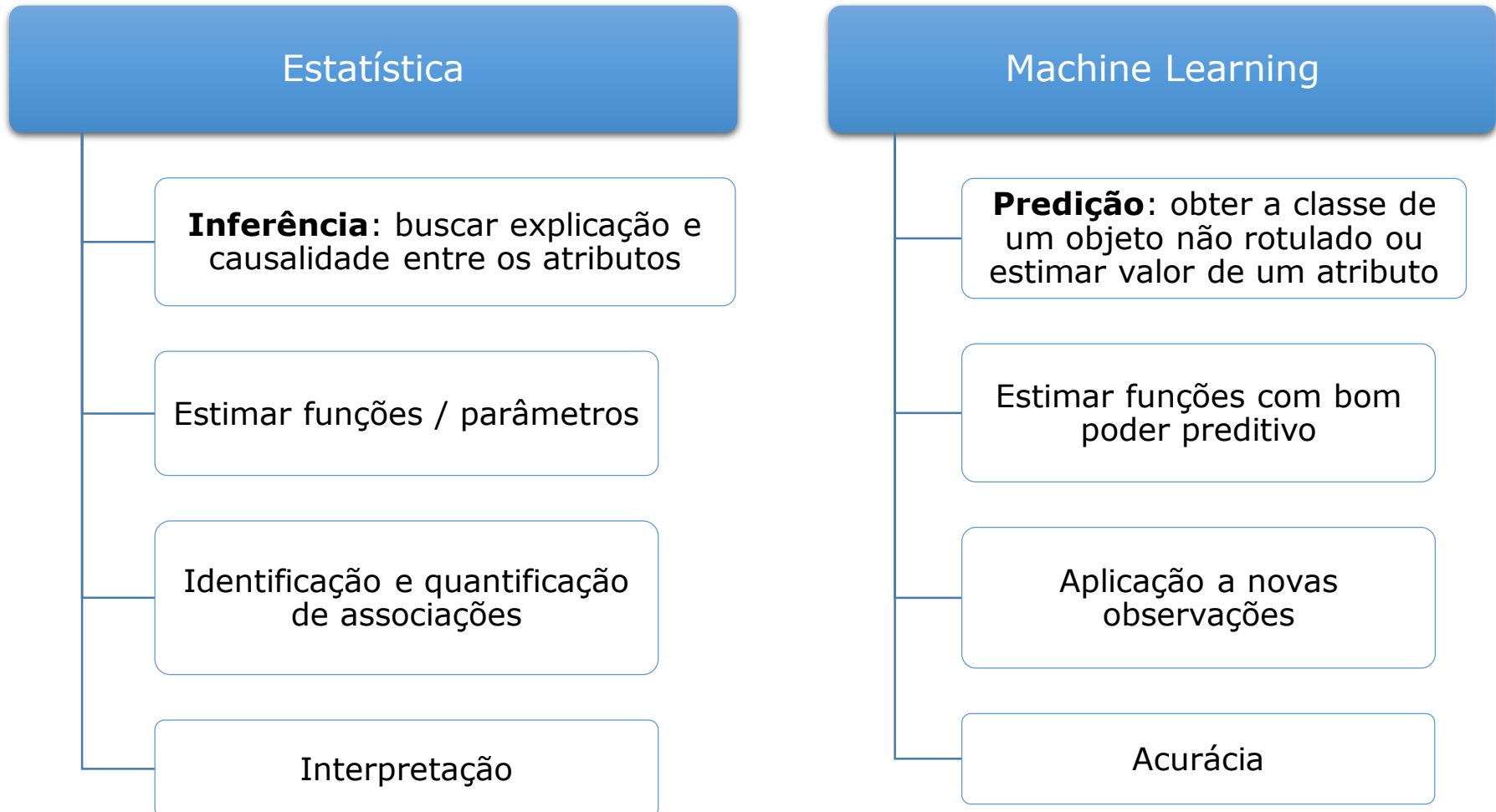
Courses	Data	Search & Data Management	Collaboration	News, Newsletters & Blogs
Boot camps	Projects & Challenges, Competitions	Machine Learning & Stats	Community & Q&A	Podcasts
Conferences	Programming Languages & Distributions	Data Visualization & Reporting		

Py Python	Js JavaScript	Vb Visual Basic	Pgs PostgreSQL	Sli SQLite	Ab Apache Hadoop	W Weka	Bml BigML	Kn Knime	Sm Spark MLlib	Pb Power BI	Obi Oracle BI	Shn Shiny	Ddl Domino Data Lab	De Data Science Experience
R R	Cp C++	Sc Scala	Ar Amazon Redshift	Bq Google BigQuery	Hw Hortonworks	O Oracle	Dar DataRobot	Lib LibSVM	Ho H2O	Bo BusinessObjects	Alt Alteryx	Mpl Matplotlib	Nt Nteract	Rs Rstudio
S SQL	Pl Perl	Ca Cassandra	Hb HBase	Td Teradata	Cl Cloudera	Mss Microsoft SQL server	Rm RapidMiner	Mat Mathematica	Th Theano	Sp Spotfire	Sav SAS Visual Analytics	Ply Plotly	Ro Rodeo	Be Beaker Notebook
B Bash	Mr Microsoft R Open	P Pig	Mdb Mongo DB	To Toad	Aem Amazon Elastic Mapreduce	Spl Splunk	Cho Chorus	Mah Mahout	Aml Azure Machine Learning	Ql Qlikview	Po PowerPivot	Me Microsoft Excel	Spy Spyder	Ze Apache Zeppelin
Mtl Matlab	Cy Canopy	Im Impala	K Kafka	Ms MySQL	Mar MapR	Sr Solr	Tf Tensorflow	St Stata	D D3	Co Cognos	Gch Google Charts	Pe Pentaho	Dst Data Science Studio	Ju Jupyter
J Java	An Anaconda	Sp Spark	Hi Hive	Idb IBM DB2	Lu Lucene	El ElasticSearch	Sk Scikit-Learn	Da Dato/Graphlab	My Microstrategy	Aa Adobe Analytics	T Tableau	B Bokeh	Db Databricks notebook	Gh Github

Dw Data.world	Q Quandl	Fte FiveThirtyEight	Sa Socrata	Gp Google Public	Dg Data.gov	K Kaggle
St Statista	Uci UCI Machine Learning Repository	Wb World Bank	At Academic Torrents	Bf Buzzfeed	Dk DataKind	Dd DrivenData
Re Reddit	So Stack Overflow	Cv Cross Validated	Qu Quora	Av Analytics Vidhya	Dse Data Science Stack Exchange	
Mu Meetup	Rdm RDataMining					

Kdn KDnuggets	Ibd insideBIGDATA
Rb R-Bloggers	Pp PlanetPython
Hn HackerNews	Dt DataTau
Dsc Data Science Central	Dsr Data Science Roundup
Dsw Data Science Weekly	Or O'Reilly
Dr Data Elixir	Pw Python Weekly
Rw R Weekly	Pd Partially Derivative
Bds Becoming a Data Scientist	Tm Talking Machines
Ds Data Stories	Dsk Data Skeptic
Ld Linear Digressions	Ns Not So Standard Deviations

Inferência *versus* Predição



DEFINIÇÃO MINERAÇÃO DE DADOS

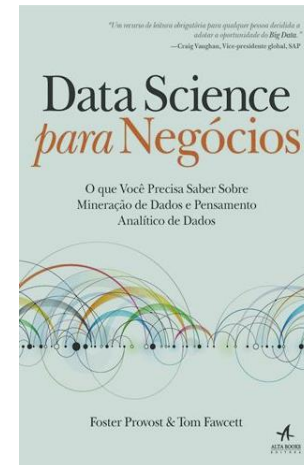
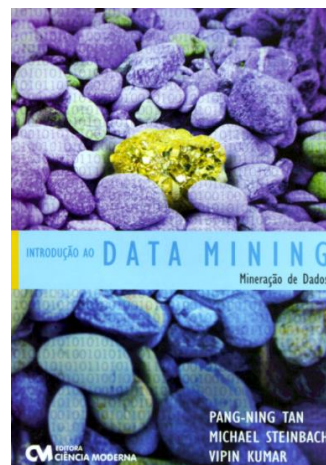
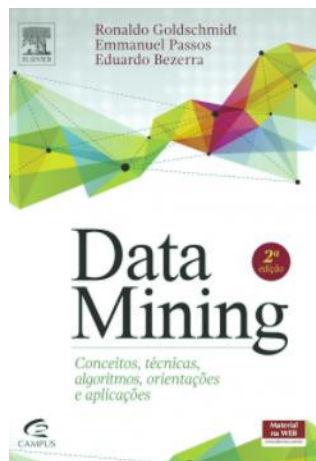
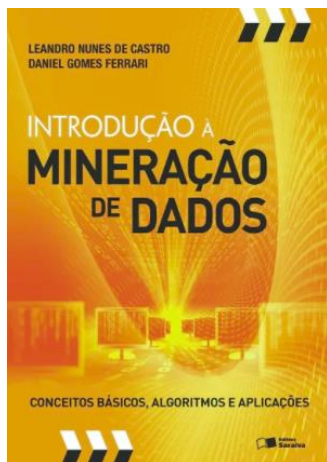
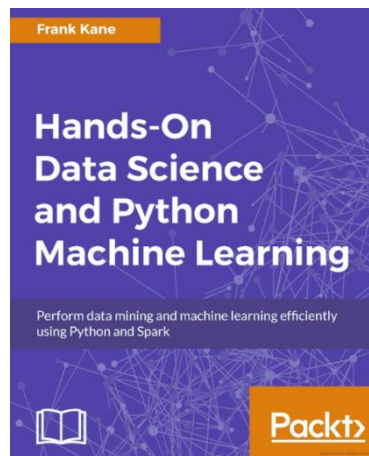
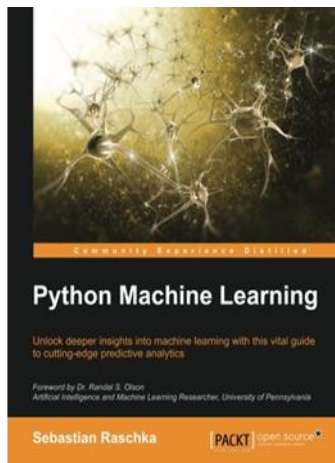
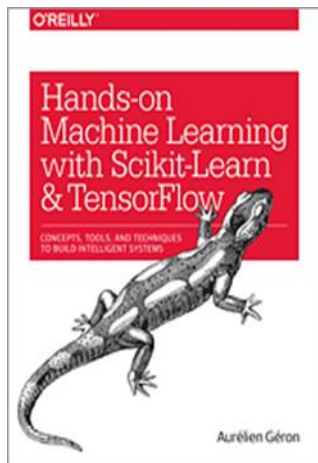
- **Etapas** do processo de **KDD**
 - **Aplicação de algoritmos** capazes de **extrair conhecimento** a partir de dados pré-processados
-
- Análise Descritiva (medidas de distribuição, tendência central e variância)
 - Análise Preditiva (classificação e regressão)
 - Análise de Agrupamento (segmentação de bases de dados)
 - Detecção de anomalias e associação

DEFINIÇÃO MACHINE LEARNING / APRENDIZADO DE MÁQUINA

- Área de pesquisa vinculada ao **KDD e Mineração de Dados**
- **Implementação de algoritmos** via programas computacionais
- Capazes de **automaticamente** melhorar seu desempenho (classificação ou regressão)
- Por meio da **experiência** de acordo com critérios prévios
- Aprendizado é um **processo interativo / iterativo**

Palavras-Chave: Big Data; Metodologia; Estatística e Dados Numéricos; Brasil

Livros recomendados



“Utilização intensiva de dados para resolução de problemas/desafios concretos”

what I really do

DEFINIÇÃO MACHINE LEARNING / APRENDIZADO DE MÁQUINA

PARADIGMAS DE APRENDIZADO

- **Aprendizado Supervisionado**

Baseado em um conjunto de objetos para os quais as saídas **desejadas são conhecidas** (exemplos Árvores de Decisão, Regressão linear e logística, k-NN, naïve Bayes, Redes Neurais Artificiais, SVM, Regras de Classificação)

- **Aprendizado Não Supervisionado**

Baseado em um conjunto de objetos para os quais as saídas desejadas **NÃO** são conhecidas ou a tarefa é de **categorização** (K-means, G-means, DBSCAN, Redes Neurais Artificiais)

Tarefas de PREDIÇÃO (Aprendizagem Supervisionada)

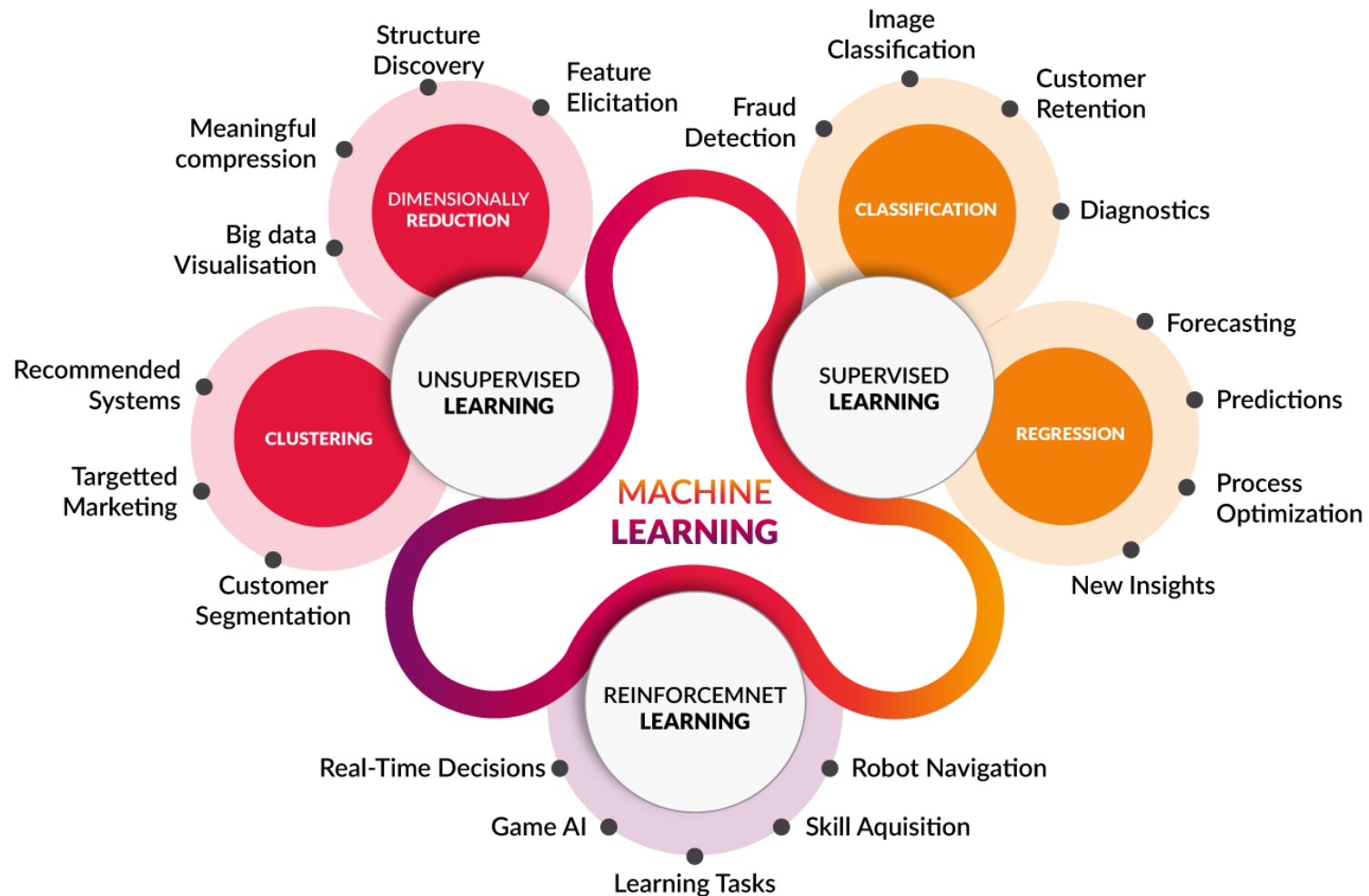
- **CLASSIFICAÇÃO:** utilizada para prever valores discretos

Exemplos: Tipo de Discriminação em Serviços de Saúde, Diagnóstico de determinada doença, Tipo de tumor, Reclassificação de óbitos com causa mal definida

- **REGRESSÃO (estimação):** utilizada para prever valores contínuos

Exemplos: Taxa de Mortalidade Infantil, Número de óbitos por causas externas,
Proporção de Internações por DSAI

Diagrama contendo as principais técnicas Machine Learning



Avaliação de desempenho (generalização) da (f) PREDIÇÃO

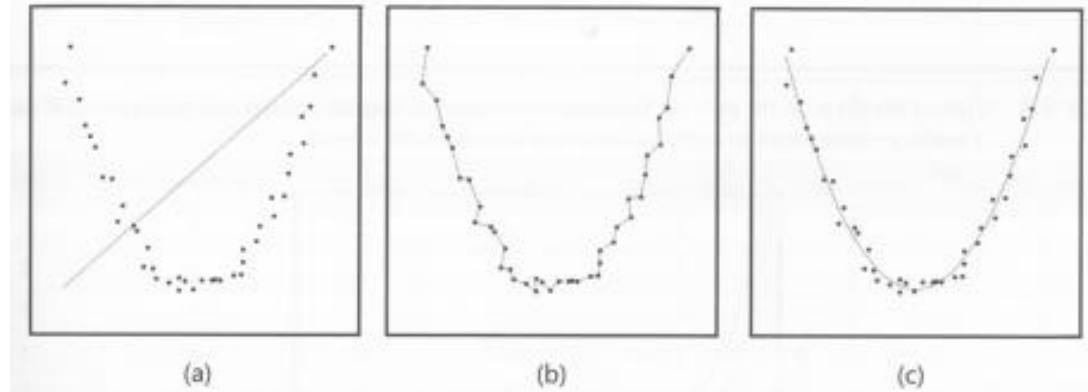
- Em geral é o **erro calculado** entre a saída fornecida pelo modelo e a saída esperada
- É construída em um **conjunto de treino** com a classe rotulada

TREINO

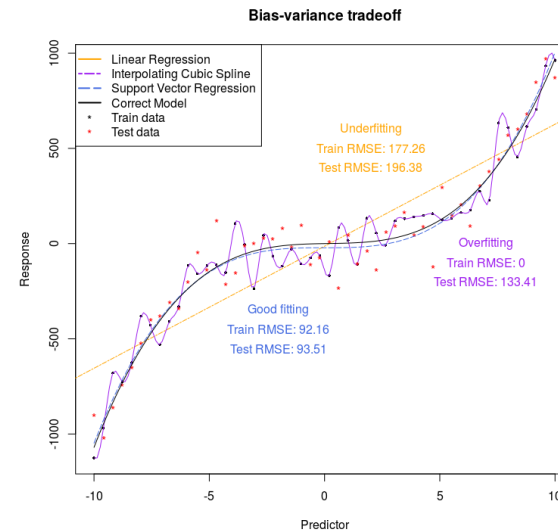
TESTE

Município	IBGE	Pop Vulneravel	Taxa de analf	% crianças extr pobres	Renda PC extr pobres	Mortalidade infantil	Mortalidade infantil Predita	
Abadia dos Dourados (MG)	310010	319.00	9.12	2.83	14.48	14.80	14.50	Ajuste da F de Predição
Abaeté (MG)	310020	1151.00	8.72	3.76	30.34	14.00	20.40	
Abre Campo (MG)	310030	1172.00	12.27	11.48	23.05	16.30	15.80	
Acaiaca (MG)	310040	315.00	13.94	11.10	30.12	16.30	18.80	
Açucena (MG)	310050	842.00	19.81	22.91	24.89	14.30	16.40	
Água Boa (MG)	310060	1692.00	27.42	18.34	36.37	18.70	19.20	
Água Comprida (MG)	310070	110.00	7.90	2.27	26.80	13.00	17.50	
Aguanil (MG)	310080	257.00	13.21	4.62	37.07	14.30	11.50	
Águas Formosas (MG)	310090	2394.00	24.40	21.89	38.93	17.50	17.50	
Águas Vermelhas (MG)	310100	1604.00	27.57	16.96	34.10	17.80	17.80	
Aimorês (MG)	310110	1531.00	14.49	9.48	38.33	17.60	14.30	
Aiuruoca (MG)	310120	412.00	14.07	3.30	29.52	12.10	18.70	
Alagoa (MG)	310130	266.00	14.44	7.06	27.93	15.40	13.00	
Albertina (MG)	310140	111.00	11.94	0.00	68.00	17.30	14.30	
Além Paraíba (MG)	310150	1912.00	6.36	5.88	37.97	13.20	17.50	
Alfenas (MG)	310160	2886.00	5.77	1.23	35.62	14.60	14.60	
Alfredo Vasconcelos (MG)	310163	589.00	9.94	6.39	35.92	15.30	15.30	
Almenara (MG)	310170	3493.00	21.80	15.72	37.43	18.30	18.30	
Alpercata (MG)	310180	533.00	16.13	7.09	30.42	16.20	16.20	
Alpinópolis (MG)	310190	1015.00	7.69	2.67	41.31	12.80	12.80	
Alterosa (MG)	310200	1255.00	10.40	5.52	8.51	14.10	14.10	
Alto Caparaó (MG)	310205	672.00	11.29	3.57	43.20	14.50	14.50	
Alto Jequitibá (MG)	315350	1285.00	10.58	11.50	50.78	20.00	20.00	
Alto Rio Doce (MG)	310210	1023.00	14.66	19.11	32.58	15.80	15.80	
Alvarenga (MG)	310220	360.00	18.13	27.84	25.07	18.80	12.80	
Alvinópolis (MG)	310230	1383.00	11.24	3.22	35.00	16.40	14.10	
Alvorada de Minas (MG)	310240	306.00	20.60	33.07	32.92	19.20	14.50	
Amparo do Serra (MG)	310250	646.00	16.81	15.86	32.64	17.50	20.00	
Andradas (MG)	310260	816.00	8.47	0.69	29.71	11.50	15.80	
Andrelândia (MG)	310280	896.00	10.86	4.72	28.17	16.80	18.80	
Angelândia (MG)	310285	1512.00	25.48	12.83	31.86	23.50	23.50	
Antônio Carlos (MG)	310290	942.00	11.56	5.79	23.16	14.90	14.90	
Antônio Dias (MG)	310300	832.00	15.74	9.77	34.16		13.90	Medida de Acurácia
Antônio Prado de Minas (MG)	310310	153.00	11.08	2.25	45.17		15.60	
Araçai (MG)	310320	150.00	10.29	5.15	32.06		18.10	
Aracitaba (MG)	310330	197.00	12.60	12.81	39.95		16.30	
Araçuaí (MG)	310340	2265.00	17.35	18.88	30.75		16.50	
Araguari (MG)	310350	3703.00	4.80	2.01	32.75		11.71	
Aranitina (MG)	310360	227.00	10.34	3.74	31.04		15.20	
Araponga (MG)	310370	1554.00	22.25	12.40	35.26		23.10	
Araporã (MG)	310375	360.00	8.29	0.21	16.16		17.30	
Arapuá (MG)	310380	114.00	7.98	1.95	20.15		12.10	
Araújos (MG)	310390	166.00	8.10	0.18	43.77		13.40	
Araxá (MG)	310400	2135.00	4.20	1.76	35.99		13.10	

- **Equilíbrio entre viés e variância**



- ERRO QUADRÁTICO MÉDIO
- SOMA DOS ERROS QUADRÁTICOS
- RAIZ DO ERRO QUADRÁTICO MÉDIO
- ERRO ABSOLUTO MÉDIO
- ERRO QUADRÁTICO RELATIVO
- RAIZ DO ERRO
- COEFICIENTE DE DETERMINAÇÃO – R^2



Algumas das principais métricas para avaliação de desempenho – Tarefa CLASSIFICAÇÃO

Matriz de Confusão (1.000 e-mails)		Classe Predit	
		Positivo (Spam)	Negativo (Normal)
Classe Observada	Positivo (Spam)	VP (100)	FN (90)
	Negativo (Normal)	FP (10)	VN (800)

- **ACURÁCIA:** proporção de predições corretas (positivas e negativas)

$$(VP + VN) / (VP + FP + VN + FN) = (100 + 800) / (100 + 10 + 800 + 90) = \mathbf{0.900}$$

- **PRECISÃO:** proporção de predições corretas (positivas)

$$(VP) / (VP + FP) = (100) / (100 + 10) = \mathbf{0.909}$$

- **RECALL (SENSIBILIDADE):** proporção de VP sobre os realmente positivos

$$(VP) / (VP + FN) = (100) / (100 + 90) = \mathbf{0.526}$$

- **ESPECIFICIDADE:** proporção de VN sobre os realmente negativos

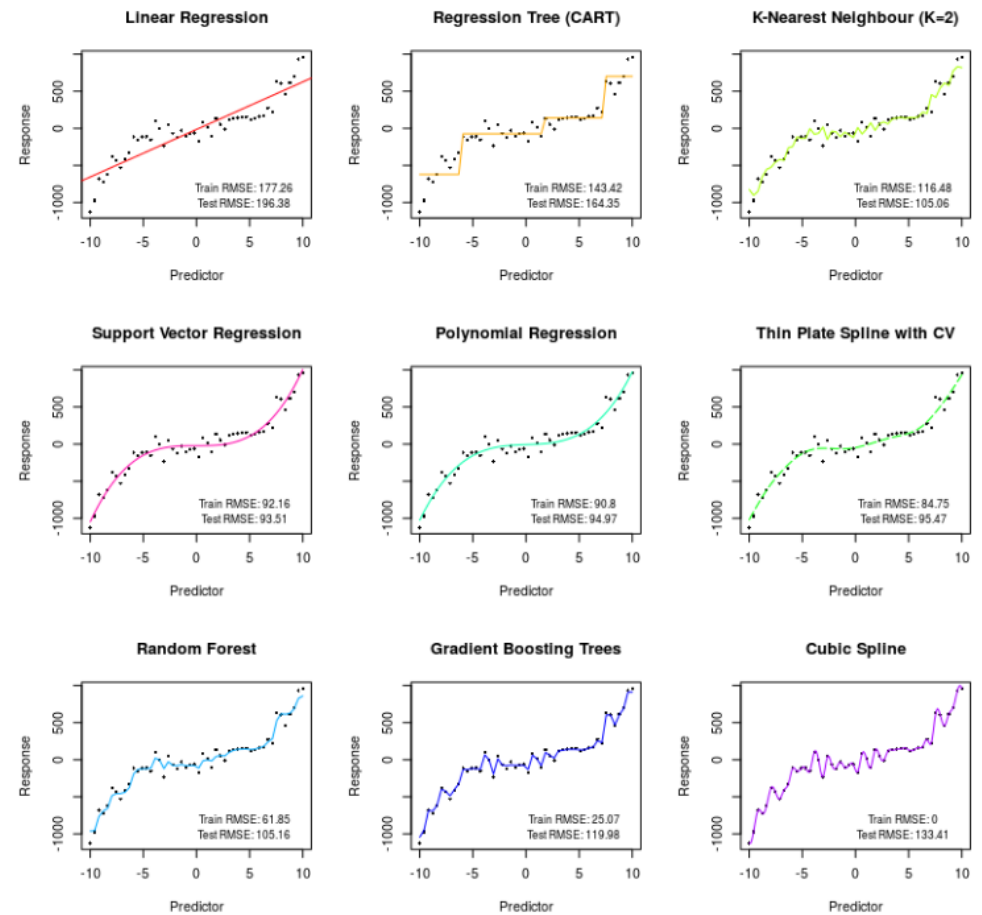
$$(VN) / (VN + FP) = (800) / (800 + 10) = \mathbf{0.987}$$

- **F1 SCORE** = Média harmônica entre PRECISÃO e RECALL

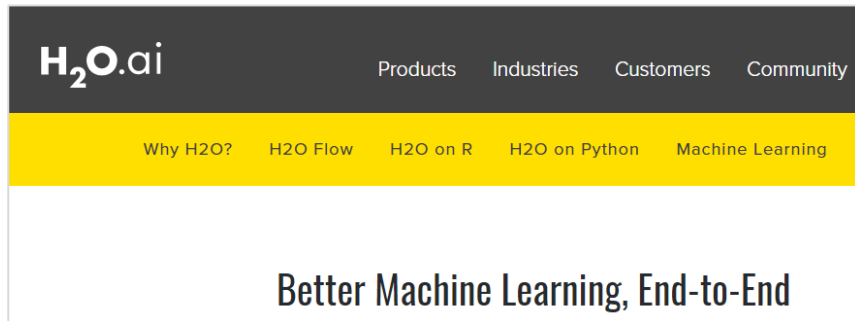
$$(2 * \text{PRECIS\~{A}\~{O}} * \text{RECALL}) / (\text{PRECIS\~{A}\~{O}} + \text{RECALL}) = \mathbf{0.666}$$

Erros em Aprendizado Supervisionado

- **ERRO DE REPRESENTAÇÃO** (aproximação ou **efeito viés**)
 - **flexibilidade** do modelo em **ajustar a função de predição** (conjunto de treinamento)
 - **adequação** do modelo selecionado para **representar e modelar** os dados disponíveis
- **ERRO DE GENERALIZAÇÃO** (estimação ou **efeito variância**)
 - Treinamento em excesso do modelo por gerar **sobre ajuste** (overfitting)
 - Incorporação dos **ruídos e inconsistências** da base de treino
 - Modelo muito rígido ou simples pode gerar **sub ajuste** (underfitting)
- Existe um *tradeoff* entre **viés** (erro no treino) e **variância** (do treino em relação ao teste)
- A qualidade do modelo de predição depende fortemente do **equilíbrio** entre viés e variância



MINERAÇÃO DE DADOS E MACHINE LEARNING VISUAL

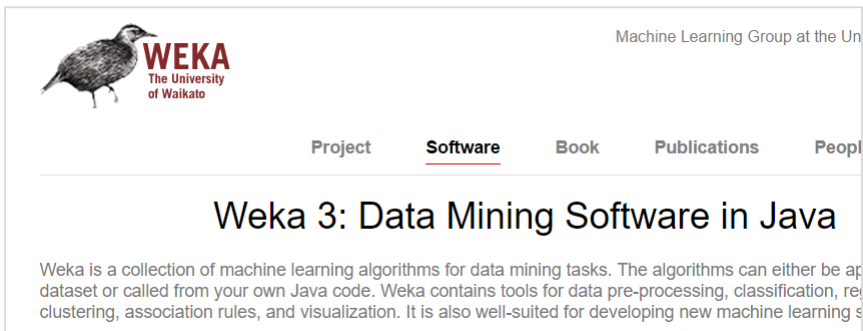



H2O.ai

Products Industries Customers Community

Why H2O? H2O Flow H2O on R H2O on Python Machine Learning

Better Machine Learning, End-to-End



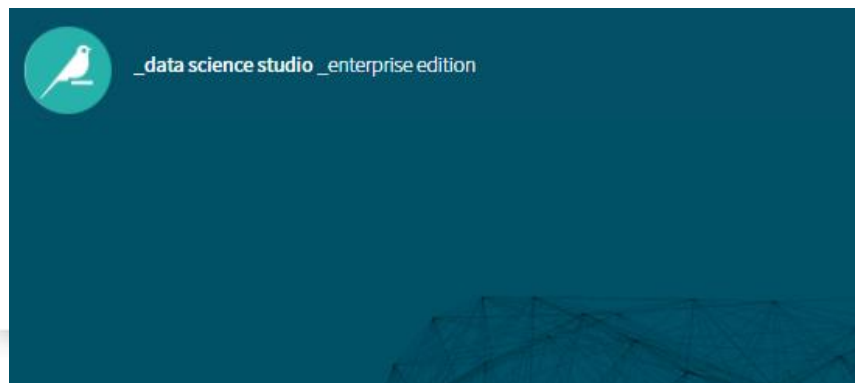
 WEKA
The University of Waikato


Machine Learning Group at the University of Waikato

Project Software Book Publications People

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning systems.



 _data science studio_ enterprise edition



Open for Innovation[®]

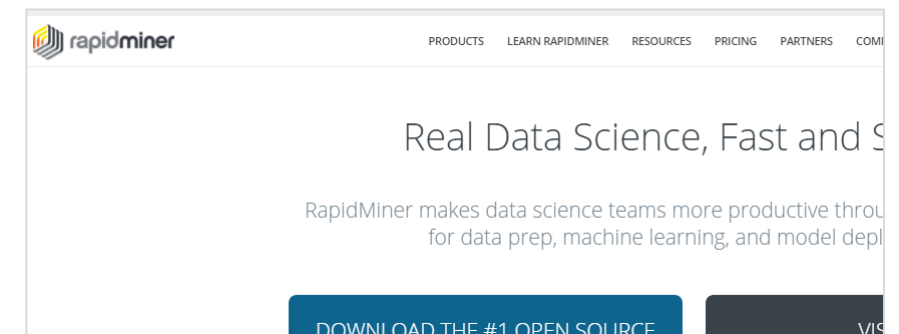
KNIME


PRODUCTS / SOLUTIONS / LEARNING / PARTNERS / COMMUNITY / ABOUT

Open for Innovation

Navigate complex data with the agility and freedom that only an open platform can bring

Learn More



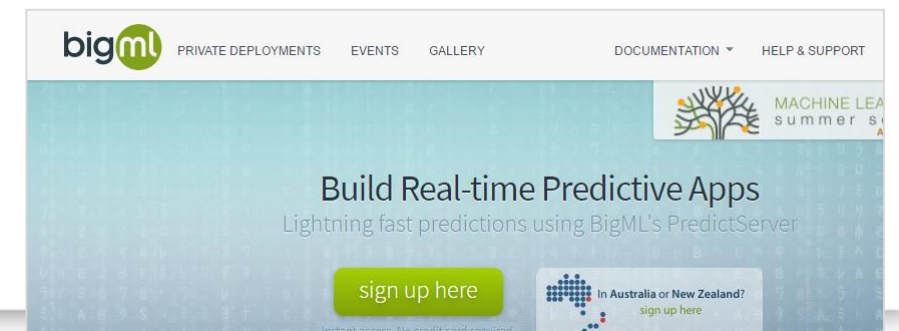
 rapidminer

PRODUCTS LEARN RAPIDMINER RESOURCES PRICING PARTNERS COMMUNITY

Real Data Science, Fast and Simple

RapidMiner makes data science teams more productive through automation for data prep, machine learning, and model deployment.

DOWNLOAD THE #1 OPEN SOURCE DATA SCIENCE TOOL VISIT US



bigml

PRIVATE DEPLOYMENTS EVENTS GALLERY DOCUMENTATION ▾ HELP & SUPPORT

MACHINE LEARNING summer school

Build Real-time Predictive Apps

Lightning fast predictions using BigML's PredictServer

sign up here

Instant access. No credit card required.

In Australia or New Zealand? sign up here

Prática

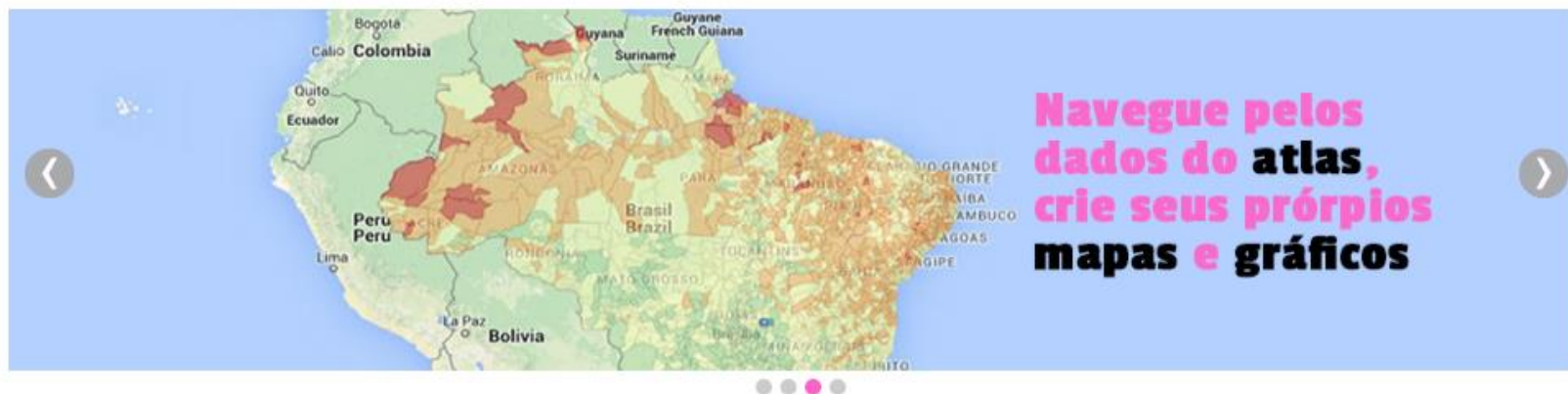
Predição Mortalidade Infantil

Prof. Dr. Marcel Pedroso
Pesquisador em Saúde Pública

FONTE DOS DADOS



Atlas do Desenvolvimento
Humano no Brasil

[HOME](#)[O ATLAS](#)[PERFIL](#)[CONSULTA](#)[MAPA](#)[RADAR IDHM](#)[ÁRVORE](#)[RANKING](#)[DOWNLOAD](#)

Perfil

Consulte o perfil
da sua localidade

estado

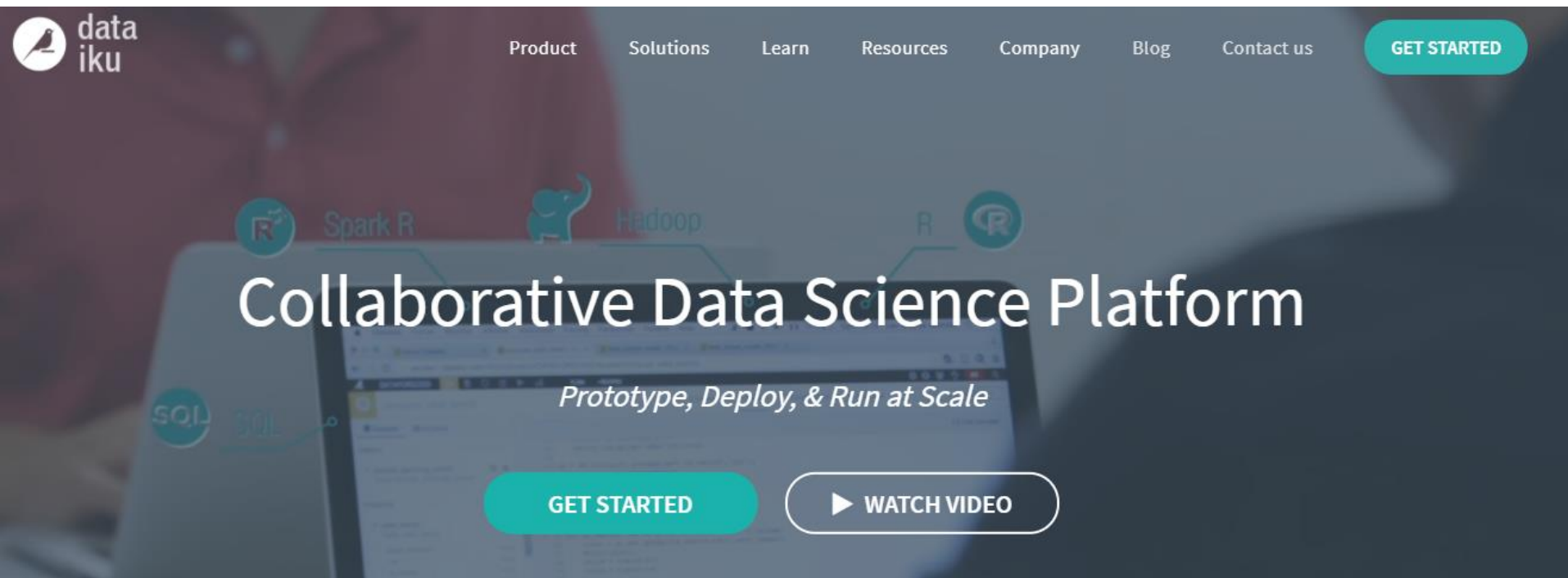
região metropolitana

município


unidade de desenvolvimento humano

Busca

FERRAMENTA PARA ANÁLISE



The hero section of the Dataiku website features a dark background with a blurred image of a person's face. Overlaid on this is a laptop screen displaying a data science interface. Various data science logos are scattered around the laptop: Spark R, Hadoop, R, and SQL. The main headline is 'Collaborative Data Science Platform' in large white text, followed by the tagline 'Prototype, Deploy, & Run at Scale' in a smaller, italicized font. At the bottom, there are two prominent buttons: a teal 'GET STARTED' button and a white 'WATCH VIDEO' button with a play icon.

 dataiku

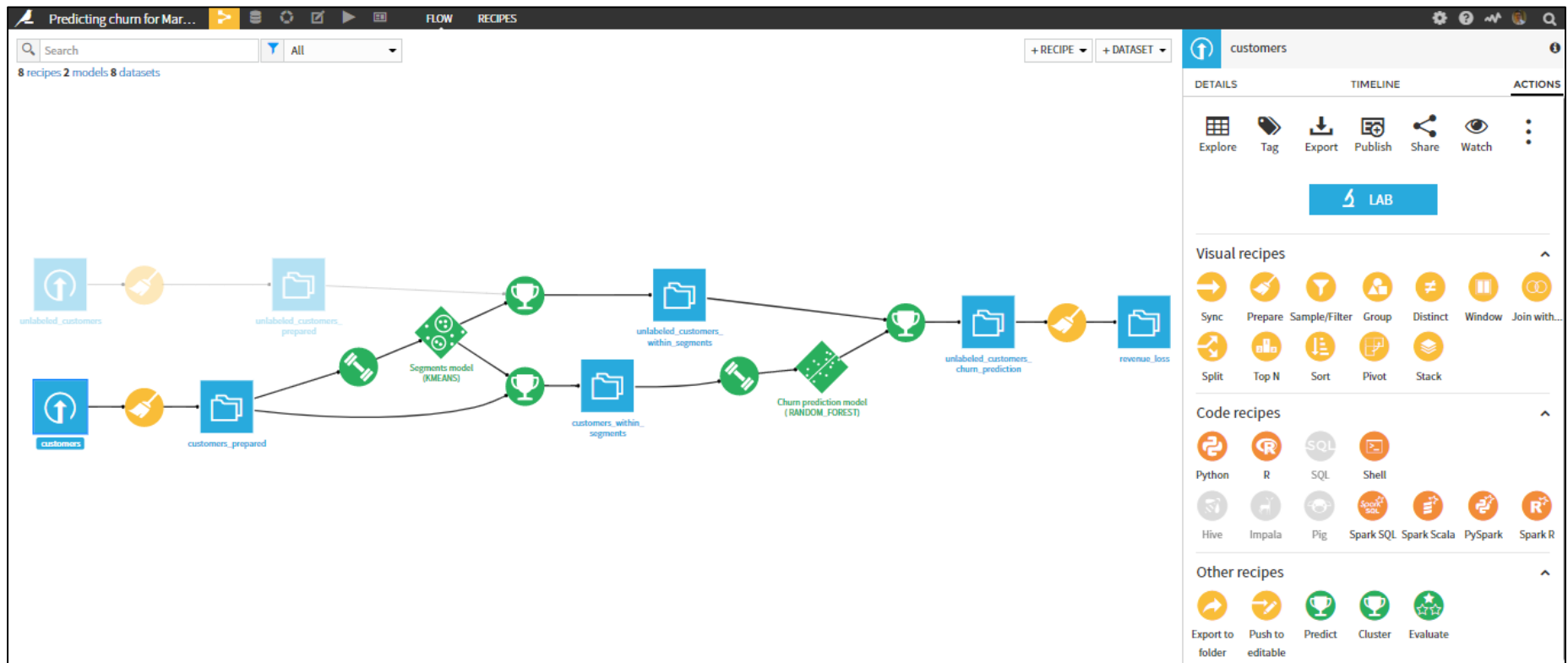
[Product](#) [Solutions](#) [Learn](#) [Resources](#) [Company](#) [Blog](#) [Contact us](#) [GET STARTED](#)

Collaborative Data Science Platform

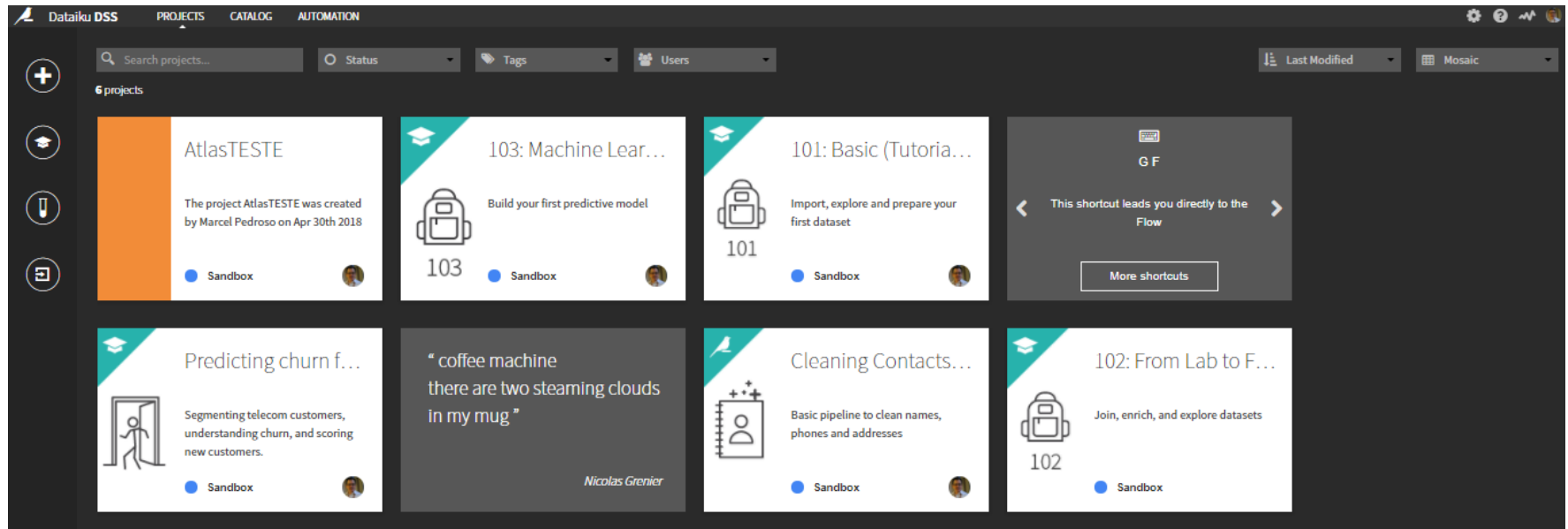
Prototype, Deploy, & Run at Scale

[GET STARTED](#) [▶ WATCH VIDEO](#)

VISUAL MACHINE LEARNING



PROJETO PREDIÇÃO TMI



The screenshot displays the Dataiku DSS Projects Catalog interface. The top navigation bar includes 'Dataiku DSS', 'PROJECTS', 'CATALOG', and 'AUTOMATION'. Below the navigation bar, there is a search bar labeled 'Search projects...' and filters for 'Status', 'Tags', and 'Users'. The main content area shows a grid of project cards, each with a title, description, and a 'Sandbox' button. The cards are arranged in two rows. The first row contains 'AtlasTESTE', '103: Machine Lear...', '101: Basic (Tutoria...', and a 'GF' card. The second row contains 'Predicting churn f...', '“ coffee machine there are two steaming clouds in my mug ”', 'Cleaning Contacts...', and '102: From Lab to F...'. Each card also features a small icon and a user profile picture.

Dataiku DSS PROJECTS CATALOG AUTOMATION

Search projects... Status Tags Users

Last Modified Mosaic

6 projects

AtlasTESTE

The project AtlasTESTE was created by Marcel Pedroso on Apr 30th 2018

Sandbox

103: Machine Lear...

Build your first predictive model

103 Sandbox

101: Basic (Tutoria...

Import, explore and prepare your first dataset

101 Sandbox

GF

This shortcut leads you directly to the Flow

More shortcuts

Predicting churn f...

Segmenting telecom customers, understanding churn, and scoring new customers.

Sandbox

“ coffee machine there are two steaming clouds in my mug ”

Nicolas Grenier

Cleaning Contacts...

Basic pipeline to clean names, phones and addresses

Sandbox

102: From Lab to F...

Join, enrich, and explore datasets

102 Sandbox

Vídeo: [A Friendly Introduction to Machine Learning](#) (UDACITY)

What is Machine Learning?

Learn from experience



data
Learn from ~~experience~~



Follow instructions





Institute for Scientific and Technological Communication and Information on Health

www.facebook.com/fiocruz.icict

[twitter.com/@Icict_fiocruz](https://twitter.com/Icict_fiocruz)

www.youtube.com/videosaudefio

www.icict.fiocruz.br