KNIME
Machine learning para não programadores

GLEBER TACIO TEIXEIRA, DSC.

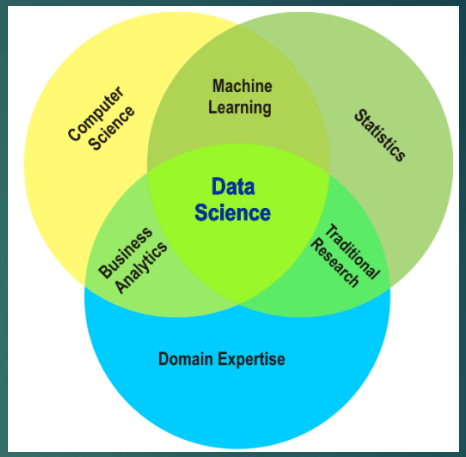
PETROBRAS

Agenda

- Introdução
- Algoritmos de Machine Learning
  - Linear regression
  - Decision tree
  - Random forest
  - MLP
- Ferramenta Knime
- Conclusões

Introdução

- O que é Machine Learning?



http://www.marsiantech.com/best-data-science-training-in-pune.php

Machine Learning = Computação + Estatística + Álgebra

Modelos

Introdução

- É possível trabalhar com *Machine Learning* sem ser matemático, estatístico ou da área de computação?

Geoffrey E. Hinton
http://www.cs.toronto.edu/~hinton/geoff6.jpg

Introdução

- É possível trabalhar com *Machine Learning* sem gostar de matemática?

Introdução

- É possível trabalhar com *Machine Learning* sem saber programar?

Algoritmos de *Machine Learning*

Método Supervisionado - regressão

Atributos

Y
Label

Casa	Numero de quartos	Suites	m2	Bairro	Valor R$
1	2	2	150	Alfa	**237.000,00**
2	3	1	189	Beta	**329.450,00**
3	4	2	220	Gama	**423.250,00**

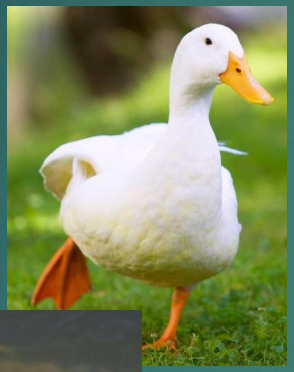
Criar Modelo

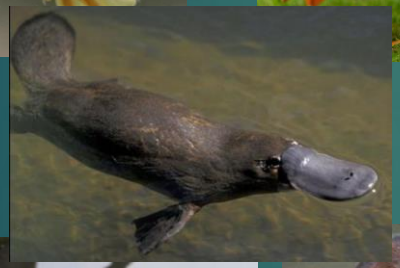
Amostra	**3**	**2**	**170**	**Alfa**	**375.670,00**

Algoritmos de *Machine Learning*

Supervisionado - classificação

Pato Pato

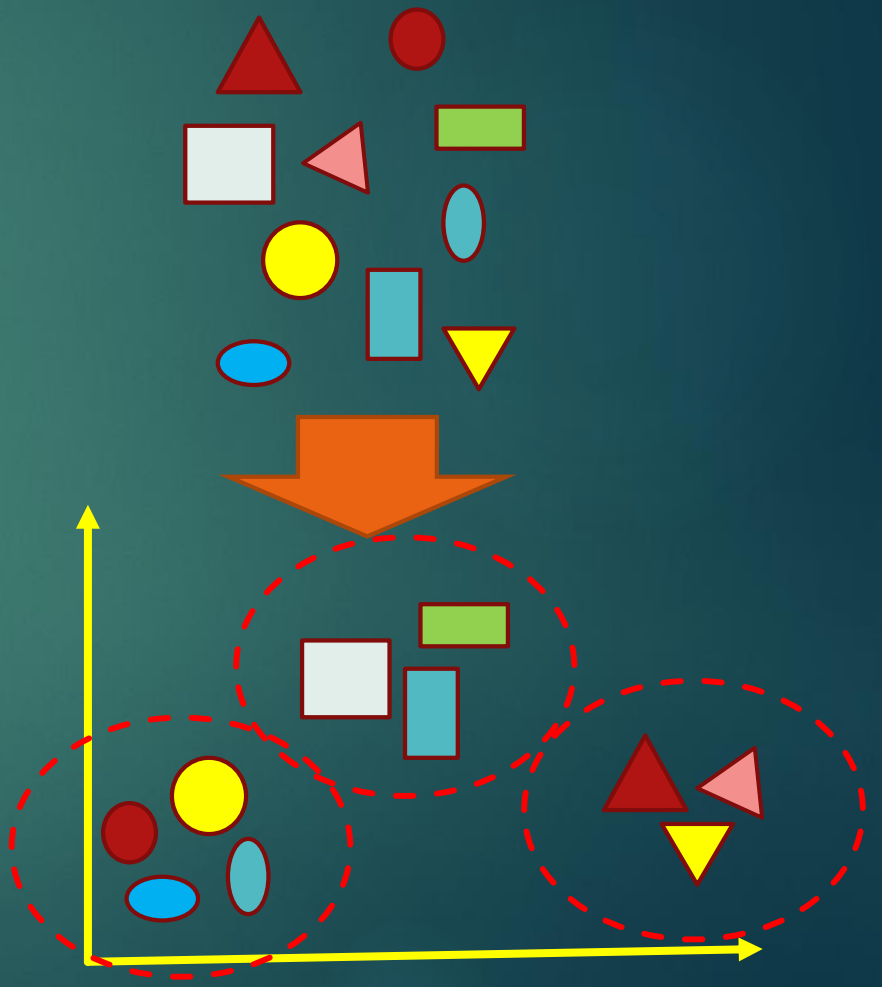




???

Castor Castor

Não Supervisionado



Algoritmos de *Machine Learning*

Variáveis contínuas

Não Supervisionado
Clusterização e redução de dimensionalidade
K-means
PCA

Supervisionado (Regressão)
Regressão Linear
Decision Tree
GB Tree
Random Forests
MLP

Variáveis categóricas

Não Supervisionado
Análise de associação
Markov Models

Supervisionado (Classificação)
Decision Tree
Random Forests
MLP
Regressão Logística
Naive-Bayes
SVM
KNN

Terminologia na modelagem preditiva

- Modelo: função explicativa.
- Dados: o que você está usando pra encaixar no modelo.
- Alvo/Target/Label/Variável dependente/y: valor que você está tentando prever.
- Características/Features/Variáveis independentes/X atributos que serão utilizados na predição.
- Métodos: algoritmos que usarão seus dados para obter um modelo.

Regressão linear

Intercepto

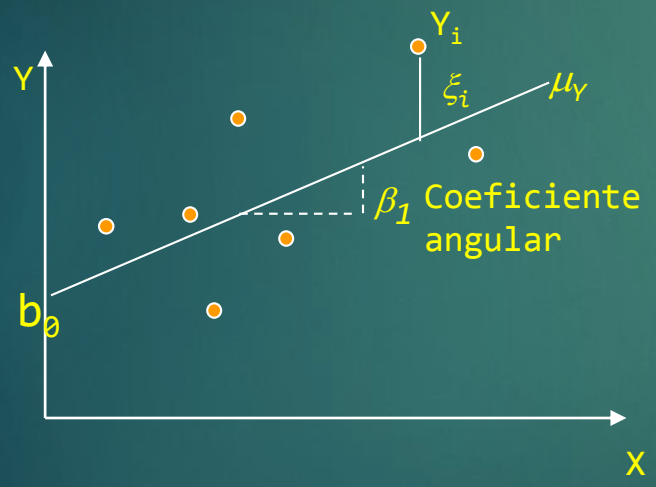
Inclinação

Variável Independente

Variável Dependente

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Erro Aleatório



$\hat{Y}_i = b_0 + b_1 X_i$ Modelo estimado

$\varepsilon_i = Y_i - \hat{Y}_i$ Resíduo

Vantagens

- Fácil de avaliar a qualidade do modelo
- Bem fundamentado
- Intuitivo

Desvantagens

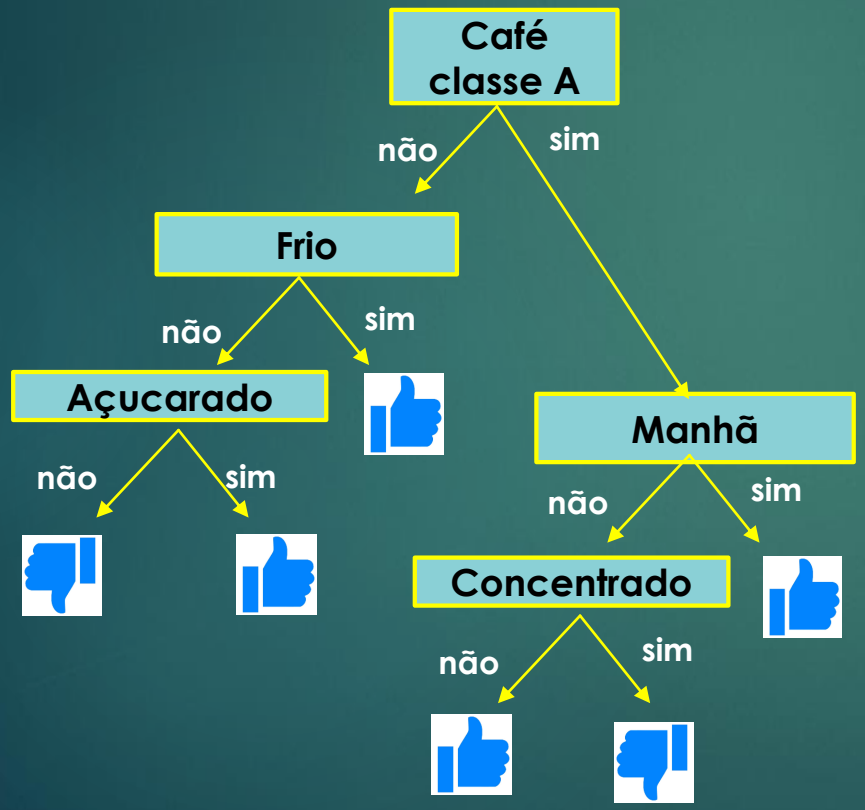
- Cuidado ao se trabalhar com variáveis categóricas (qualitativa)
- Deve-se cumprir os pressupostos (Linearidade, homocedasticidade....)

Classe A	Dia Frio	Manhã	Concentrado	Açucarado	Resultado
Sim	Sim	Não	Não	Sim	
Não	Sim	Sim	Sim	Não	
Não	Não	Sim	Sim	Sim	
Sim	Nao	Não	sim	Sim	

• • •

Decision Tree

- "Aprende" criando ramos ou subdivisões do dataset de maneira recursiva com ganho de informação (supervisionado)



Vantagens
- Simples de usar e interpretar (gráfico da árvore)
- Robustas
- Variáveis categorias ou numéricas
- Pouca preparação
- Trabalha bem com muitos dados

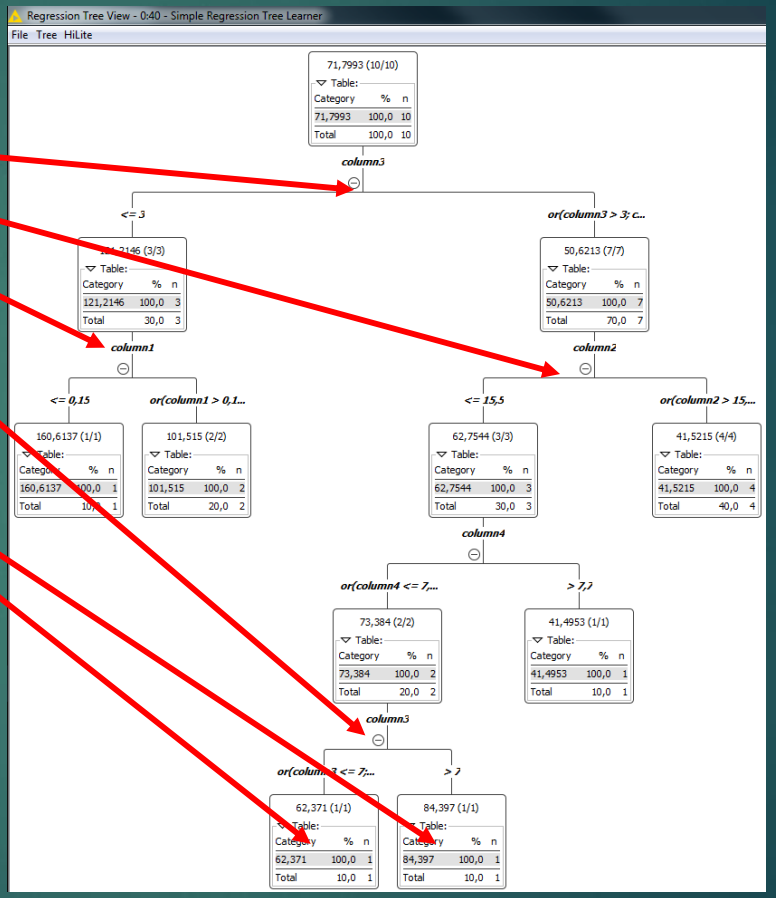
Desvantagens
- Podem não ser muito precisas
- Podem sofrer overfitting.
- Variáveis categóricas com várias divisões podem induzir desvio.

Árvore de decisão

Número de nós

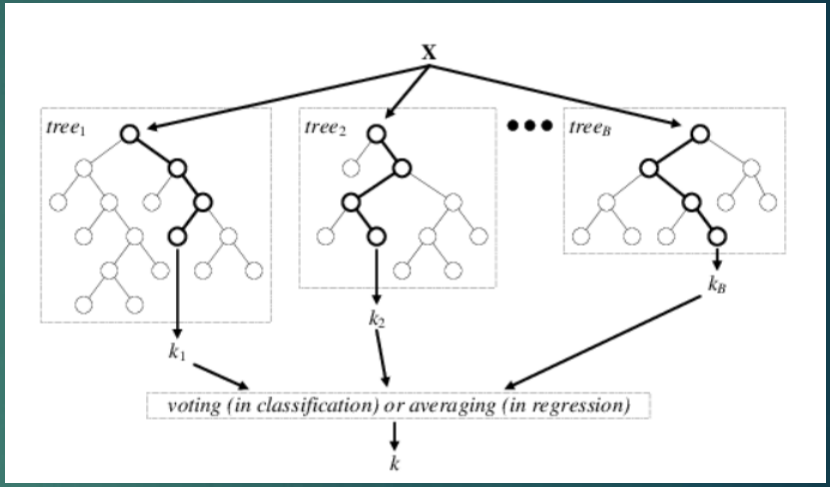
Número de divisões de nós (filhos)

Prof. da árvore, 4 camadas



Random forest

- Bom pra regressão e classificação
- Visualização do modelo
- Menor risco de overfitting
- Exige pouco preparo das amostras
- Exige pouco ajuste de hyperparamentros
- Rápido pra treinar mas lento pra usar (real time)
- Bagging de árvores (decision tree)
  - Combinação de vários modelos pode gerar um modelo melhor
  - Votação

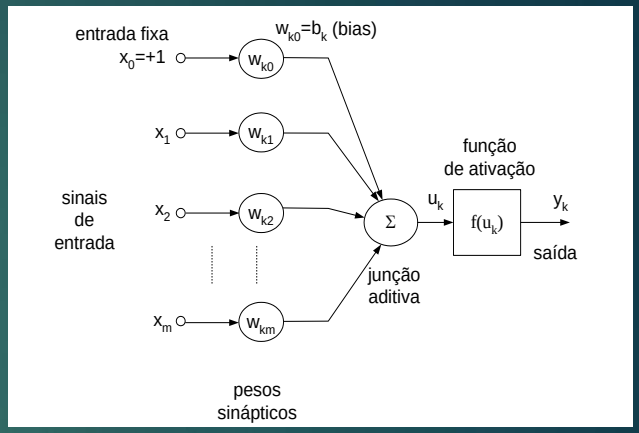


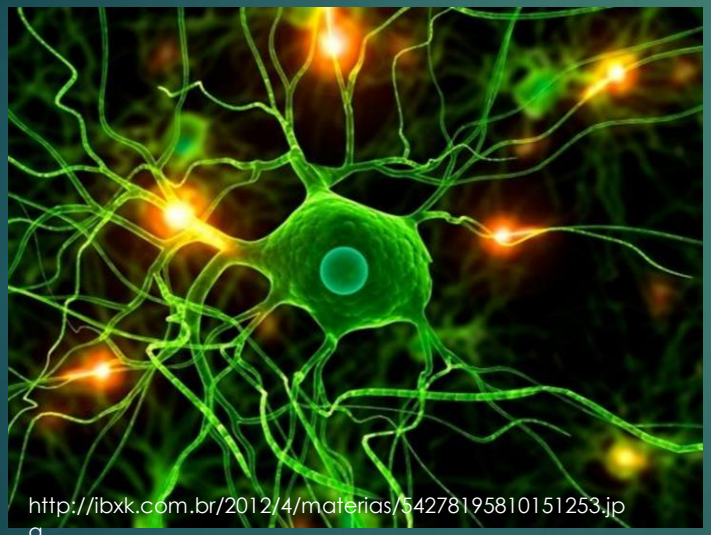
https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643

Rede Neural (Rprop - MLP)

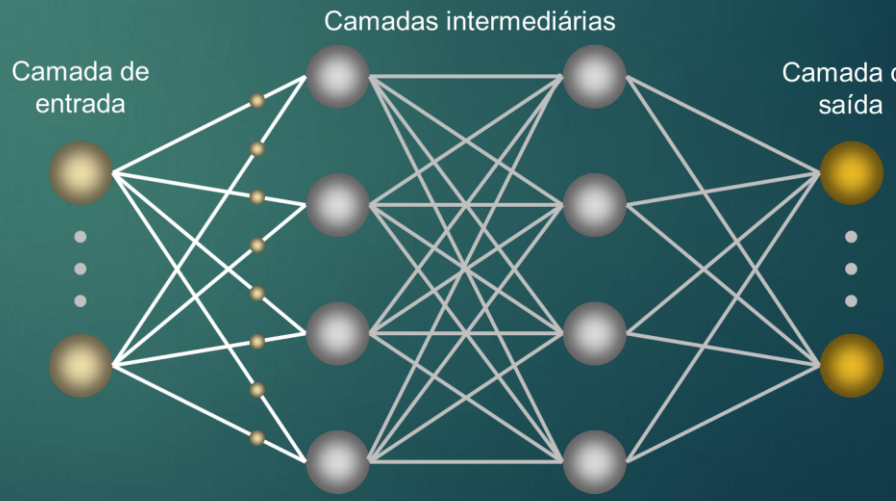
Técnica inspirada no funcionamento do cérebro, onde neurônios artificiais, conectados em rede, são capazes de aprender e de generalizar.

Técnica de aproximação de funções por regressão não linear.





http://ibxk.com.br/2012/4/materias/54278195810151253.jpg



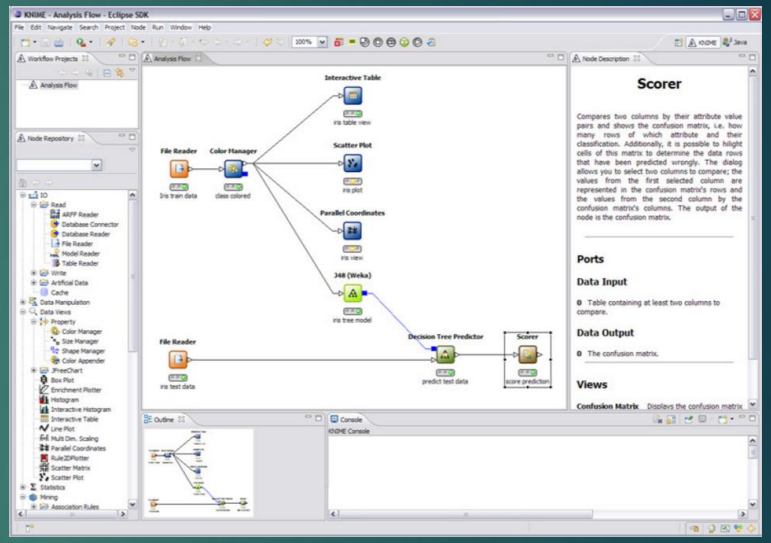
www.dcc.ufla.br/~lacerda/download/palestras/rna/palestra_01.ppt

Open for Innovation ®
KNIME

KNIME Analytics Platform

KNIME® Analytics Platform is the leading open solution for data-driven innovation, helping you discover the potential hidden in your data, mine for fresh insights, or predict new futures. Our enterprise-grade, open source platform is fast to deploy, easy to scale and intuitive to learn.

With more than 1000 modules, hundreds of ready-to-run examples, a comprehensive range of integrated tools, and the widest choice of advanced algorithms available, KNIME Analytics Platform is the perfect toolbox for any data scientist. Our steady course on unrestricted open source is your passport to a global community of data scientists, their expertise, and their active contributions.



www.knime.com

- Free
- Intuitiva
- Fácil
- Ágil
- Open Source
- Constante Atualização
- Ampla aplicação

Knime - Konstanz Information Miner

- Universidade de Konstanz – início em 2004 (Michael Berthold)
  - Proposta de plataforma de integração
- Primeira versão 2006
  - Companhias farmacêuticas e de softwares científicos
- Hoje - Plataforma de integração de vários projetos / empresas

- Desenvolvimentos atuais:
  - Big Data
  - Spark

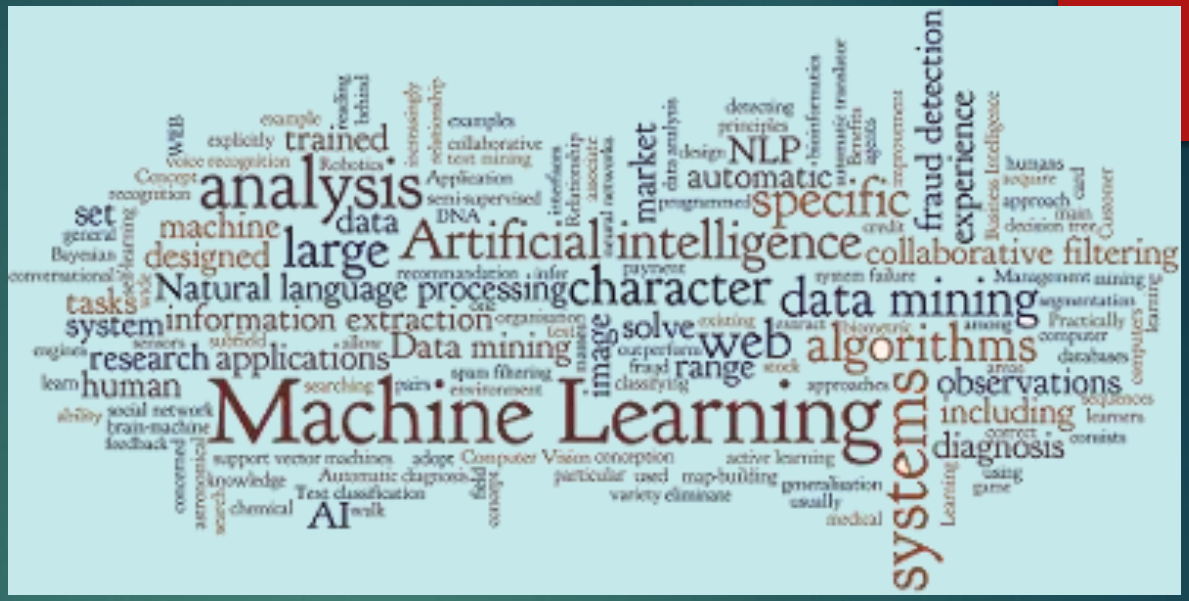
https://en.wikipedia.org/wiki/KNIME

Knime - Konstanz Information Miner

- Plataforma analítica
  - Machine Learning
  - Data mining
  - ETL
- Plataforma de integração e de reporting.
- Mínima programação
  - Aplicação final
  - Prototipagem
  - Ensino

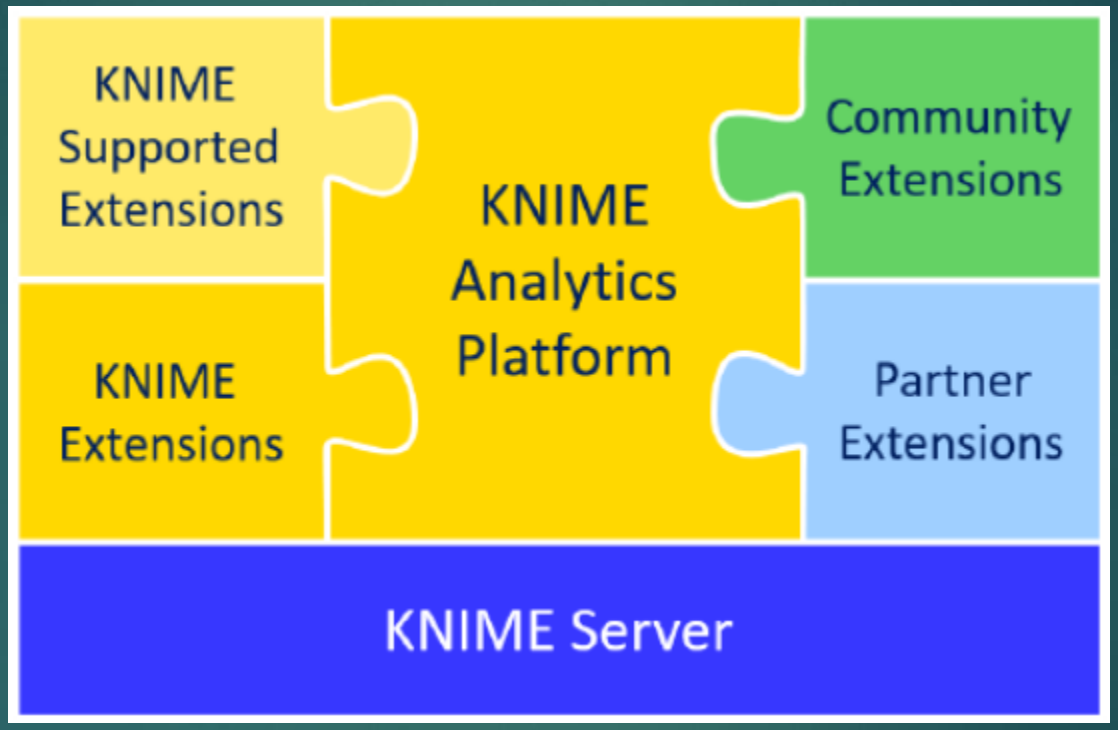
Fig. 1: Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms

- Data Science
- Machine Learning
- Artificial Inteligence
- Preditive Modeling
- Deep Learning
- Data Mining
- Multivariate Analysis
- Pattern recognition
- Chemometrics
- BI
- Econometrics
- Biostatistics
- Proteomics / Genomics



http://s.hswstatic.com/gif/ai-arms-race-part-1.jpg

Knime - Konstanz Information Miner

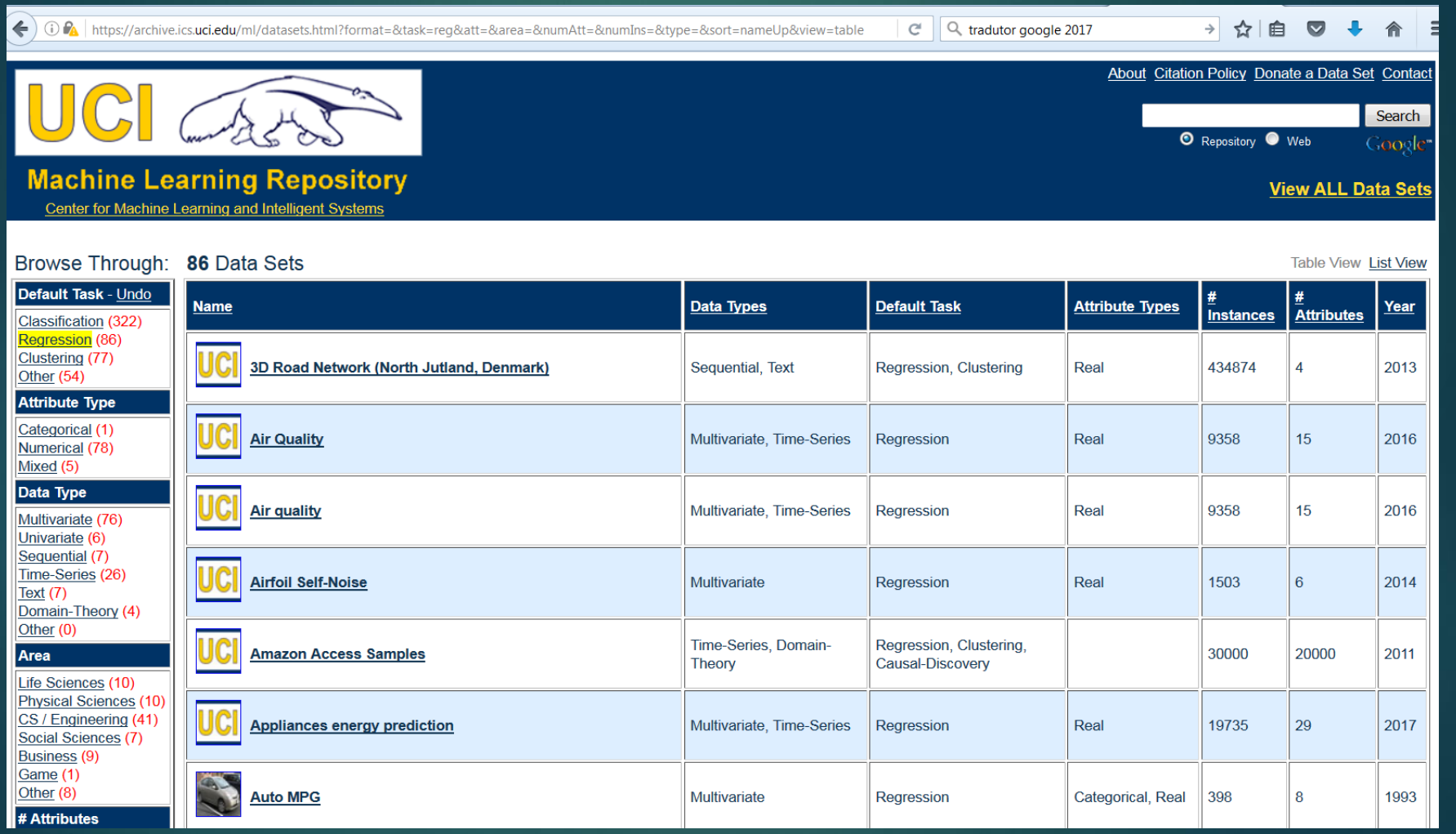


www.knime.com

Knime - Konstanz Information Miner

Perl

Ao trabalho ...
Regressão



Regressão – Resistência compressiva do concreto

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy

Reposit

Concrete Compressive Strength Data Set

Download: Data Folder, Data Set Description

Abstract: Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1030	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	9	**Date Donated**	2007-08-03
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	130670

Source:

Original Owner and Donor
Prof. I-Cheng Yeh
Department of Information Management
Chung-Hua University,
Hsin Chu, Taiwan 30067, R.O.C.
e-mail:icyeh '@' chu.edu.tw
TEL:886-3-5186511

Date Donated: August 3, 2007

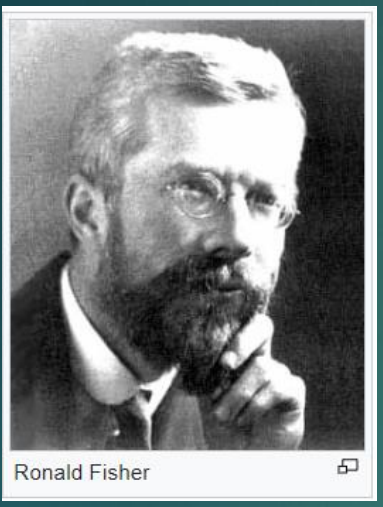
Attribute Information:

Given are the variable name, variable type, the measurement unit and a brief description. The concrete co
the rows of the database.

Name -- Data Type -- Measurement -- Description

Cement (component 1) -- quantitative -- kg in a m3 mixture -- Input Variable
Blast Furnace Slag (component 2) -- quantitative -- kg in a m3 mixture -- Input Variable
Fly Ash (component 3) -- quantitative -- kg in a m3 mixture -- Input Variable
Water (component 4) -- quantitative -- kg in a m3 mixture -- Input Variable
Superplasticizer (component 5) -- quantitative -- kg in a m3 mixture -- Input Variable
Coarse Aggregate (component 6) -- quantitative -- kg in a m3 mixture -- Input Variable
Fine Aggregate (component 7) -- quantitative -- kg in a m3 mixture -- Input Variable
Age -- quantitative -- Day (1~365) -- Input Variable
Concrete compressive strength -- quantitative -- MPa -- Output Variable

Classificação – Iris flower dataset



Ronald Fisher

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1954925

Source:

Creator:

R.A. Fisher

Donor:

Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)

Classificação – Qualidade do vinho

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

tradutor google 2017

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Con

Repository Web

View ALL Data S

Wine Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	676156

Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

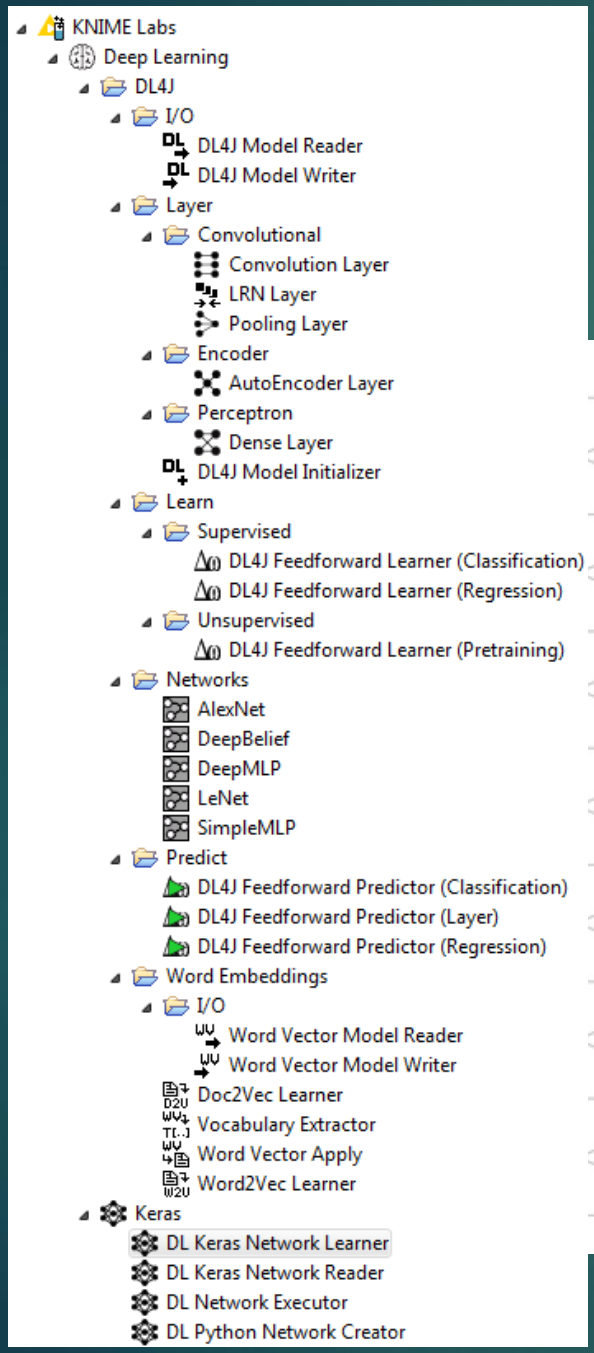
Source:

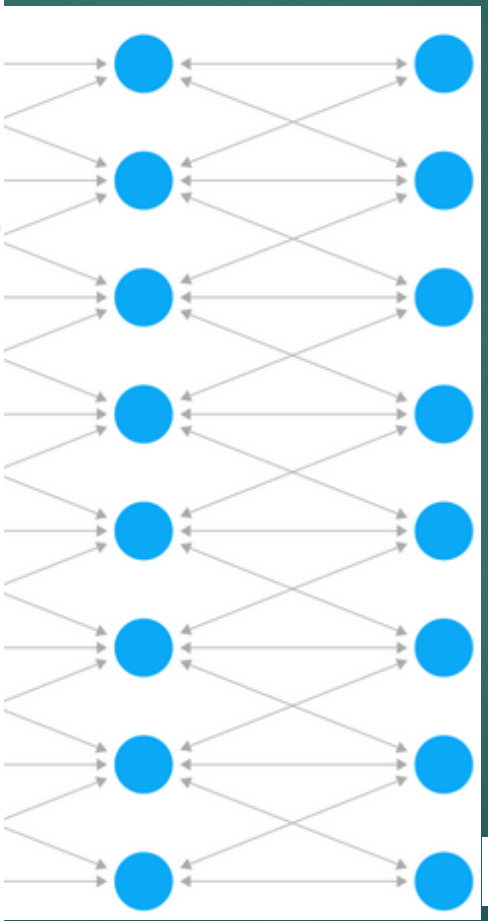
Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

Data Set Information:

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

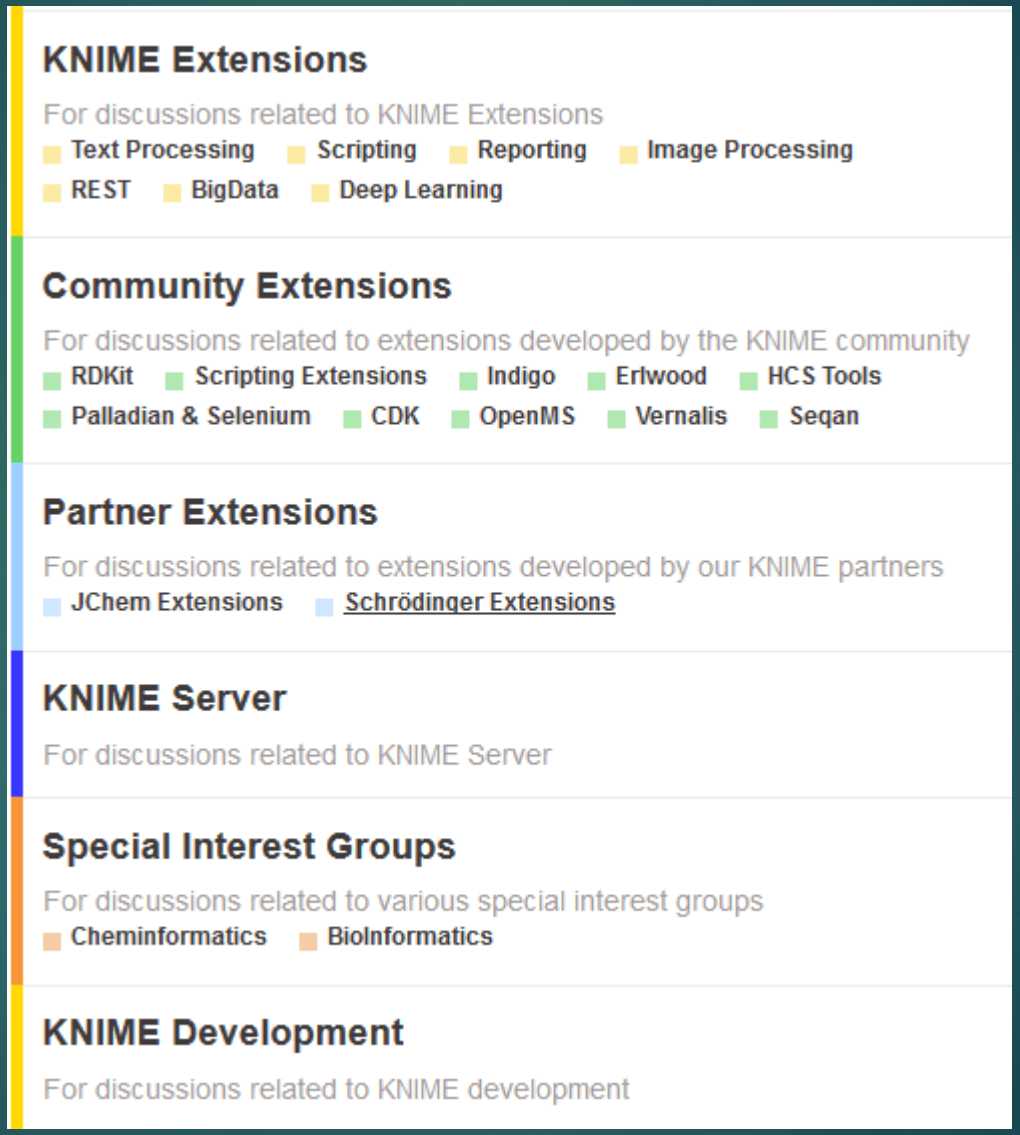




Keras

TensorFlow™

DEEPLEARNING4J



www.knime.com

E agora... Pra onde vou!!

- Aprofundar nos conceitos matemáticos dos algoritmos.
- Aprofundar nas outras aplicações e nos outros nós disponíveis no KNIME.
- E...

Se for do seu interesse aprenda também a programar:

Open for Innovation KNIME + SQL + BI

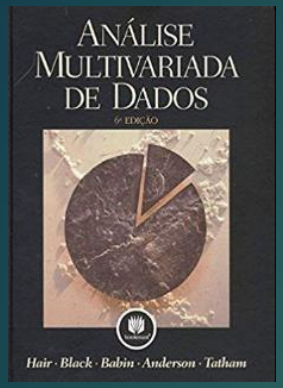
+

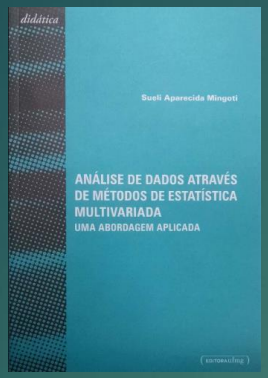
python ou R ou Scala

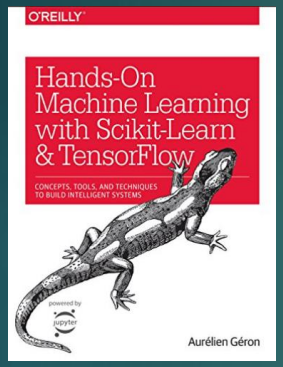
Scala e python e R

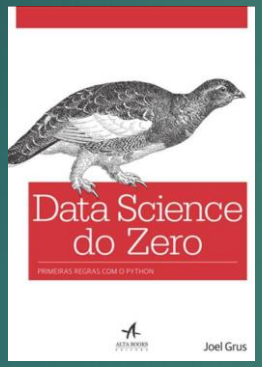
Scala python R Spark hadoop

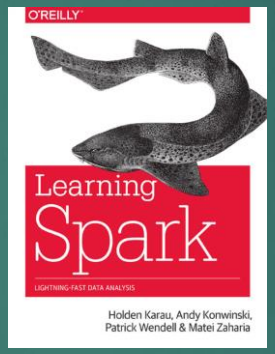
JS TensorFlow ANACONDA jupyter Regex

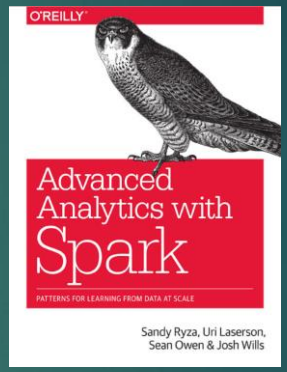


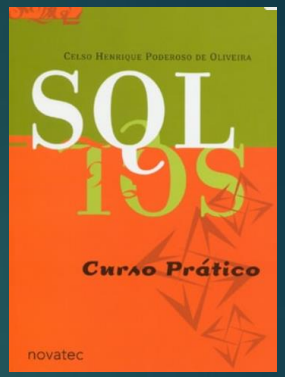












http://professores.dcc.ufla.br/~lacerda/
Fonte: Prof. Wilian Soares Lacerda
Depto. Ciência da Computação

- Deep Learning in Python: Master Data Science and Machine Learning with Modern Neural Networks written in Python, Theano, and TensorFlow (Machine Learning in Python) (English Edition) eBook Kindle
- KACHIGAN, Sam Kash. Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods. New York: Radius Press, 1986, 589 p.

“Ninguém sabe tanto que não possa aprender e nem tão pouco que não possa ensinar.”

Anônimo

Gleber Teixeira, PhD
Technical Consultant at Petrobras