
Data
Science

O que é e o que faz, um cientista de dados?

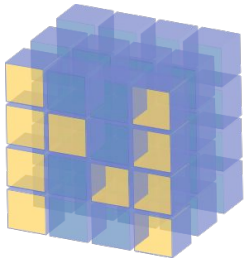
Um cientista de dados é alguém que sabe como extrair significado e interpretar dados, o que exige ferramentas e métodos de estatística e aprendizado de máquina, Ele passa muito tempo no processo de coleta, limpeza e seleção de dados,

Carreiras:

- Engenheiro de dados
- Analista de dados
- Engenheiro de ML
- Estatístico
- Arquiteto de Data Warehouse
- Analista de negócios

Codificando Data Science!

As bibliotecas Python mais utilizadas para Data Science



NumPy

+

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

- **NumPy** é uma biblioteca fundamental para computação científica com Python

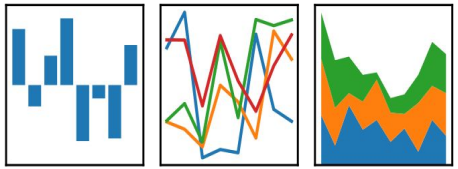
- Utiliza umas estrutura de arrays de **N-dimensões**

NumPy

+

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- **pandas** é uma biblioteca open-source que utiliza estrutura de dados e ferramentas de análise de dados para a linguagem Python

- pandas é fácil de usar

Hora de codar !

Seaborn

&

- O Matplotlib é uma biblioteca de plotagem 2D do Python e de fácil utilização em ambientes interativos.

- Seaborn é uma biblioteca de visualização de dados Python baseada no matplotlib . Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos.

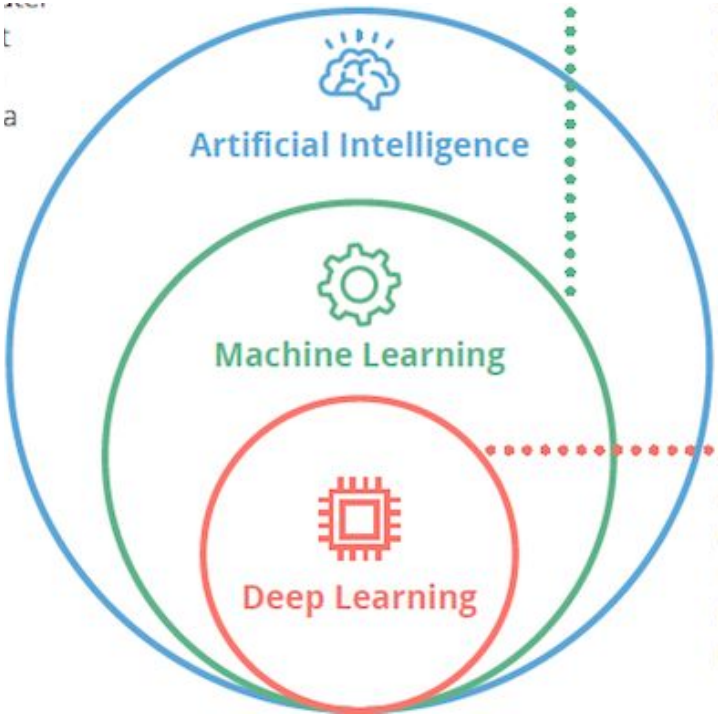
Inteligência Artificial

Conceitos, paradigmas e… irá nos matar e dominar o mundo?

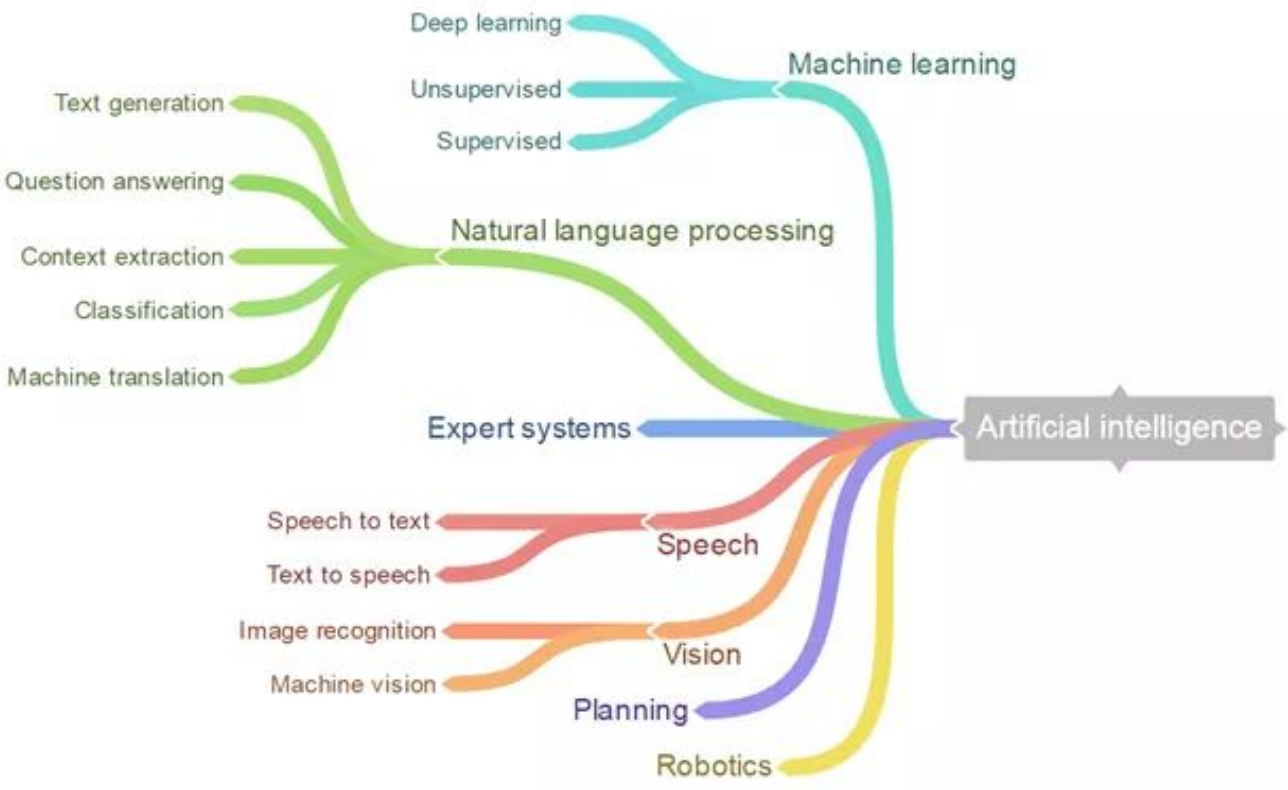
O que é?
Onde vive?
Para que serve?

Inteligência Artificial

“Inteligência Artificial é a simulação dos processos da inteligência humana realizados por uma máquina, mais especificamente por sistemas de computadores”



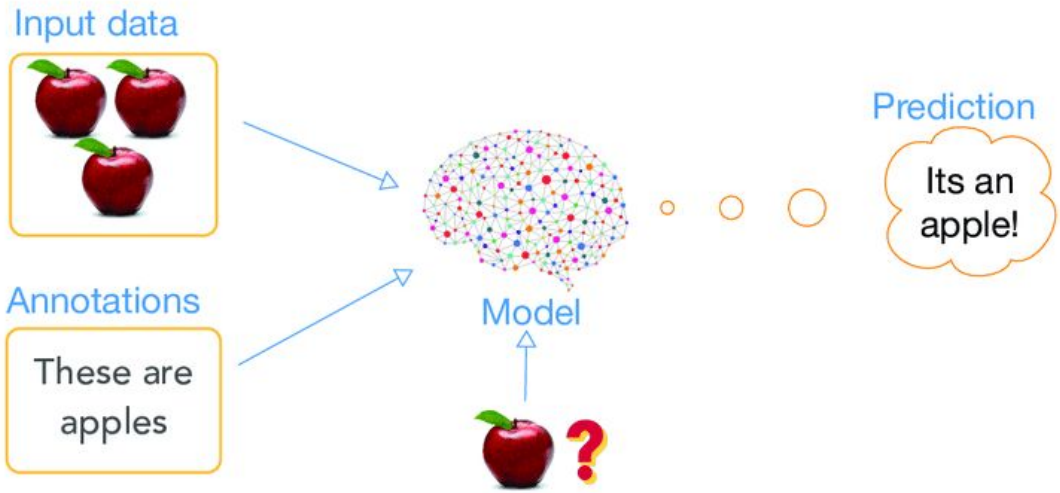
Ramos da IA


Deep learning
Unsupervised
Supervised
Machine learning
Text generation
Question answering
Context extraction
Classification
Machine translation
Natural language processing
Expert systems
Artificial intelligence
Speech to text
Text to speech
Speech
Image recognition
Machine vision
Vision
Planning
Robotics

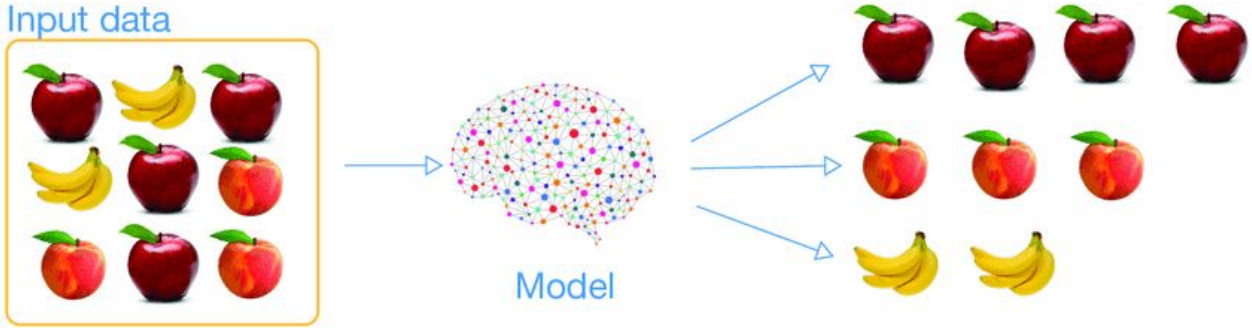
Problemas do Aprendizado de Máquina (ML)

Supervisionado. Não Supervisionado. Por Reforço.

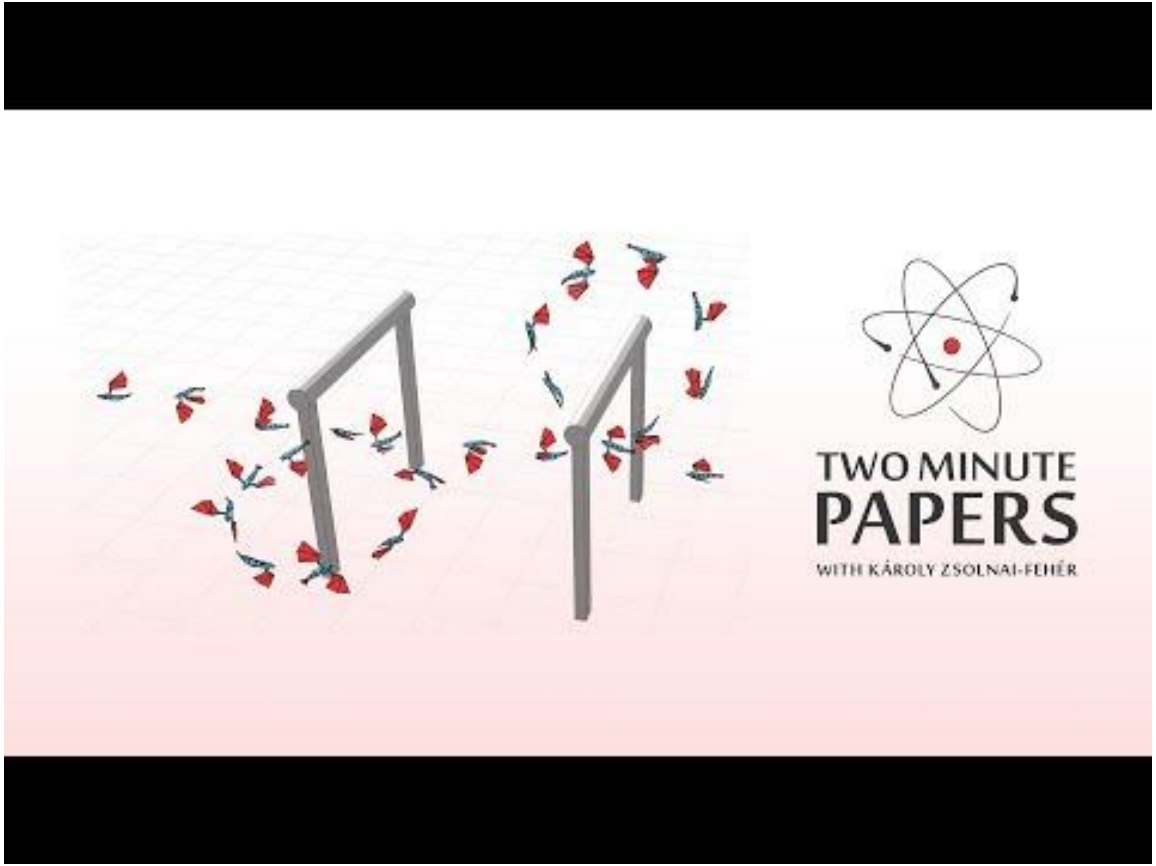
Aprendizado Supervisionado

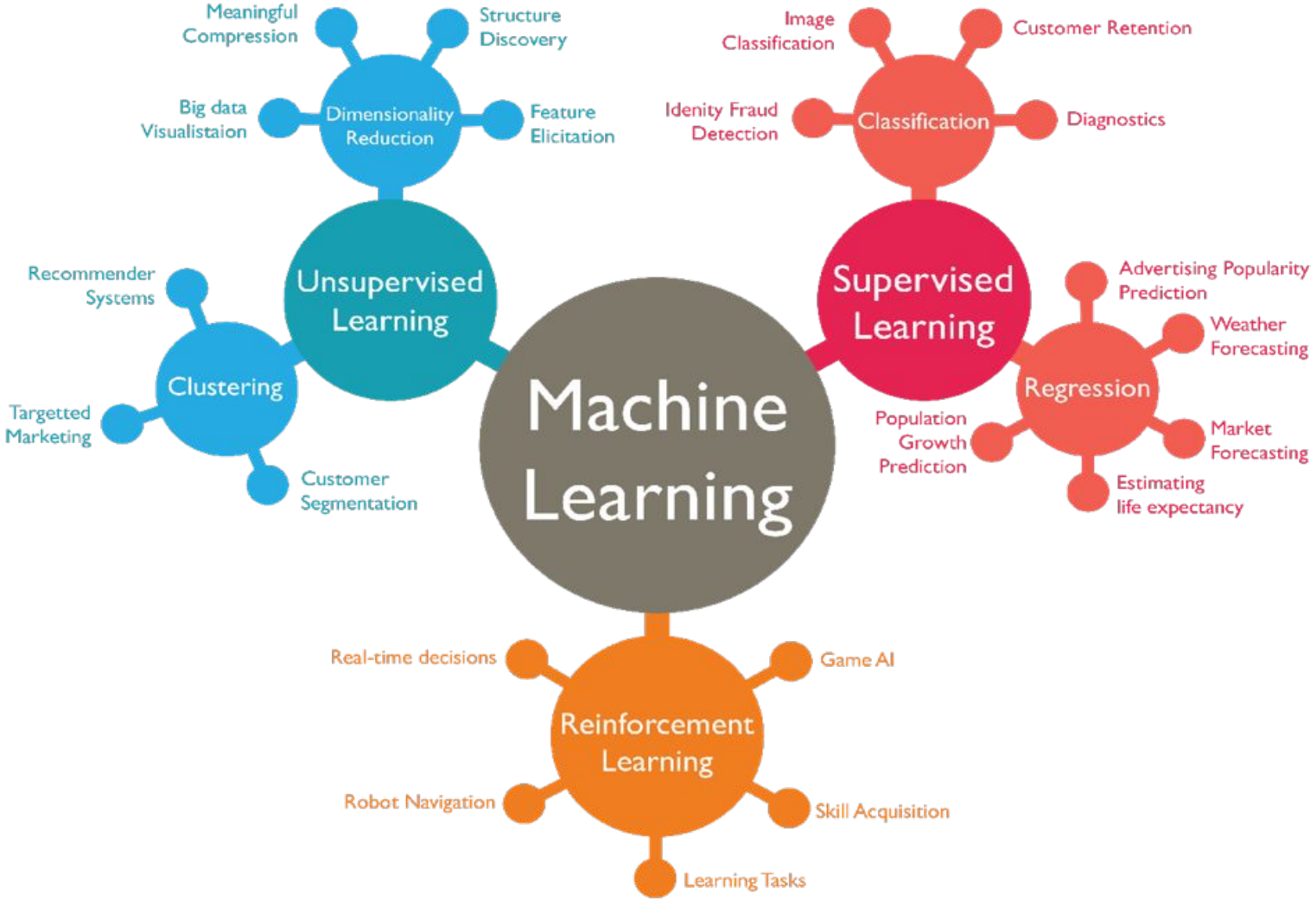


Aprendizado Não Supervisionado



Aprendizado Por Reforço




Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

KNN - K Nearest Neighbors

K Vizinhos mais Próximos

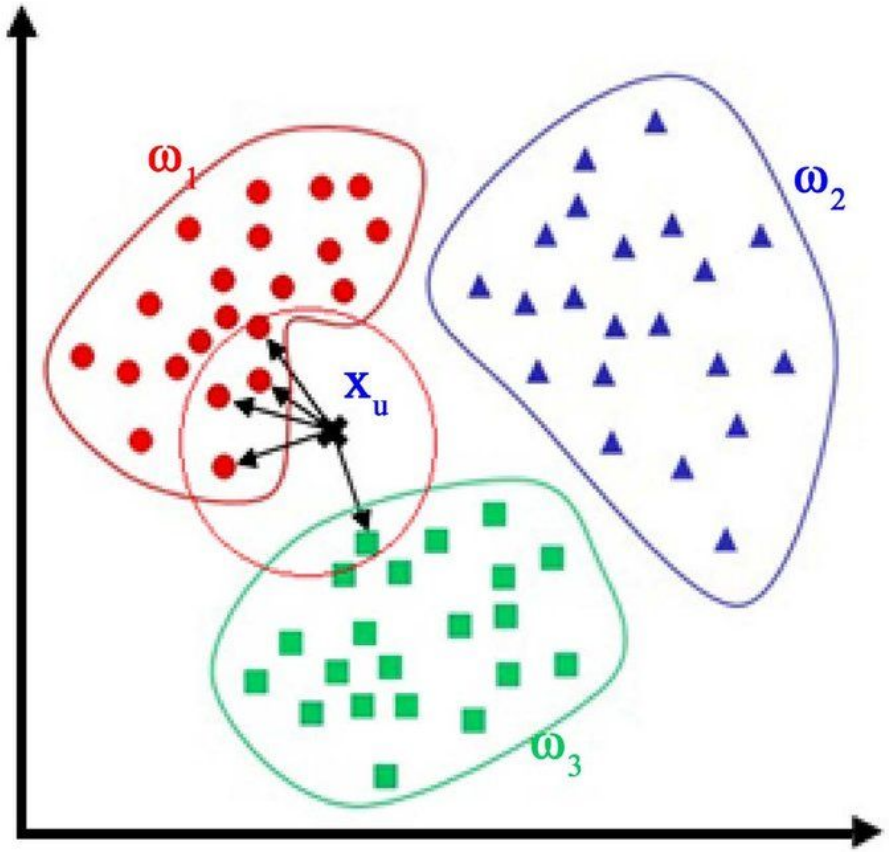
KNN

- Algoritmo de aprendizagem supervisionada
- Utilizado em problemas de classificação ou regressão
- Lazy Learner (aprendiz preguiçoso)

Intuição

Definição

É um algoritmo simples que armazena todos os casos disponíveis e classifica os novos dados ou casos com base em uma medida de similaridade. É usado principalmente para classificar um ponto de dados com base em como seus vizinhos são classificados.

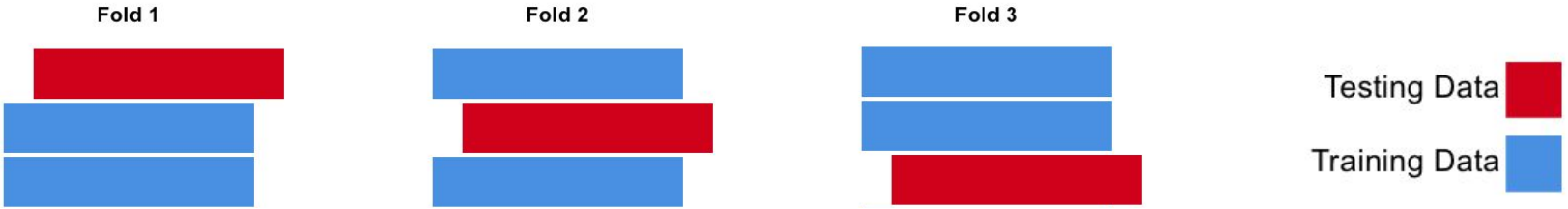


Etapa de Validação


Test
Train

K-Fold (Cross Validation)

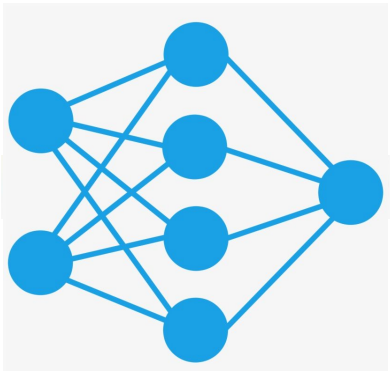
O dataset é dividido em K partes de tamanhos iguais. E então o treinamento é feito K vezes. Separando em cada um, um segmento diferente para realizar o teste, e os demais para o treinamento.



Variações:

- Leave one out
- Stratified

Métricas de avaliação



É cachorro!

Matriz Confusão

Métricas

Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;

$$\frac{VP + VN}{Total}$$

Precisão: dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;

$$\frac{VP}{VP + FP}$$

Recall/**Sensibilidade**: dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas;

$$\frac{VP}{VP + FN}$$

F1-Score: média harmônica entre precisão e sensibilidade.

$$2 * \frac{Precisão * Recall}{Precisão + Recall}$$

Vamos codar!

Tratamento de Dados

Dados Categóricos
Dados Faltantes

Dados Faltando...

E AGORA?

Abordagens para dados faltando...

- Deletar os dados que estão com variáveis faltantes;
- Atribuir a média/mediana aos valores que estão faltando (se for numérico);
- Atribuir os dados faltando como mais uma categoria (se for categórico);
- Rodar modelos preditivos para preencher os dados faltando
    - KNN e Decision Tree são os mais utilizados para isso

Show me the code!

Dados Categóricos, como lidar?

Estado Civil
Solteiro(a)
Casado(a)
Divorciado(a)

Dados Categóricos

Estado Civil
Solteiro(a)
Casado(a)
Divorciado(a)

Estado Civil
1
2
3

- **Label Encoder / Dummy Encoder**
- **One Hot Encoder**
- **Binary Encoder**

One Hot Encoder

Estado Civil
Solteiro(a)
Casado(a)
Divorciado(a)

est_civ_solte	est_civ_cas	est_civ_divor
1	0	0
0	1	0
0	0	1

Binary Encoder

Estado Civil
Solteiro(a)
Casado(a)
Divorciado(a)

Estado Civil
1
2
3

est_civ_2	est_civ_2
0	1
1	0
1	1

Code Time!

Classes desbalanceadas

- Os dados estão mal distribuídos entre as classes.
  - Há uma classe com muito mais dados do que a outra

- Problema:
  - O algoritmo vai não aprenderá o suficiente da classe com menos amostras
  - Tenderá a classificar todos os dados novos como da classe majoritária

- Mudar o algoritmo

- Reamostragem:
  - Oversample
  - Undersample
  - Amostras sintéticas

Dicas para o mundo dos dados!

Podcasts

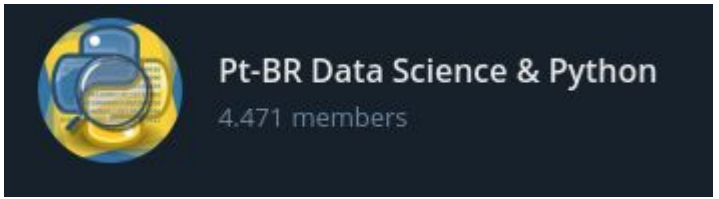
- Hipsters Ponto Tech
- Pizza de Dados
- Data Hackers
- Deep Mind: The Podcast

Plataformas

- Kaggle
- Data Camp
- Udacity

Comunidades

- *Slack* - Data Hackers
- *Telegram* - Data Science & Python
- *Telegram* - Pizza de Dados
- *Medium* - Towards Data Science



Contatos


fmdss@ic.ufal.br


wsf@ic.ufal.br


francamacdowell


linkedin.com/in/wagner-fontes


francamacdowell


WagnerFLL


francamacdowell

PERGUNTAS?

OBRIGADO !!!