99% de acurácia... tá, mas e aí?

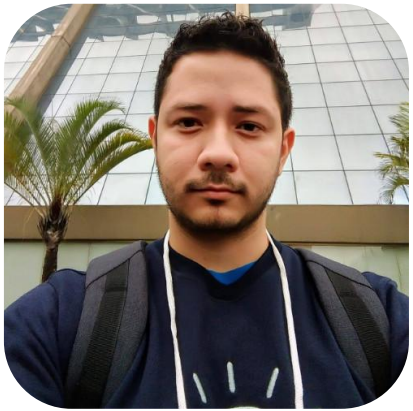
By Rodrigo C. Moraes

https://github.com/rodrigocmoraes

rdcmdev@gmail.com

kaggle https://www.kaggle.com/rdcmdev

Quem sou eu?



UEA
UNIVERSIDADE DO ESTADO DO AMAZONAS

itriad.
Instituto Triad de Pesquisa e Desenvolvimento

Engenheiro de Machine Learning

Graduando em Engenharia de Computação

Ex Maratonista de Programação

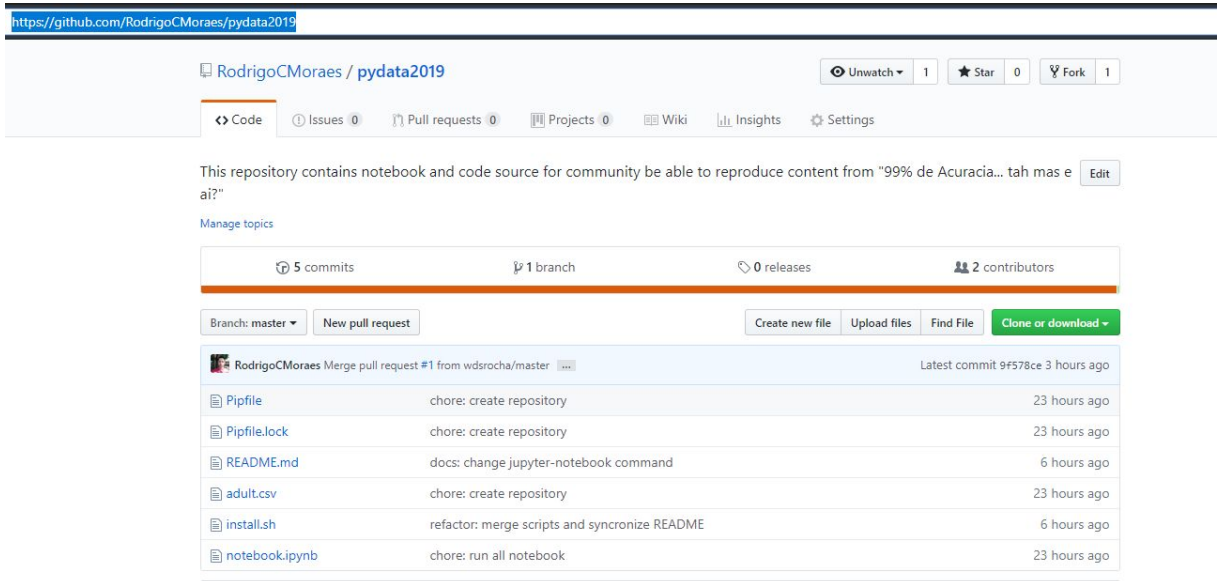
...

Como validar um modelo de Machine Learning?

Programação do Minicurso

1. Apresentação
2. Código/Hands-on

Acesso ao material do minicurso

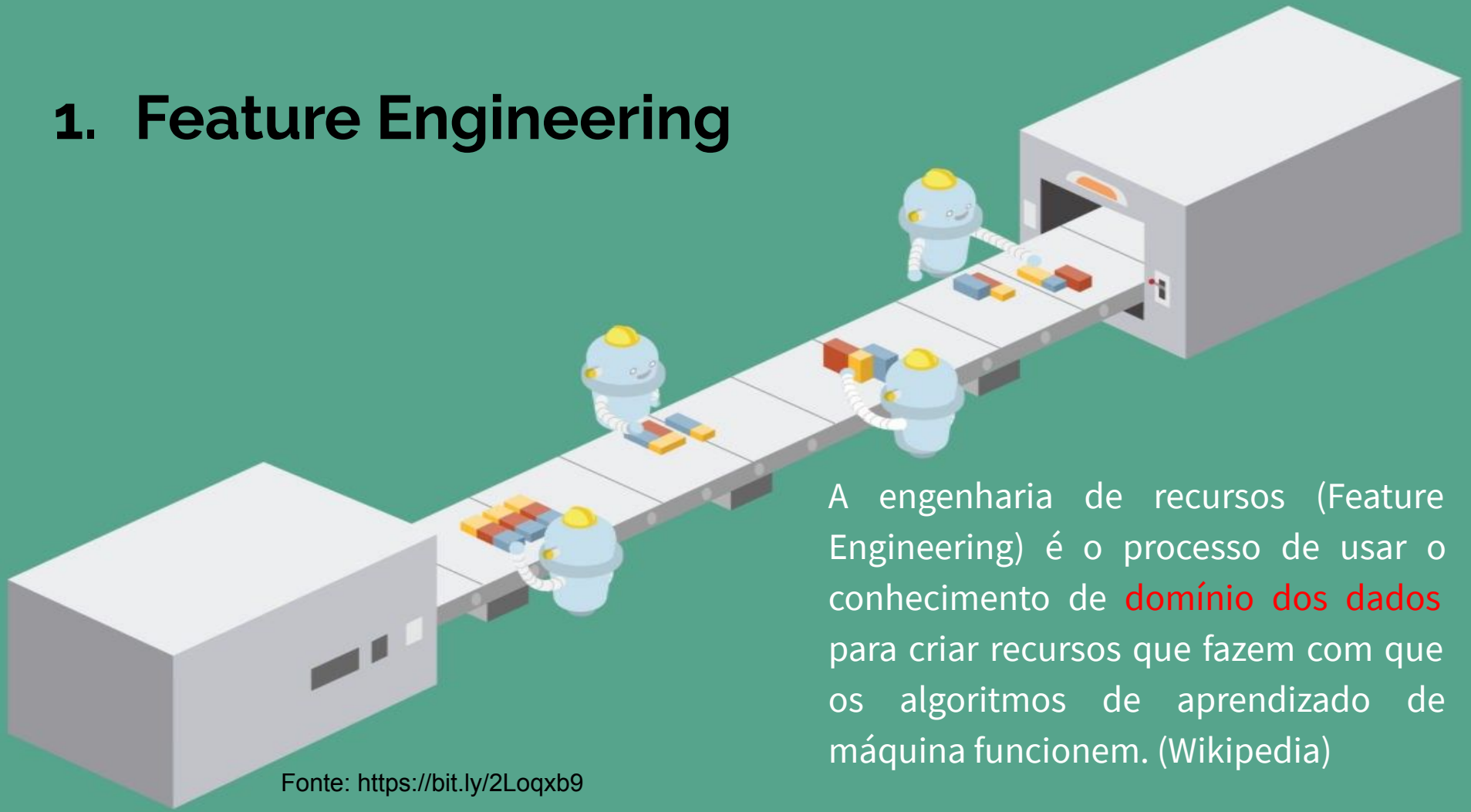


https://github.com/RodrigoCMoraes/pydata2019

Conceitos

1. Feature Engineering
2. Split do Dataset
3. PCA
4. Métricas de validação
5. Modelo

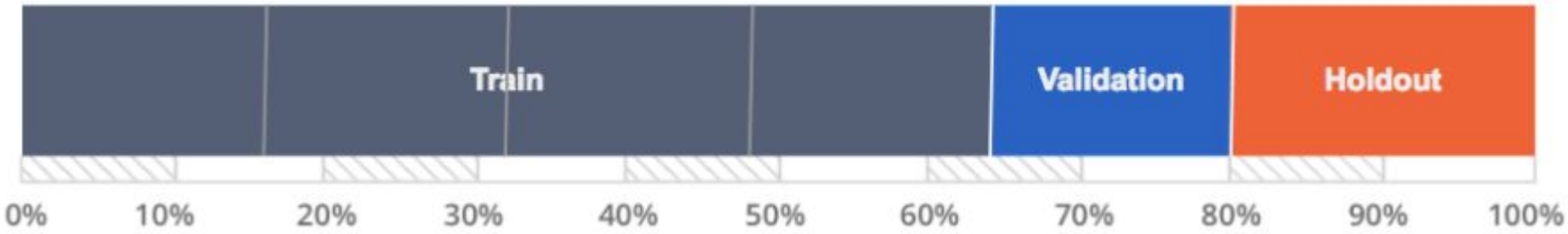
1. Feature Engineering



Fonte: https://bit.ly/2Loqxb9

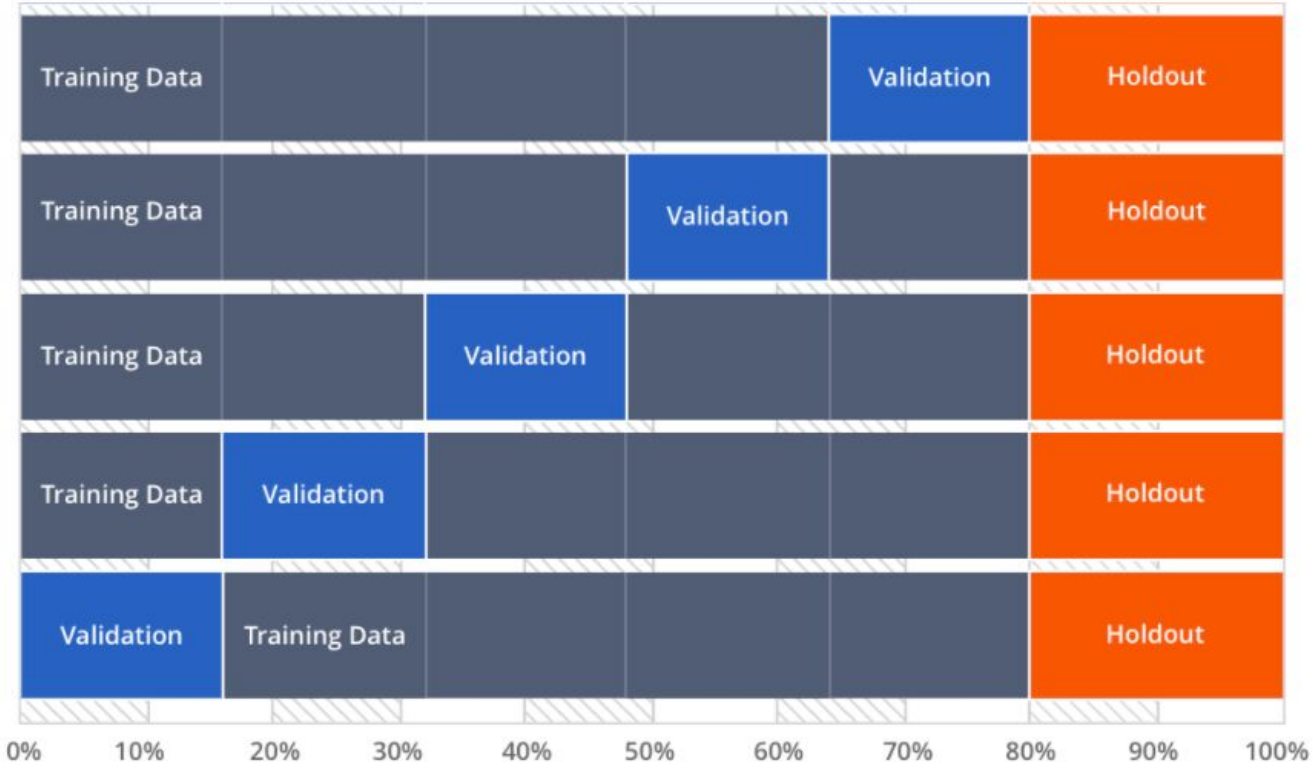
A engenharia de recursos (Feature Engineering) é o processo de usar o conhecimento de domínio dos dados para criar recursos que fazem com que os algoritmos de aprendizado de máquina funcionem. (Wikipedia)

2. Split do Dataset: Holdout



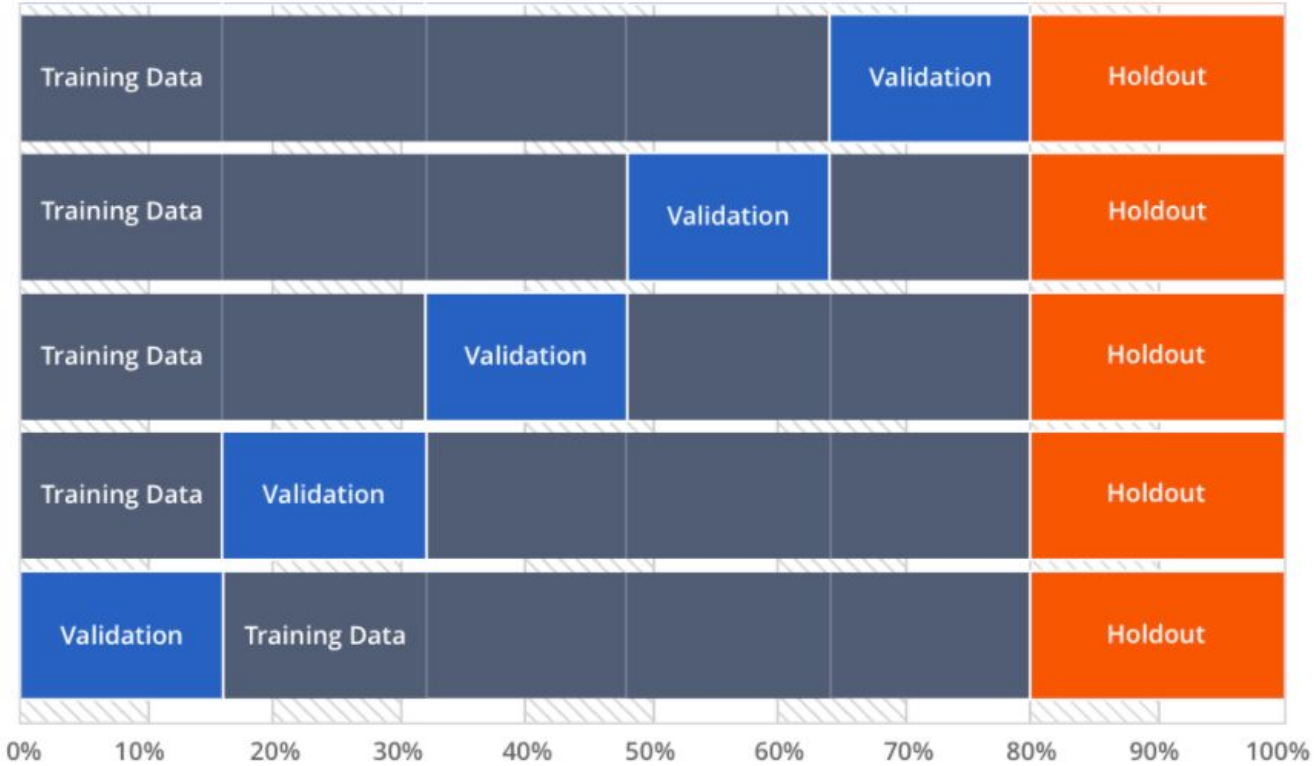
Fonte: https://bit.ly/2JnwjHJ

2. Split do Dataset: K-Fold+Validação Cruzada



Fonte: https://bit.ly/2JnwjHJ

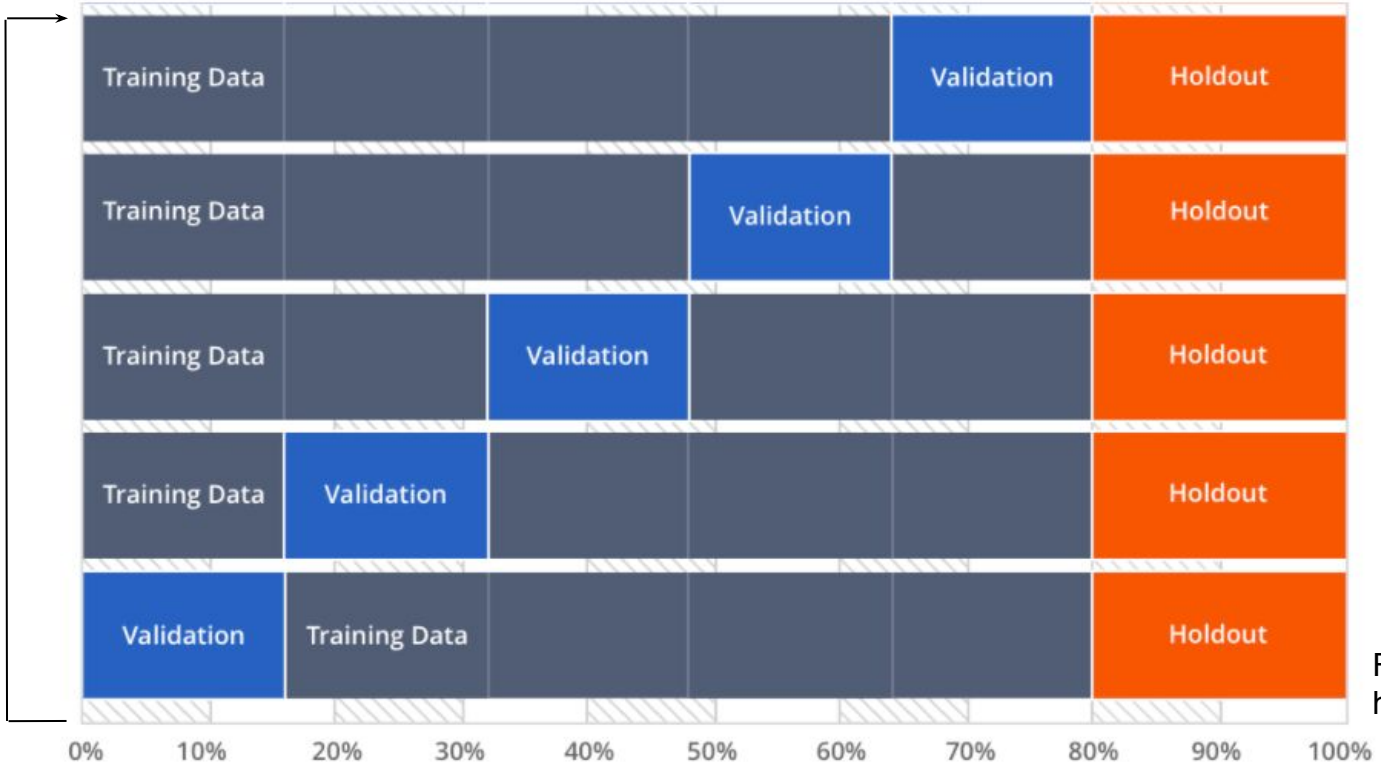
2. Split do Dataset: K-Fold+Validação Cruzada



Fonte: https://bit.ly/2JnwjHJ

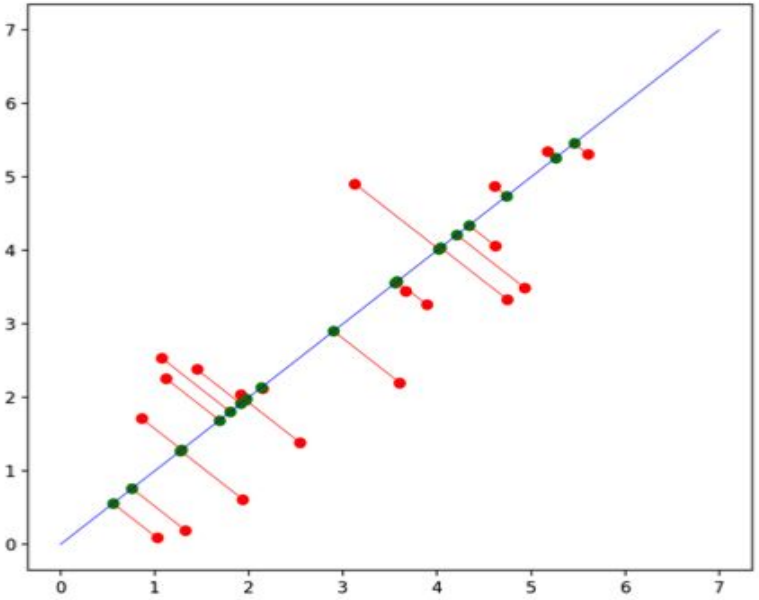
2. Split do Dataset: RSKF+Validação Cruzada

Randomize Dataset

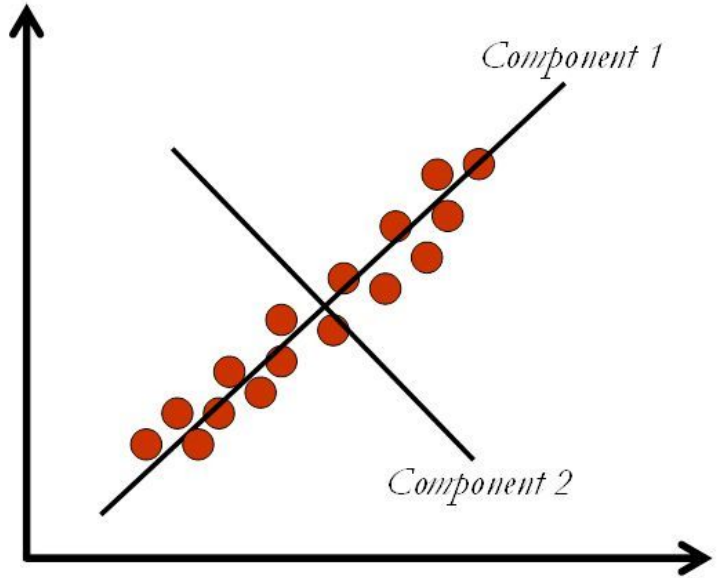


Fonte: https://bit.ly/2JnwjHJ

3. PCA - Principal Component Analysis



Fonte: https://bit.ly/2J0PVSi



Fonte: https://bit.ly/2Y5ytj1

4. Métricas de Validação

$$\text{Acurácia} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precisão} = \frac{tp}{tp + fp}$$

$$\text{Revocação} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precis} \cdot \text{revoc}}{\text{precis} + \text{revoc}}$$

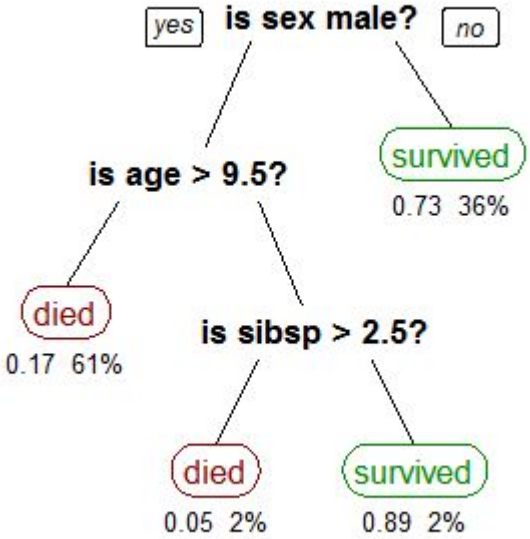
True Positive (TP)

True Negative (TF)

False Positive (FP)

False Negative (FN)

5. Modelo



Fonte: https://bit.ly/2jnSH5w