

# 99% de acurácia... tá, mas e aí?

By Rodrigo C. Moraes



<https://github.com/rodrigocmoraes>



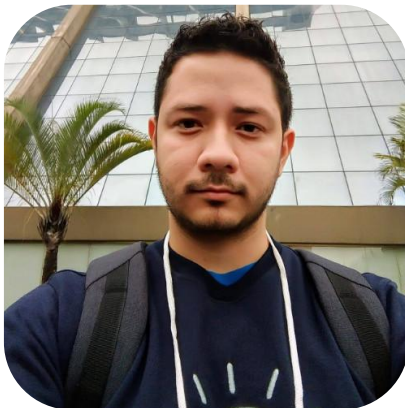
[rdcmdev@gmail.com](mailto:rdcmdev@gmail.com)

kaggle

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



# Quem sou eu?



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>

RodrigoCMoraes / pydata2019

Unwatch 1 Star 0 Fork 1

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

This repository contains notebook and code source for community be able to reproduce content from "99% de Acurácia... tah mas e ai?"

Manage topics

5 commits 1 branch 0 releases 2 contributors

Branch: master New pull request

Create new file Upload files Find File Clone or download

RodrigoCMoraes Merge pull request #1 from wdsrocha/master Latest commit 9f578ce 3 hours ago

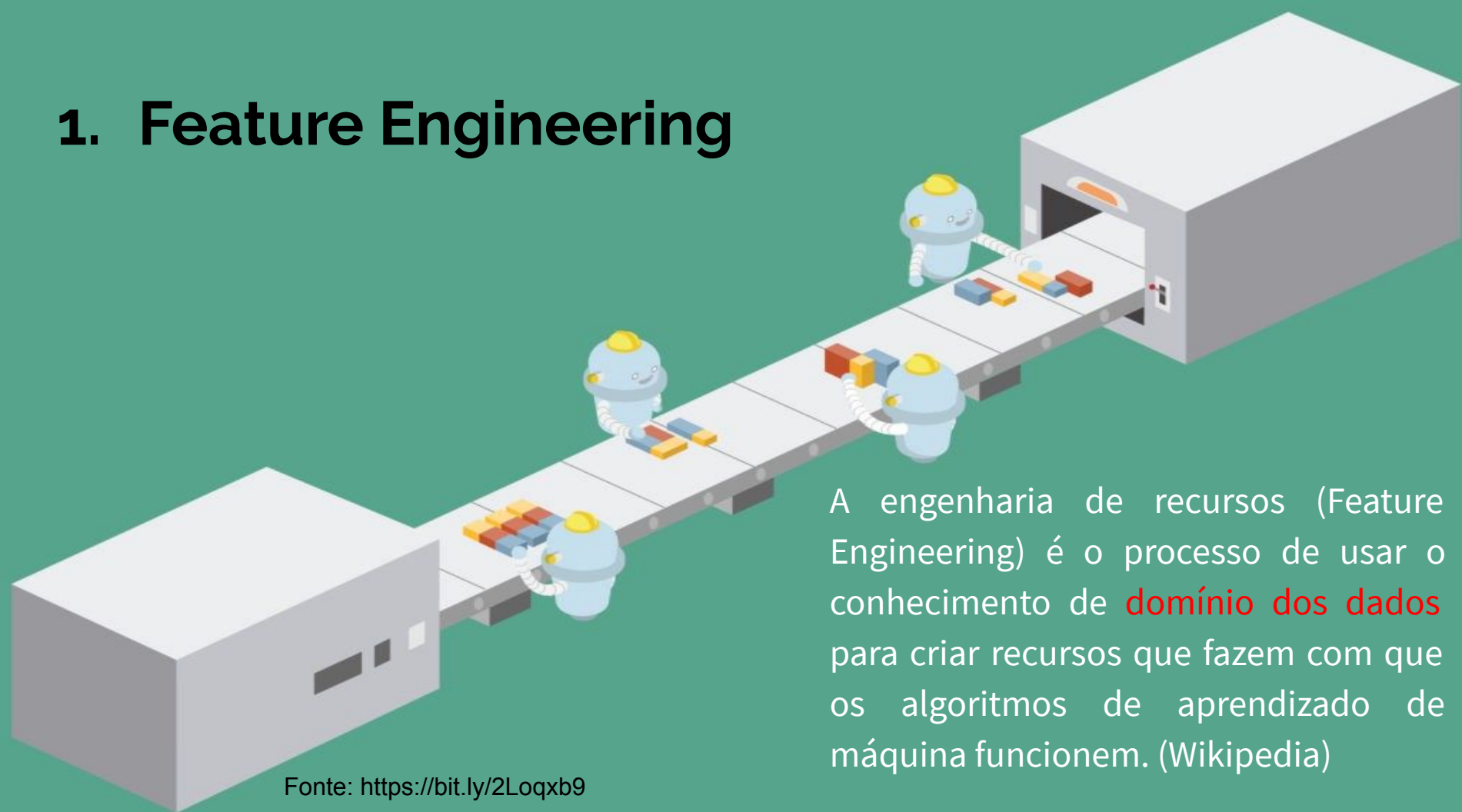
Pipfile	chore: create repository	23 hours ago
Pipfile.lock	chore: create repository	23 hours ago
README.md	docs: change jupyter-notebook command	6 hours ago
adult.csv	chore: create repository	23 hours ago
install.sh	refactor: merge scripts and synchronize README	6 hours ago
notebook.ipynb	chore: run all notebook	23 hours ago

<https://github.com/RodrigoCMoraes/pydata2019>

# Conceitos

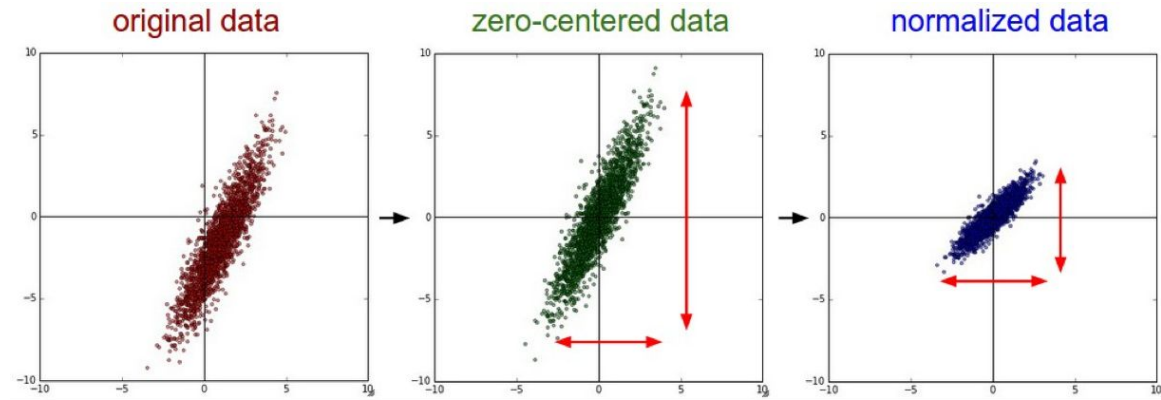
1. Feature Engineering
  2. PCA
  3. Split do Dataset
  4. Métricas de validação
  5. Modelo
  6. Overview
-

# 1. Feature Engineering



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)

# 1. Feature Engineering - Reorganizing



Fonte: <https://bit.ly/2Y7NYXY>



# 1. Feature Engineering - Polynomial Features

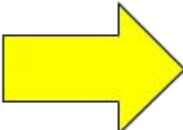
## Examples

```
>>> X = np.arange(6).reshape(3, 2)
>>> X
array([[0, 1],
       [2, 3],
       [4, 5]])
>>> poly = PolynomialFeatures(2)
>>> poly.fit_transform(X)
array([[ 1.,  0.,  1.,  0.,  0.,  1.],
       [ 1.,  2.,  3.,  4.,  6.,  9.],
       [ 1.,  4.,  5., 16., 20., 25.]])
```

Fonte: <https://bit.ly/2Y1Hx8w>

# 1. Feature Engineering - Encoding

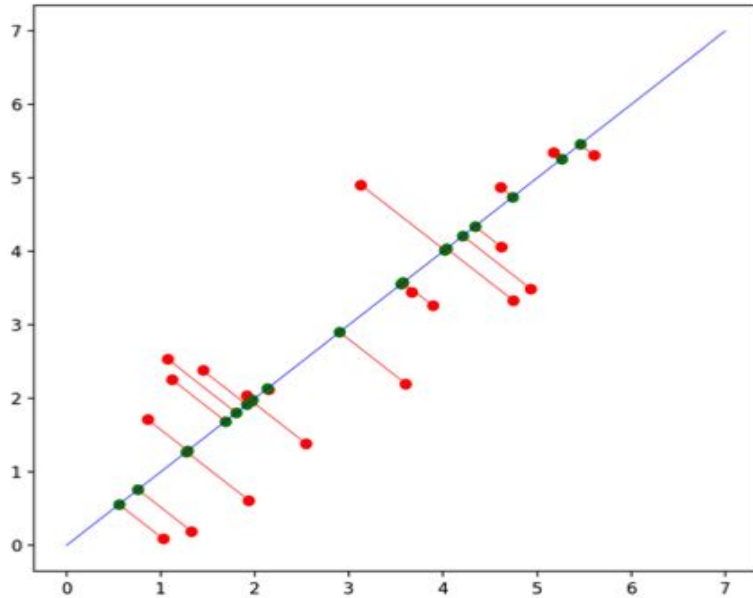
Color	
Red	
Red	
Yellow	
Green	
Yellow	



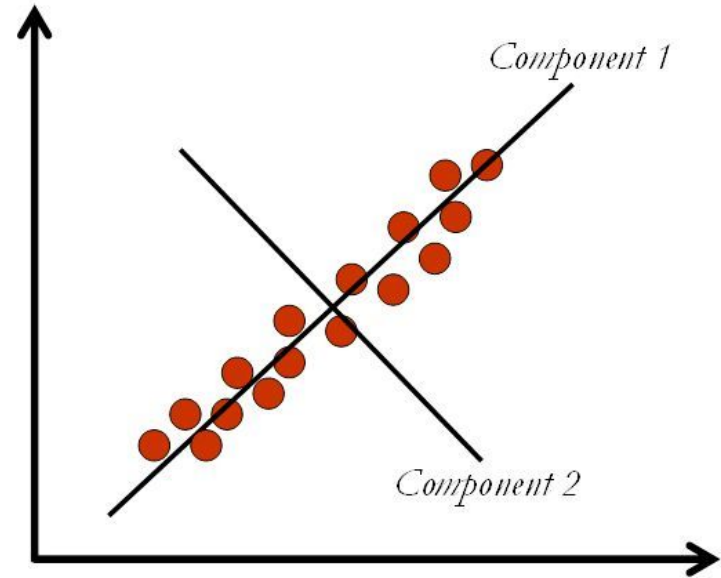
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Fonte: <https://bit.ly/2V0D443>

## 2. PCA - Principal Component Analysis

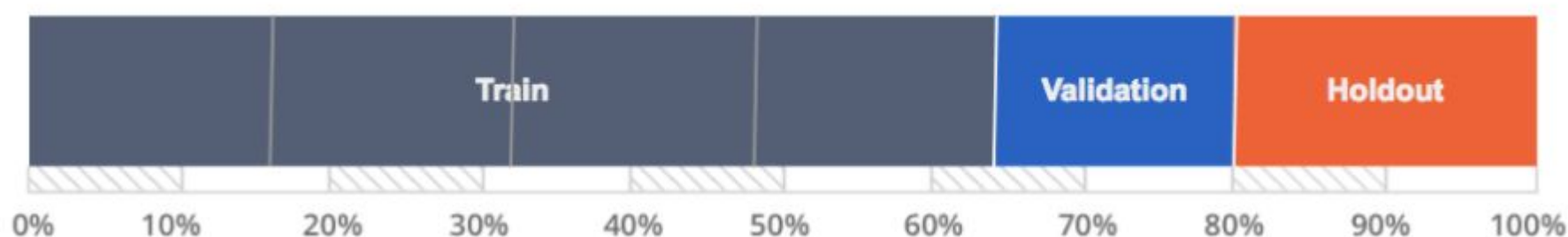


Fonte: <https://bit.ly/2J0PVSi>



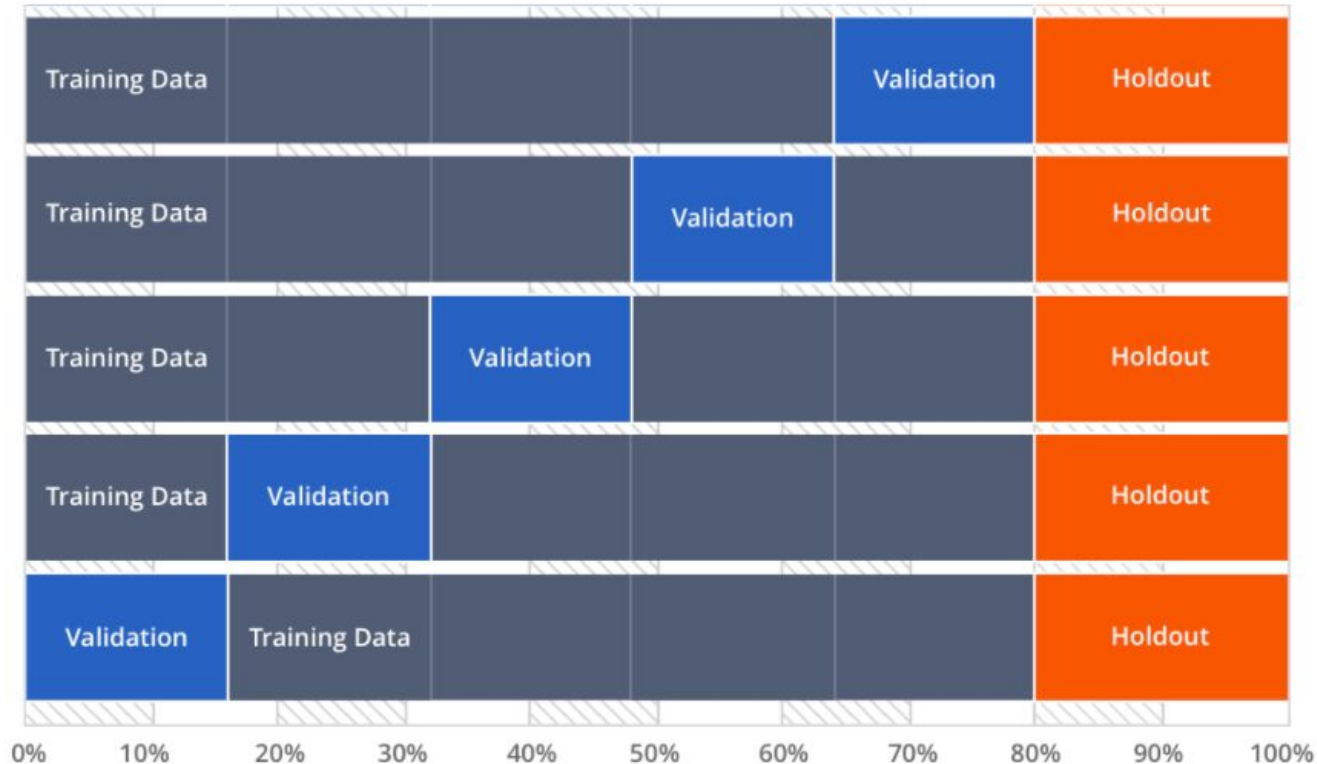
Fonte: <https://bit.ly/2Y5ytj1>

### 3. Split do Dataset: Holdout



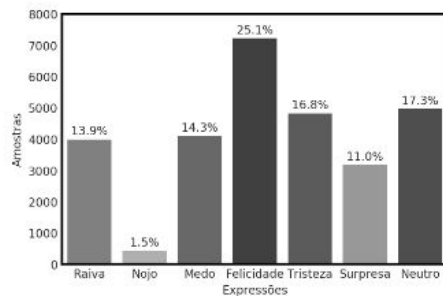
Fonte: <https://bit.ly/2JnwjHJ>

### 3. Split do Dataset: K-Fold+Validação Cruzada

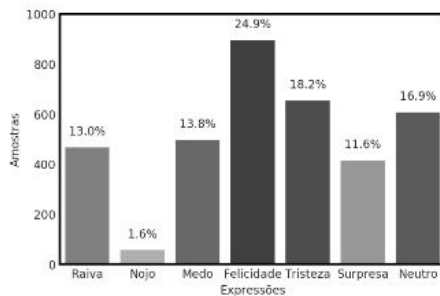


Fonte: <https://bit.ly/2JnwjHJ>

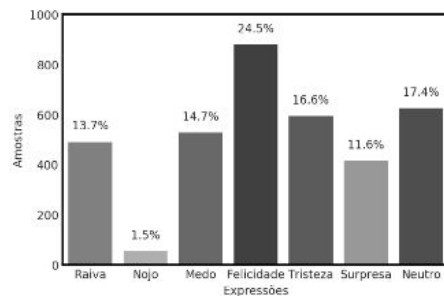
### 3. Split do Dataset: Stratificação



(a) Treinamento



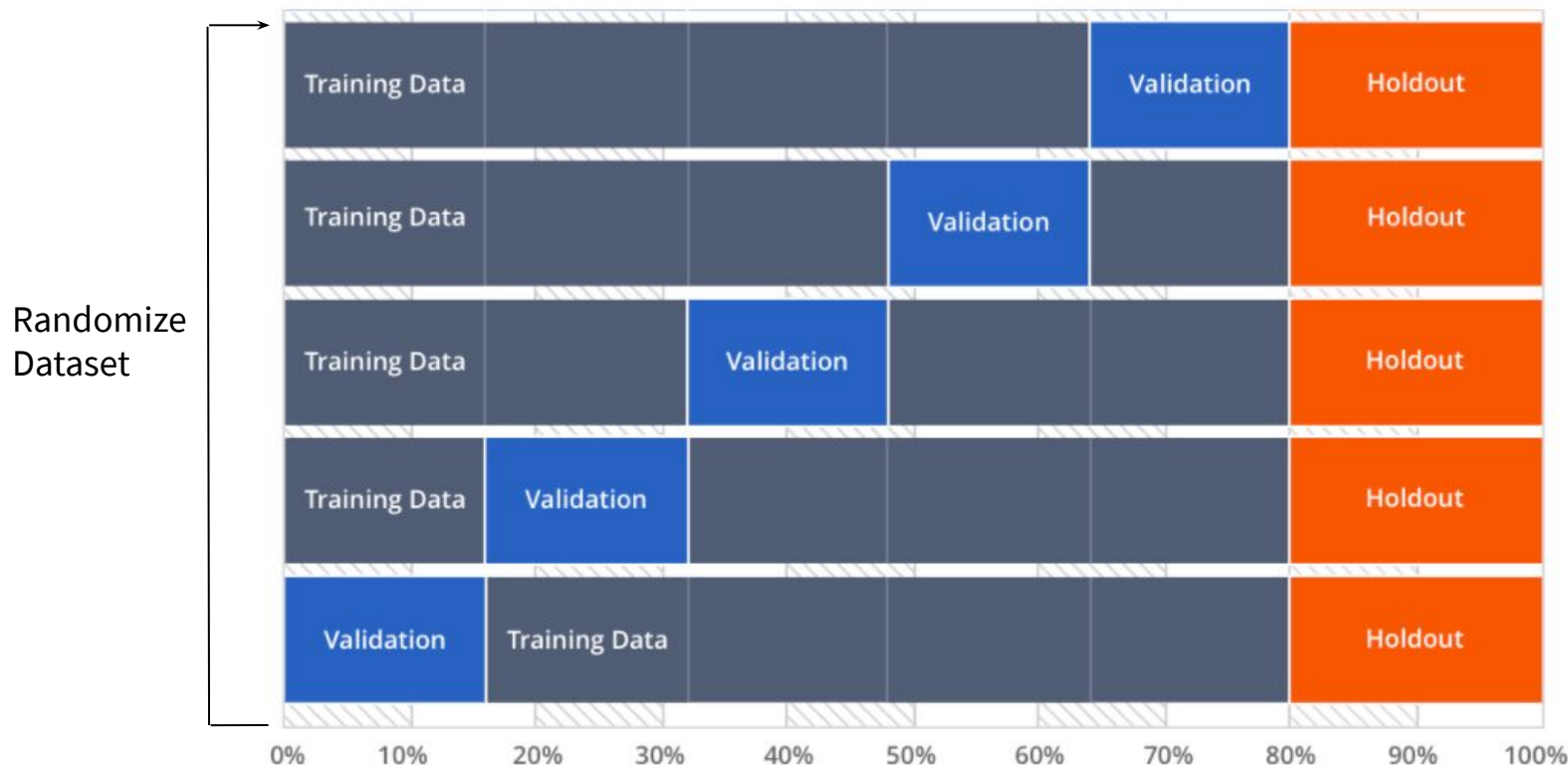
(b) Validação



(c) Teste

Fonte: Própria

### 3. Split do Dataset: RSKF+Validação Cruzada



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

## 4. Métricas de Validação

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

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

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

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

Fonte: <https://bit.ly/2ZXLYDi>

True Positive (TP)

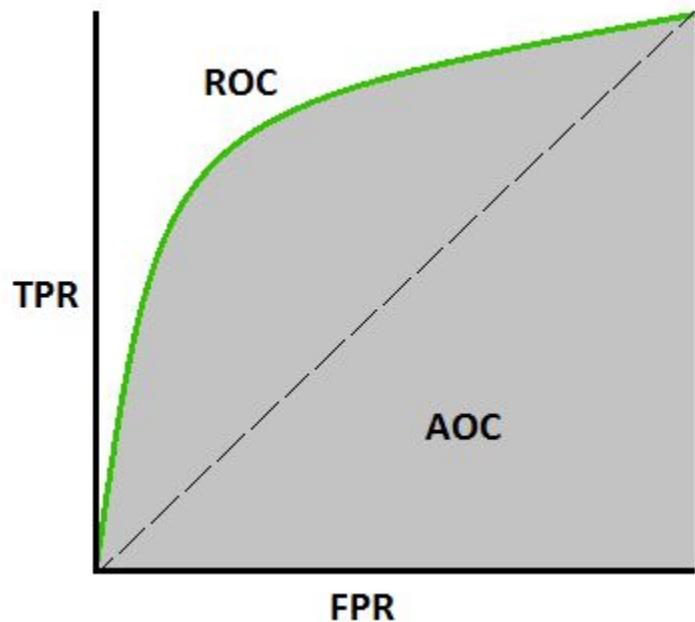
True Negative (TF)

False Positive (FP)

False Negative (FN)



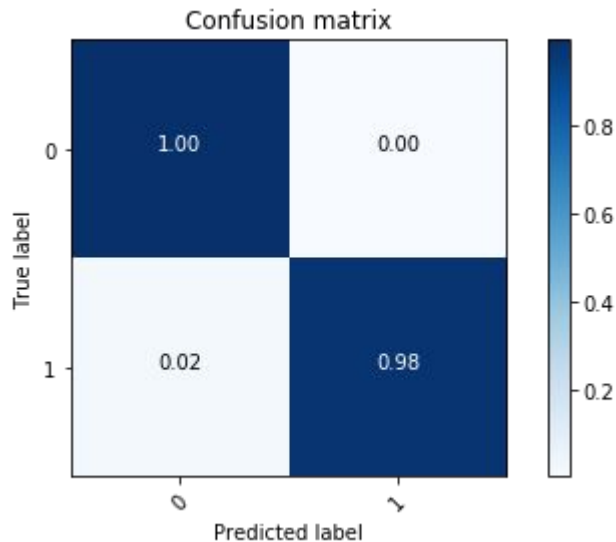
## 4. Métricas de Validação - AUC



Fonte: <https://bit.ly/2E0YdqU>

AUC - Area Under  
the Curve

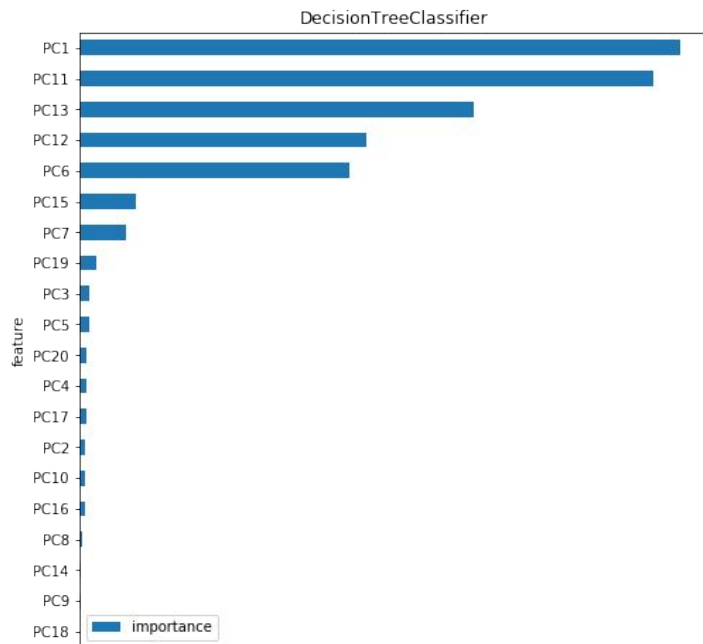
## 4. Métricas de Validação - Confusion Matrix



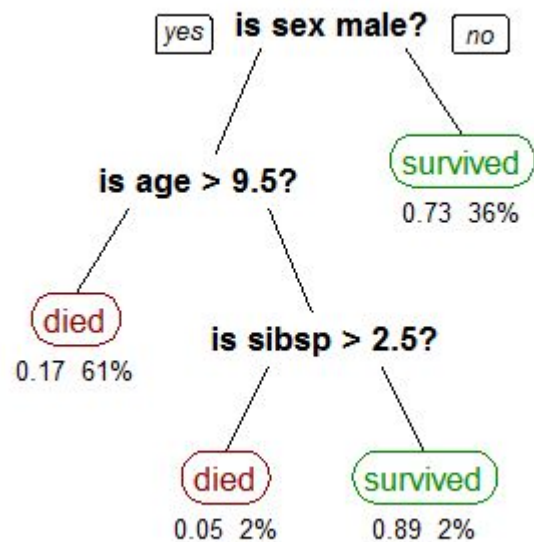
Fonte: própria

Confusion Matrix

# 5. Modelo

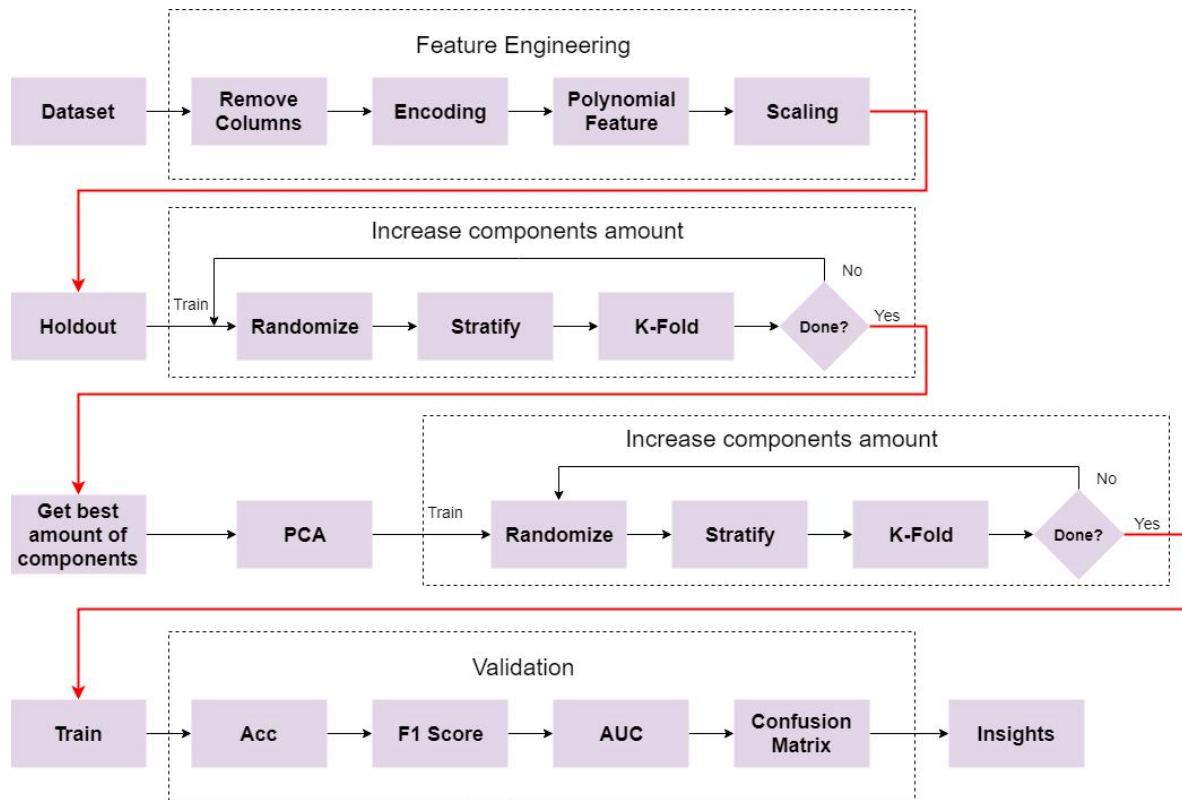


Fonte: Própria



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

# 5. Overview



**Bora codar?**

[Notebook](#)

