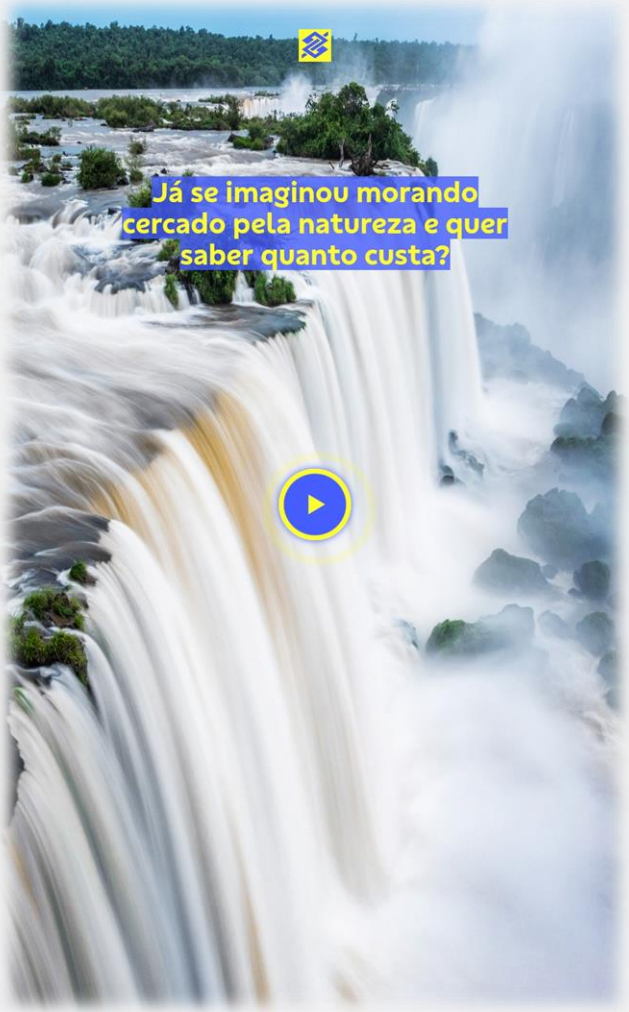
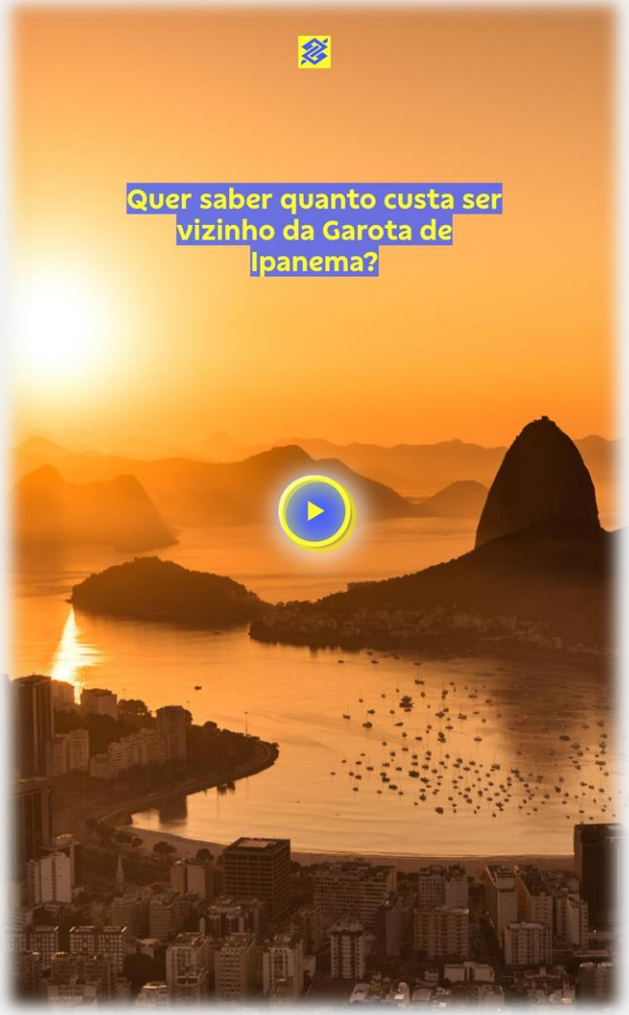
Análise de Dados em Python para Não Programadores

Patrícia Aguiar Moreira

Estudo de Caso: Valoração de Imóveis

Quanto custa o imóvel dos seus sonhos?





O que vamos ver?

- Acesso a base de dados
- Insights a partir de visualizações
- Desenvolvimento de um modelo preditivo
- Simulação de uso do modelo

Onde vamos realizar as análises?

Linguagem de Programação e Ambiente

https://colab.research.google.com/github/mpatricia/workshop/blob/main/Notebook.ipynb

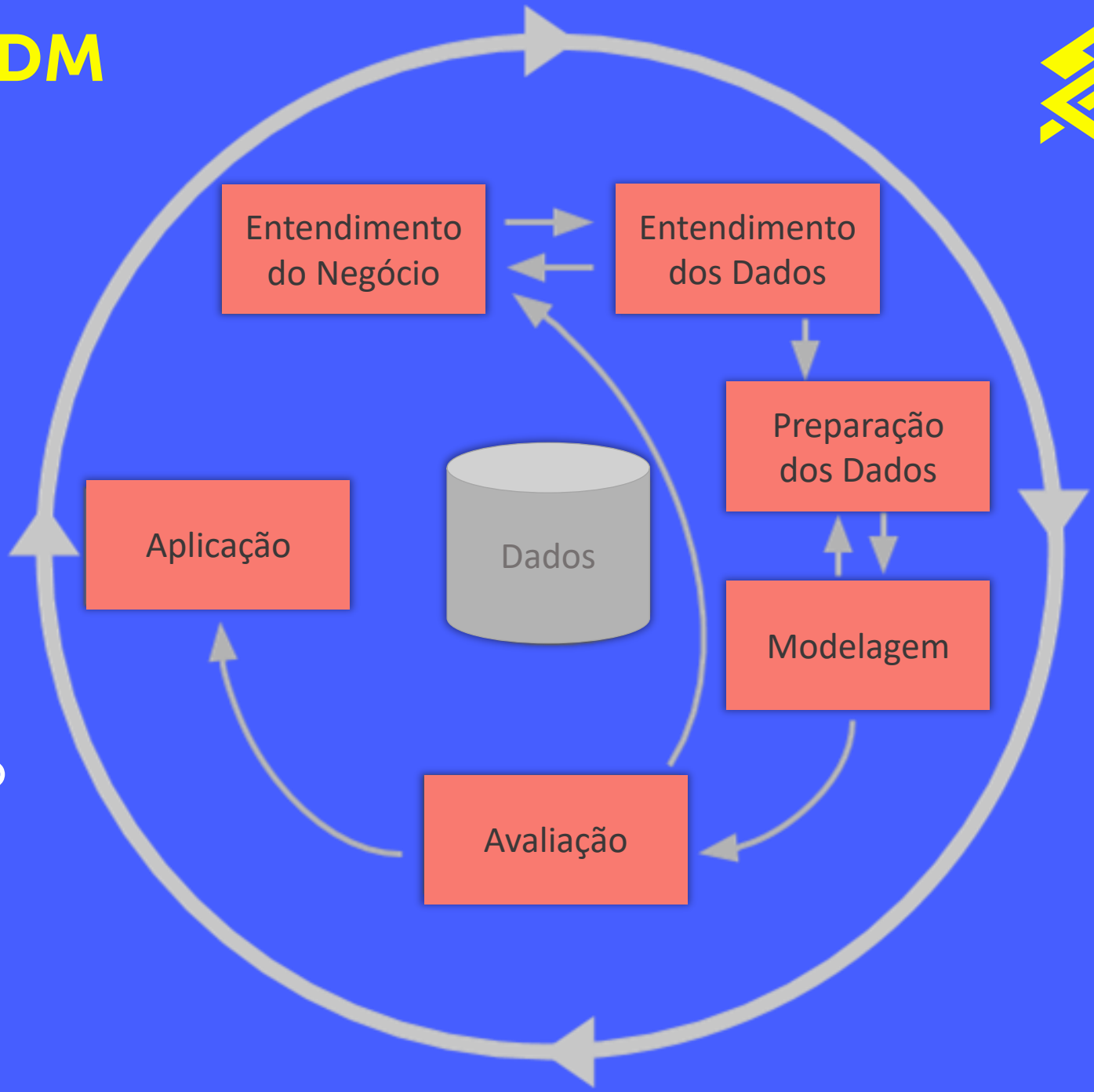
https://labblite.bb.com.br/

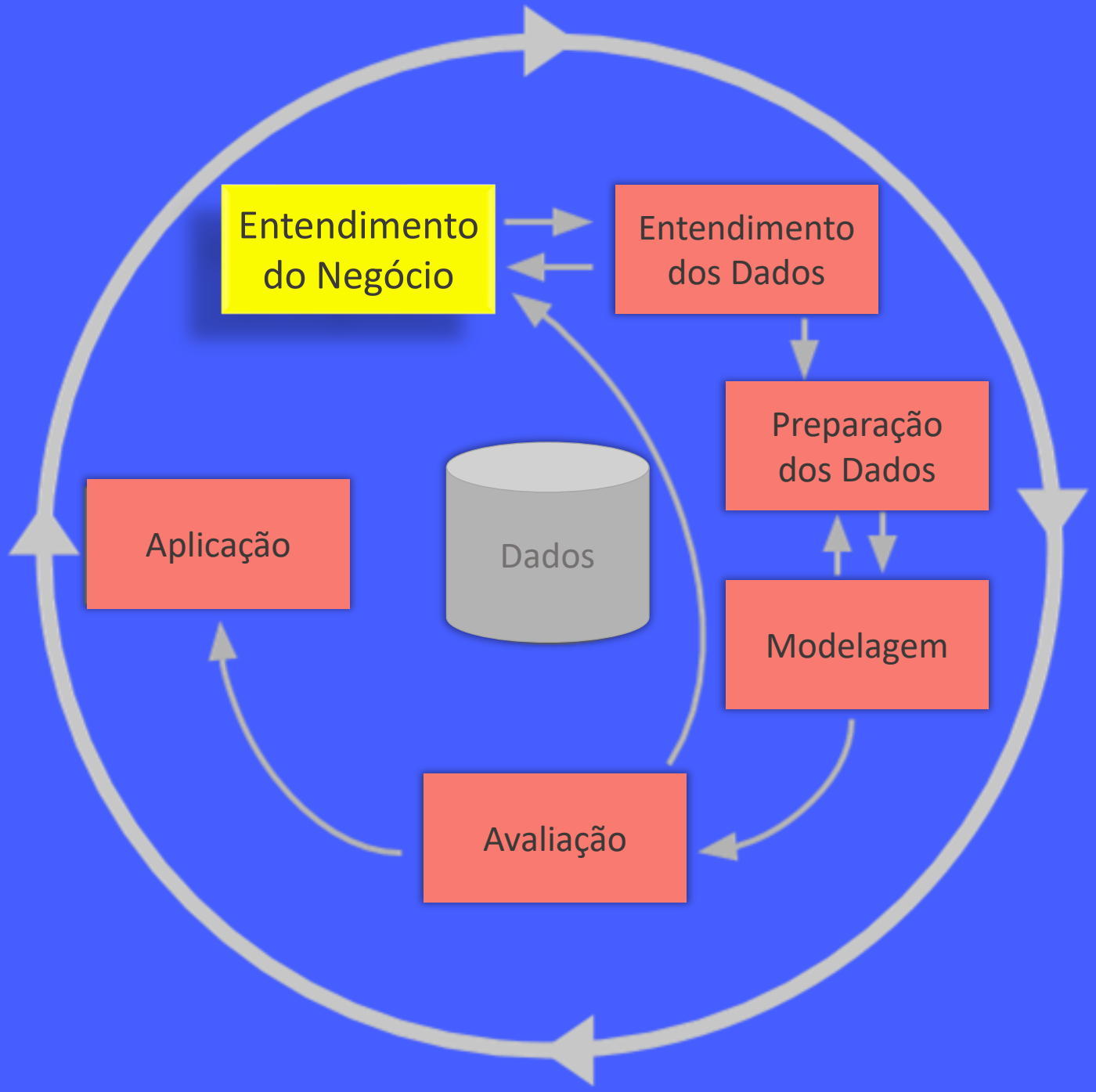
Fonte da imagem: https://makeameme.org/meme/are-you-ready-26dd8a

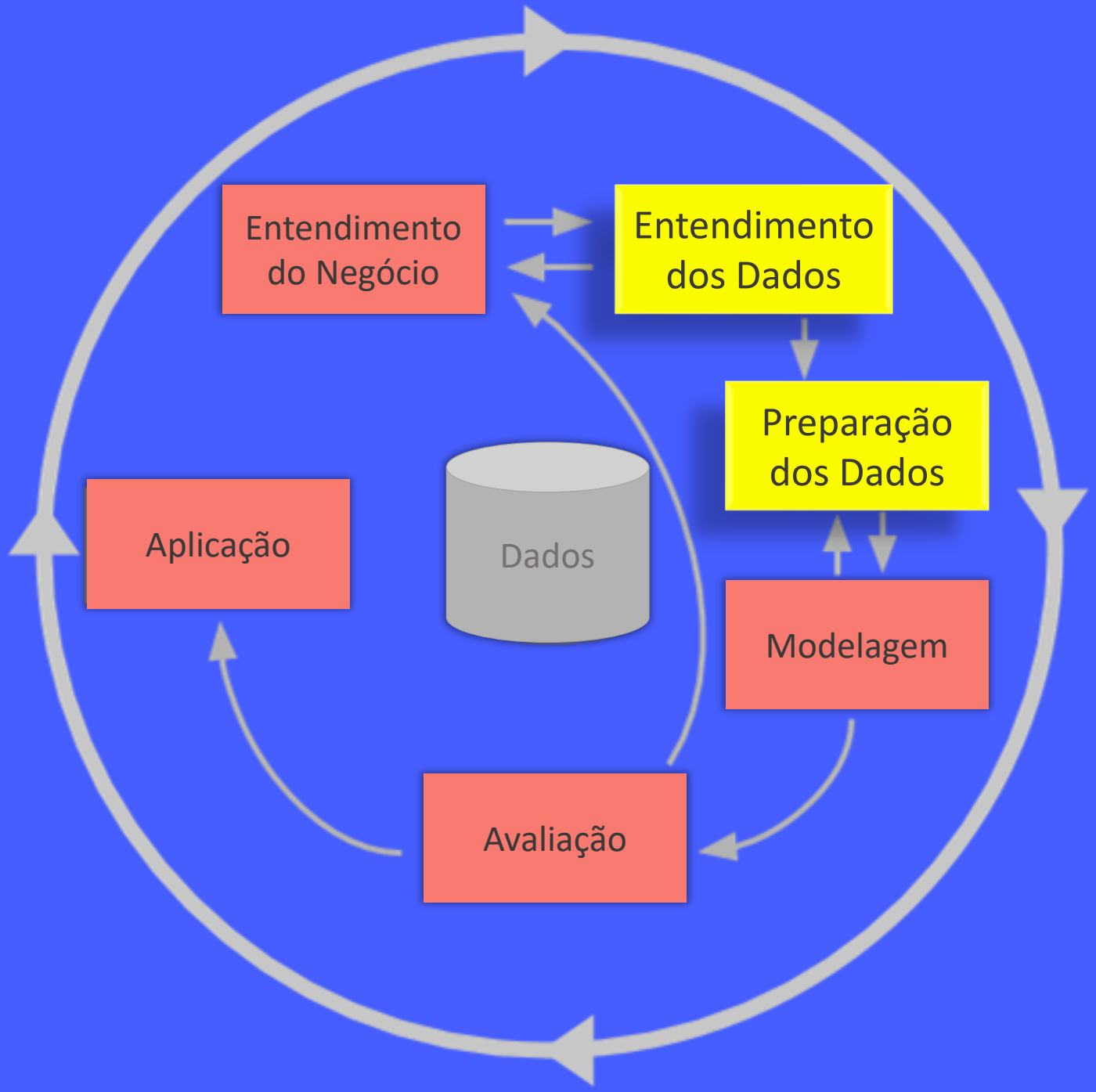
Metodologia CRISP-DM

CRoss
Industry
Standard
Process for
Data **M**ining

É um modelo de processo para análise de dados




Entendimento do Negócio
Entendimento dos Dados
Preparação dos Dados
Modelagem
Avaliação
Aplicação
Dados


Entendimento do Negócio
Entendimento dos Dados
Preparação dos Dados
Modelagem
Avaliação
Aplicação
Dados

Entendimento e Preparação dos Dados

Fontes de Origem → Dados Estruturados

Atributos/Características
nas colunas

Observações/Exemplos
nas linhas

ID	Tem vista para orla?	Índice de Qualidade	Número de Quartos	Preço
1	S	3	2	500.000,00
2	N	7	1	600.000,00
3	N	3	4	2.000.000,00
4	S	11	6	5.000.000,00
5	N	5	2	700.000,00

Identificador
da observação

Inputs
Variáveis Explicativas

Output
Variável Resposta/Target

Entendimento e Preparação dos Dados

Qual o tratamento adequado?

Excluir atributos **não relevantes**

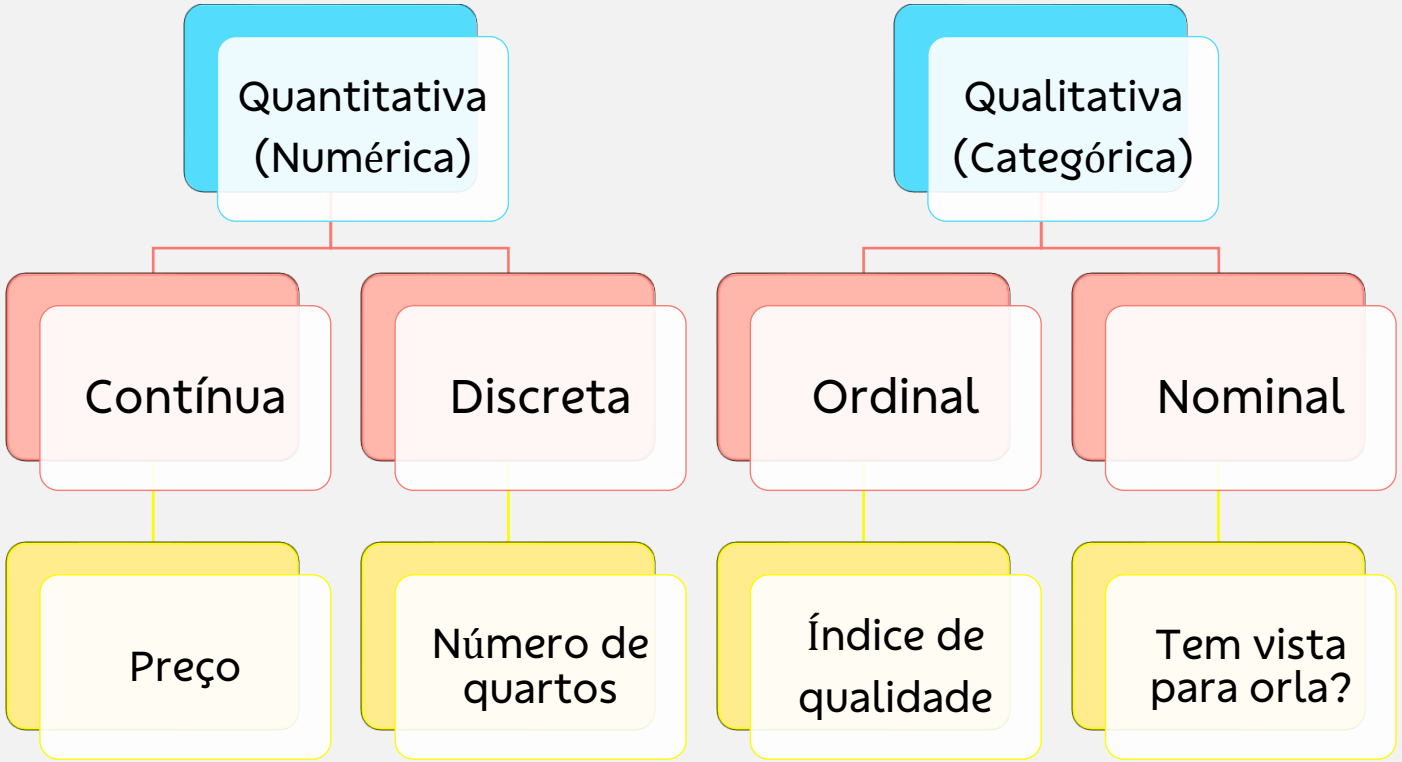
Os algoritmos não interpretam **textos** (**variáveis qualitativas**), é preciso codificá-los em números.

ID	Tem vista para orla?	Índice de Qualidade	Número de Quartos	Preço
1	S	3	2	500.000,00
2	N	7	1	600.000,00
3	N	3	4	2.000.000,00
4		11	6	5.000.000,00
5	N	5	2	700.000,00

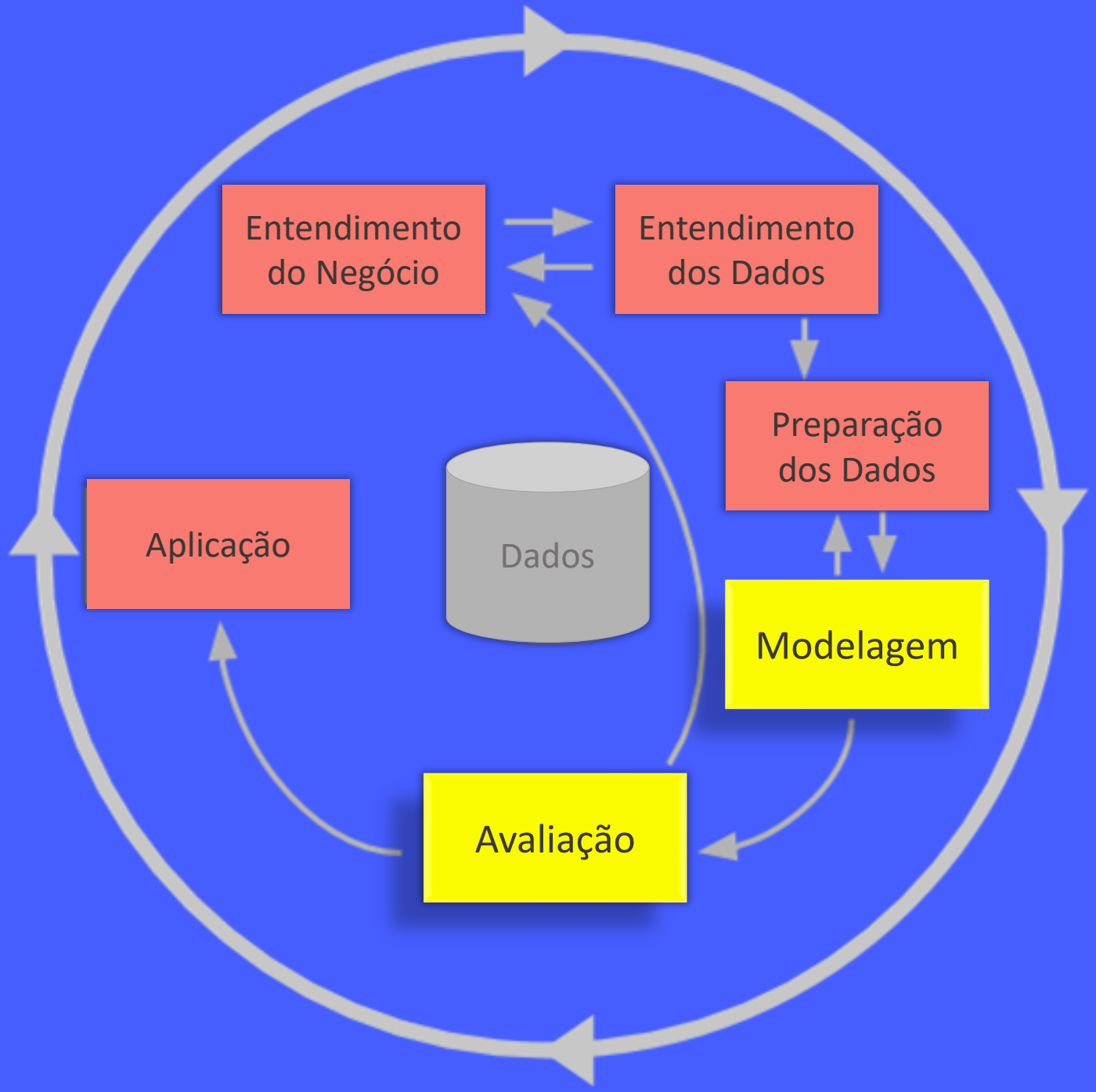
Valores ausentes são informação, se foram observados nos dados treino, é provável observar nos novos dados. Como tratá-los?

Entendimento e Preparação dos Dados

Qual o tipo das variáveis?

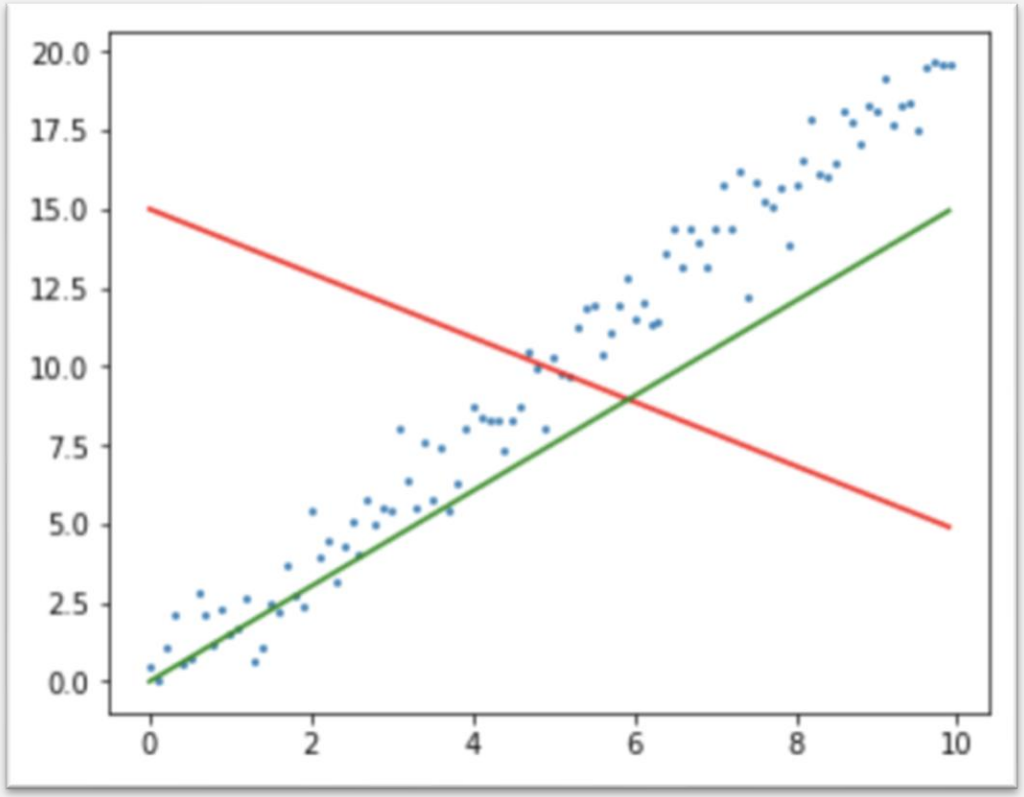


Fonte da imagem: http://www.quickmeme.com/meme/3q2vq6


Entendimento do Negócio
Entendimento dos Dados
Preparação dos Dados
Modelagem
Avaliação
Aplicação
Dados

Modelagem

Como prever y a partir de x?



Qual reta melhor descreve a relação entre x e y?

$$y = 15 - 1x$$

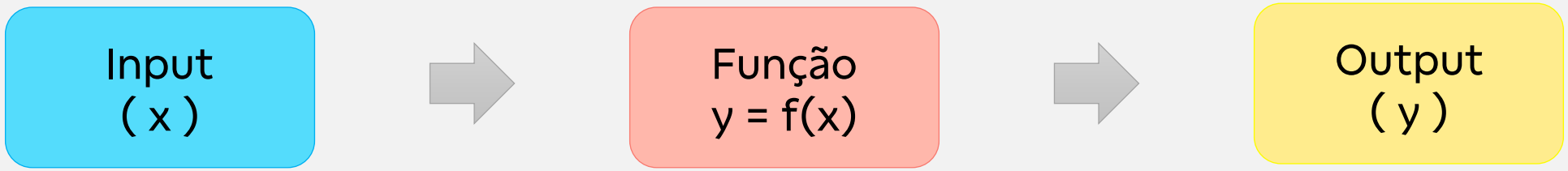
$$y = 1,5x$$

Fonte da imagem: https://sigmoidal.ai/como-implementar-regressao-linear-com-python/

Modelagem

Treinamento do Modelo Preditivo

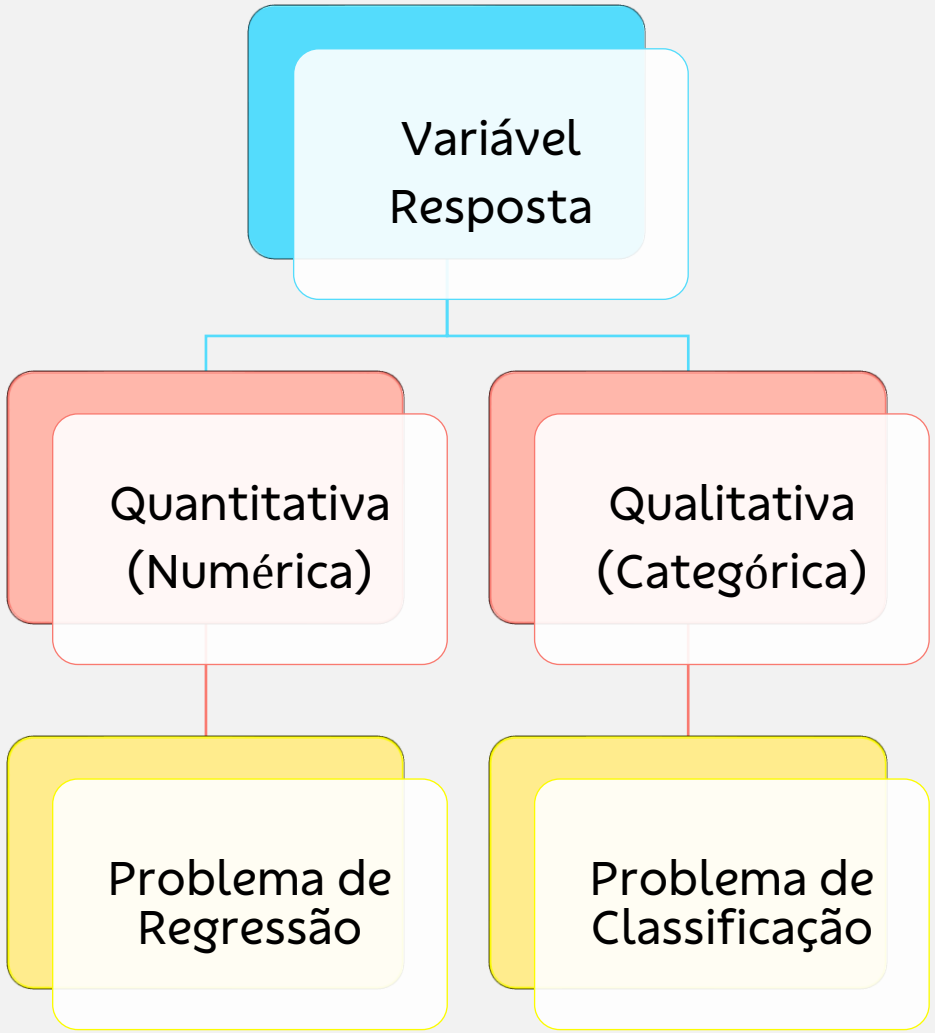
Temos um fenômeno em estudo onde cada valor **y** pode ser encontrado por uma função desconhecida **y = f(x)**:



O objetivo do treinamento é estimar uma função h (hipótese) que se aproxime da função verdadeira f.

Modelagem

Estimadores

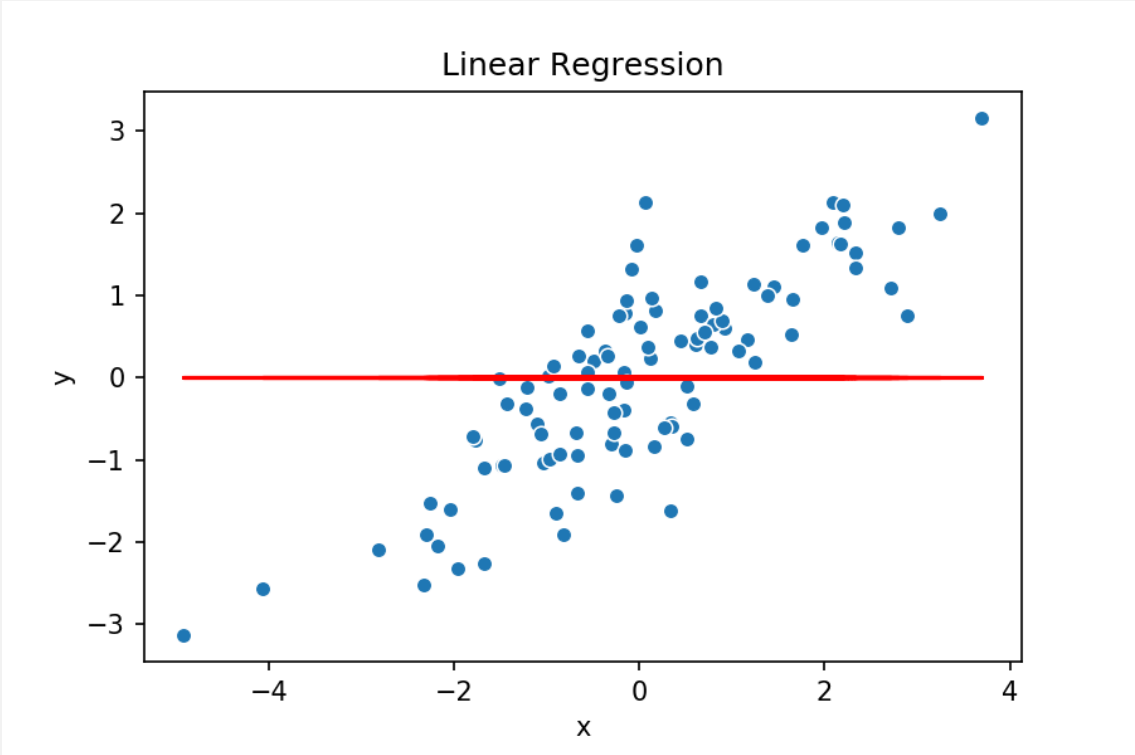


Estimadores para Regressão:

- Regressão Linear
- KNN
- Árvore de Decisão
- Random Forest
- Grandient Boosting

Modelagem

Regressão Linear



Regressor que faz previsões a partir de uma **combinação linear** dos preditores:

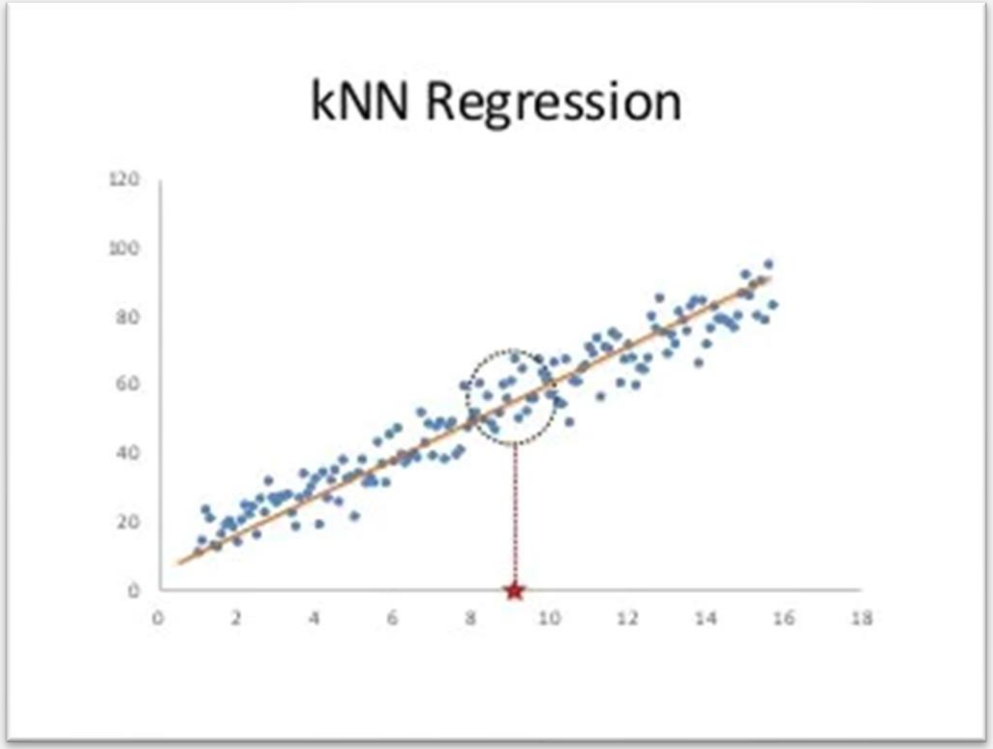
$$y = \alpha + \beta X$$

Fonte da imagem: https://sigmoidal.ai/como-implementar-regressao-linear-com-python/

Modelagem

KNN - K Nearest Neighbors

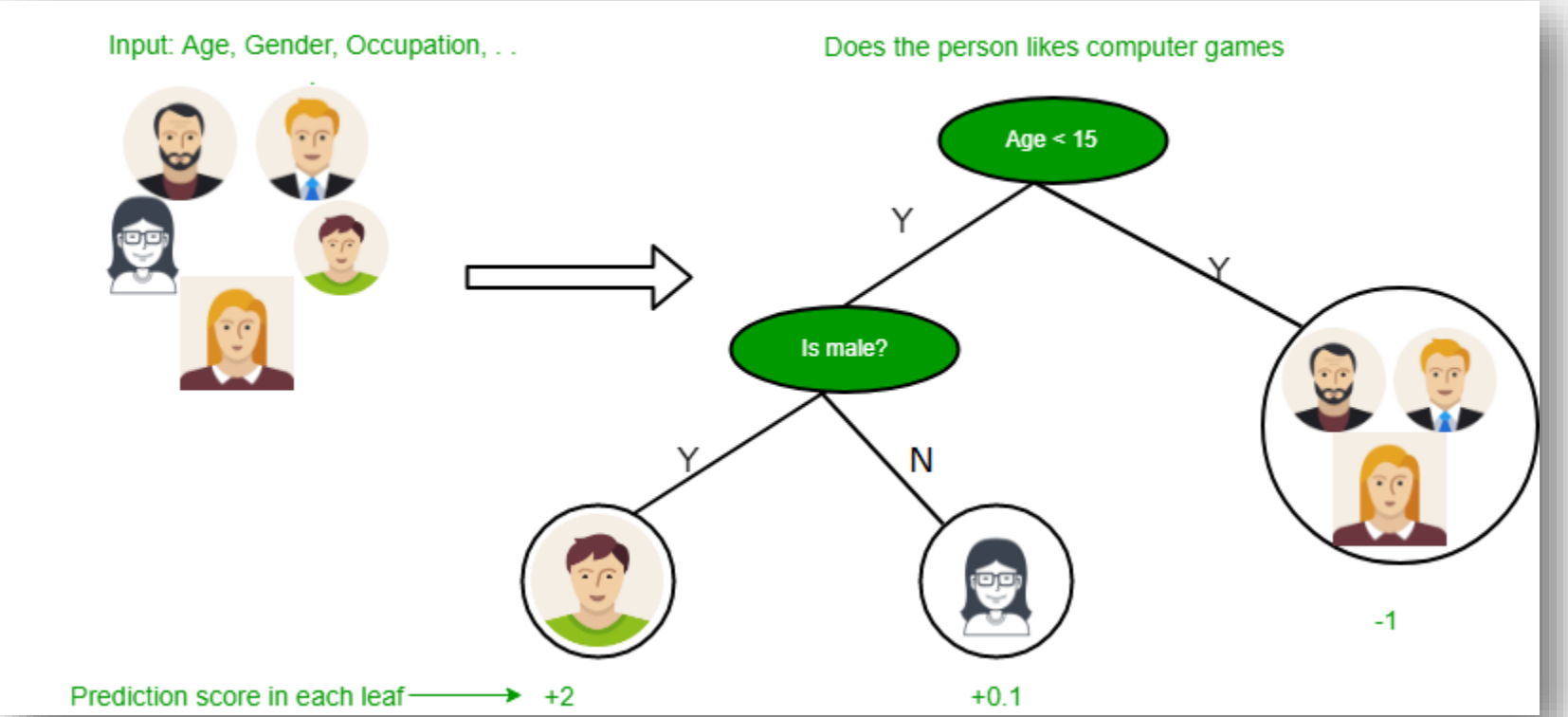
Regressor que faz previsões baseadas nos K vizinhos mais próximos.



Fontes das imagens: https://pt.slideshare.net/amirudind/k-nearest-neighbor-presentation e https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Modelagem

Árvore de Decisão



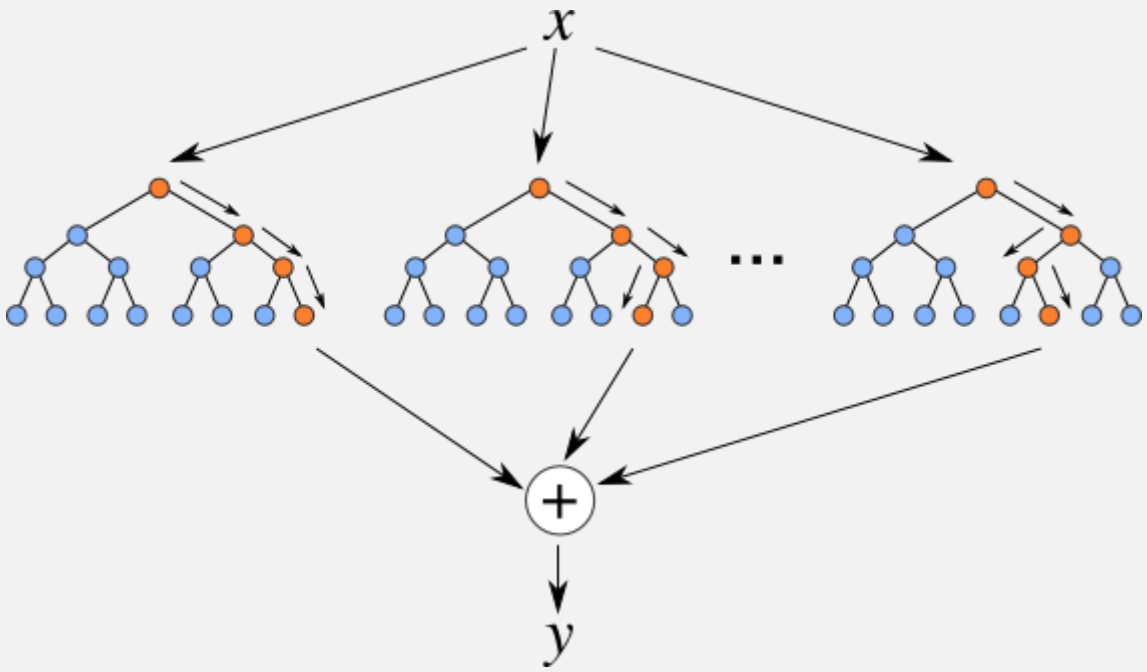
Regressor que, a cada passo, divide as observações em grupos de acordo com a característica que melhor separa os valores da variável target.

Fonte da imagem: https://www.geeksforgeeks.org/decision-tree-introduction-example/

Modelagem

Random Forest

Regressor que consulta vários regressores “fracos” treinados em amostras de variáveis e observações, e faz a previsão com base em uma média das previsões dos regressores “fracos”.

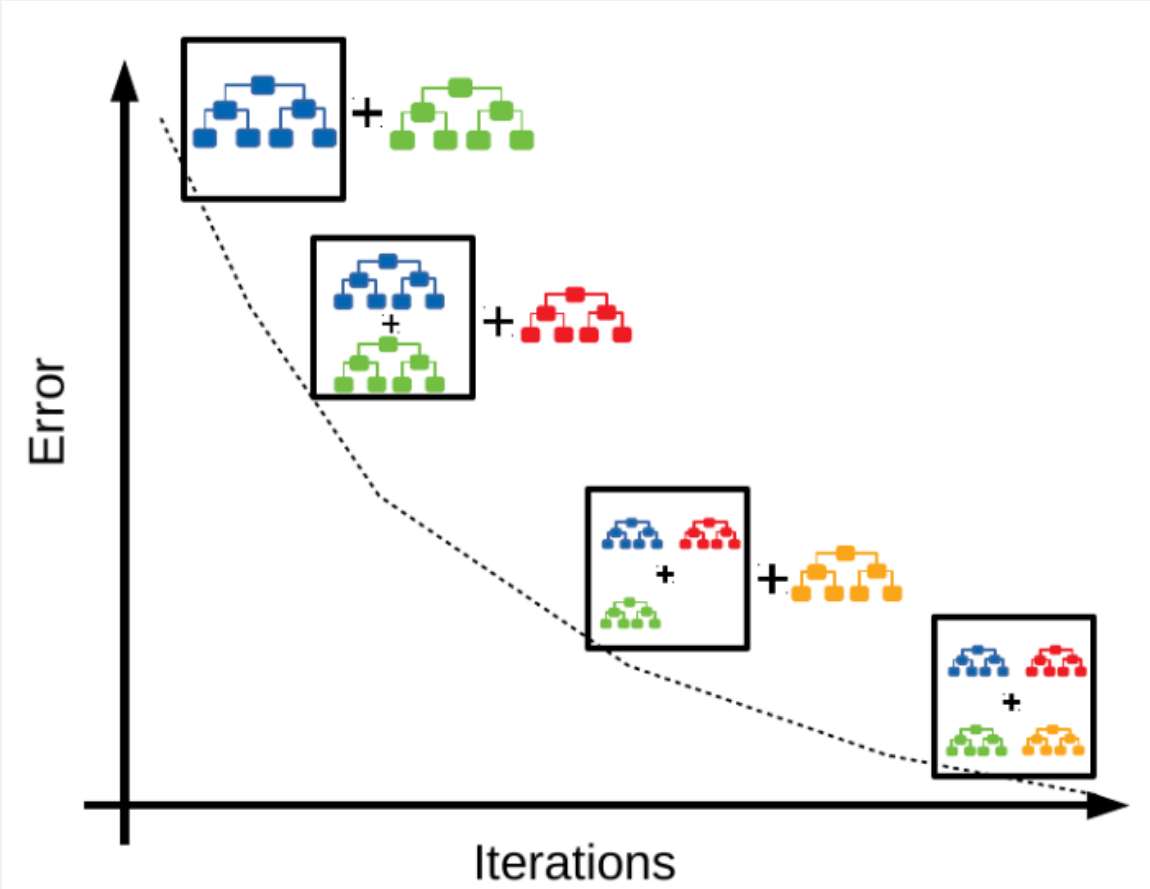


Fonte da imagem: https://miro.medium.com/max/567/1*Mb8awDiY9T6rsOjtNTRcIg.png

Modelagem

Gradient Boosting

Regressor que faz uso de vários regressores “fracos” em sequência, sendo que cada regressor adicionado é treinado para corrigir o erro dos anteriores.

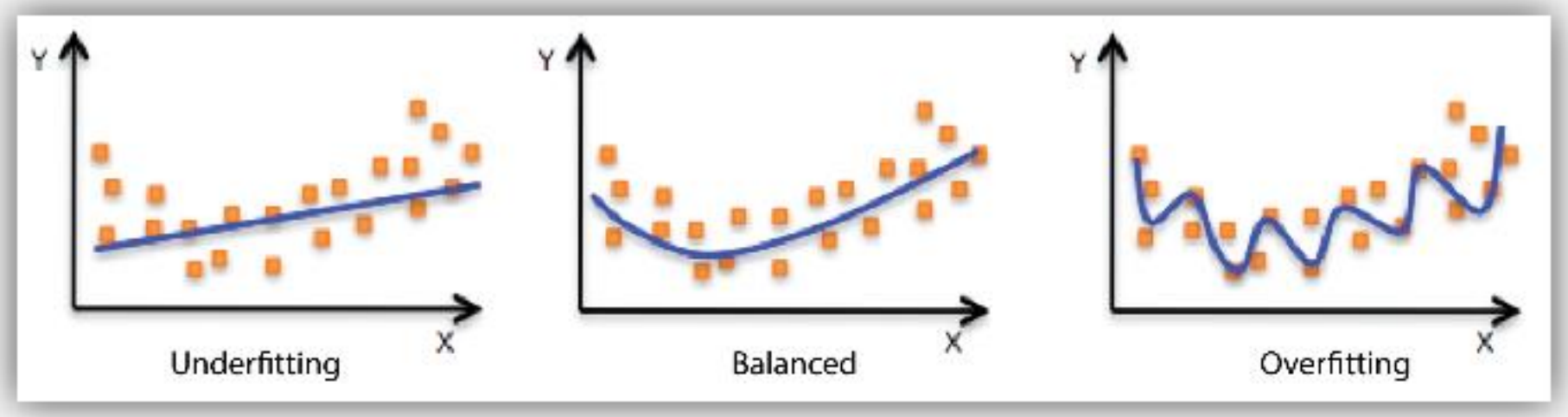


Fonte da imagem: http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html

Modelagem

Underfitting / Overfitting

O estimador vai treinar um modelo que minimiza o erro das previsões nos dados de treino, mas será que é capaz de generalizar para novos casos?

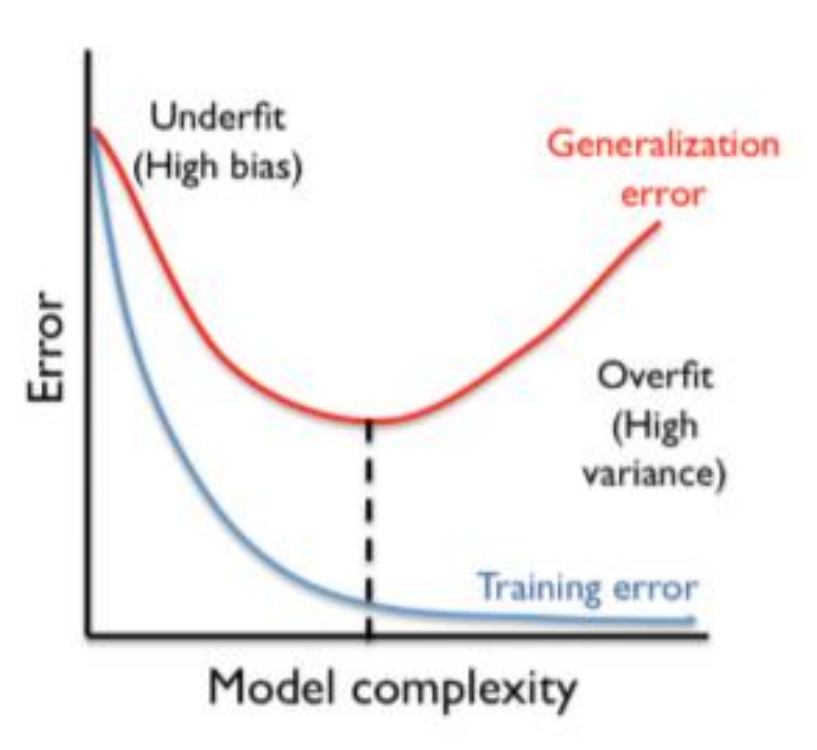


Fonte da imagem: https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/images/mlconcepts_image5.png

Modelagem

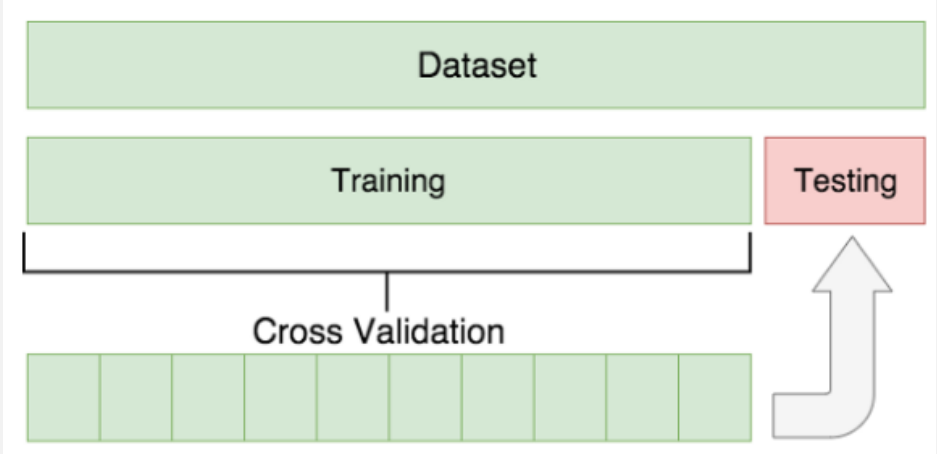
Validação Cruzada

Underfitting/Overfitting



Solução

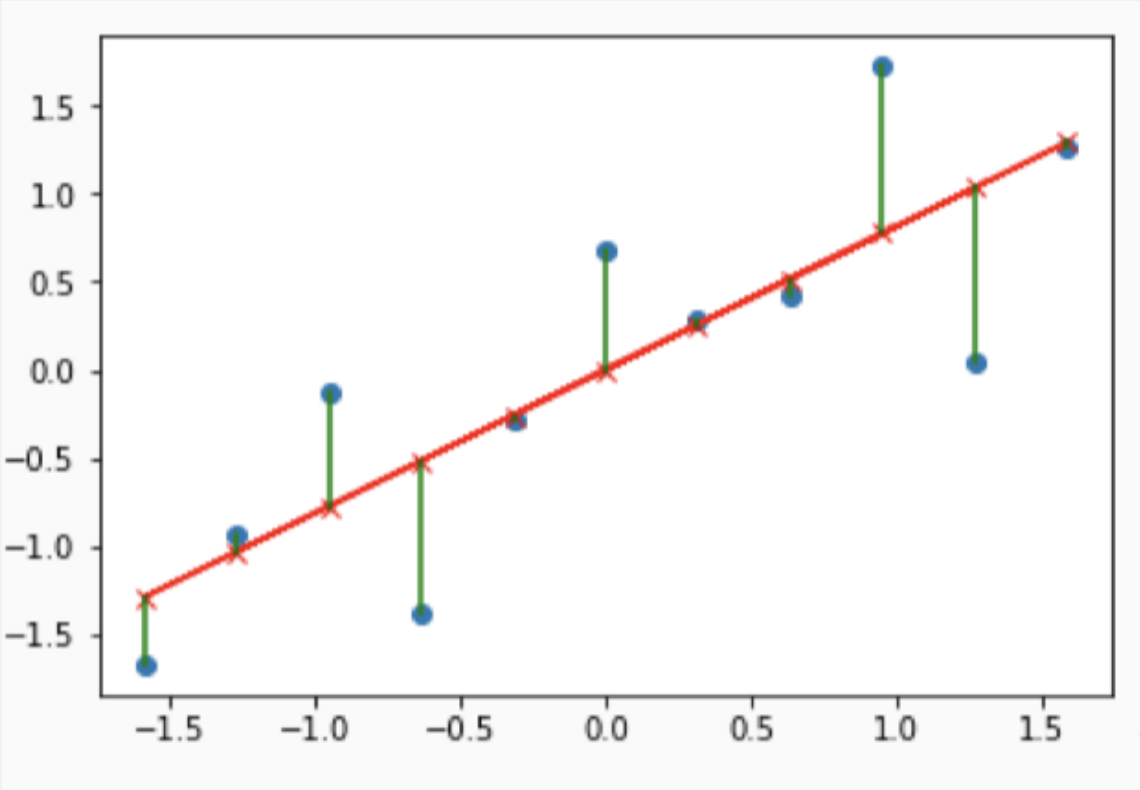
Treinar algoritmo de aprendizado em parte dos dados e reservar a outra parte para mensurar a métrica de performance.



Fonte da imagem: https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Avaliação

Métricas de Performance



Erro absoluto mediano:

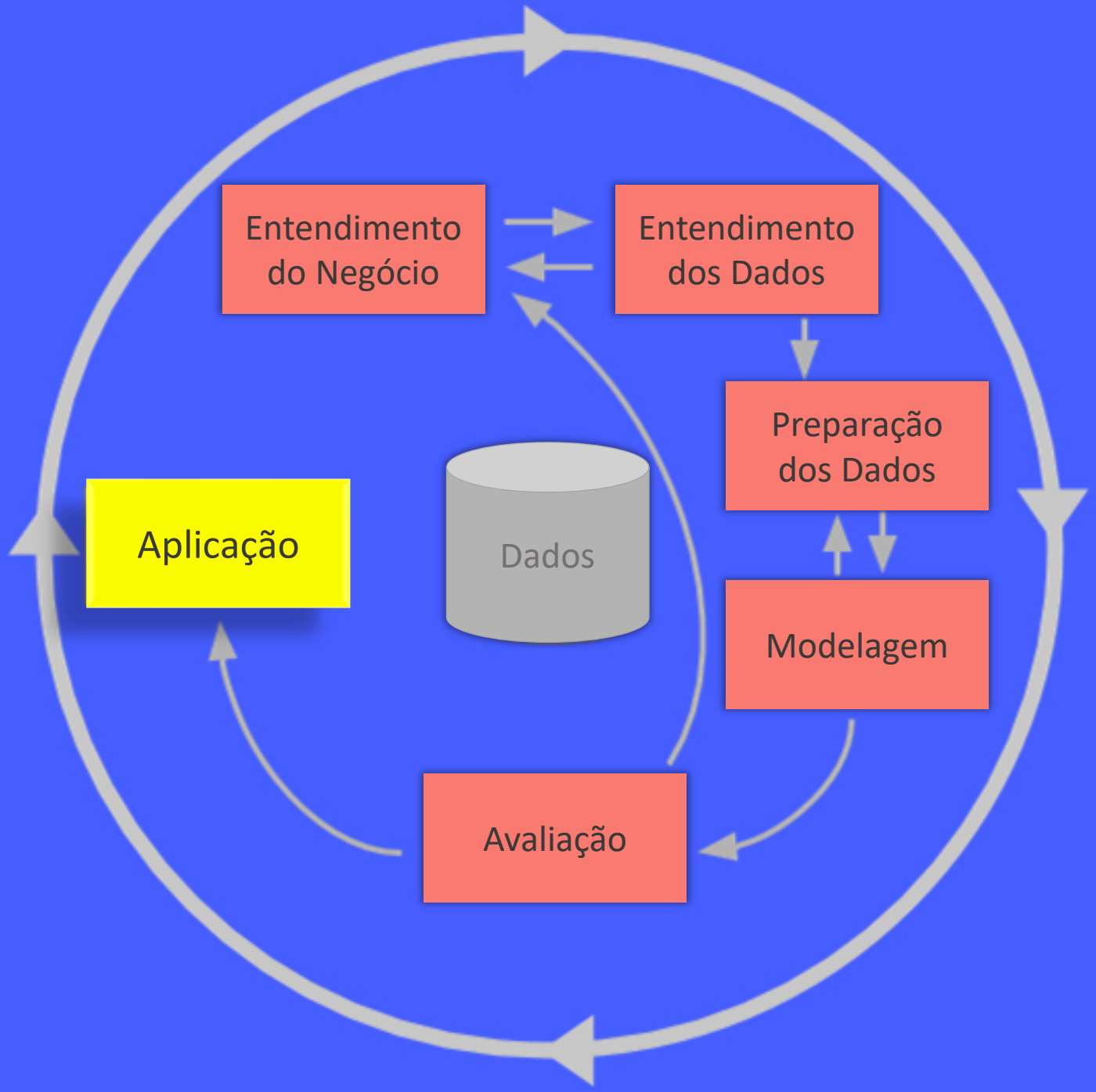
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

R^2: quanto da variância da variável resposta é explicado pelas variáveis explicativas

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

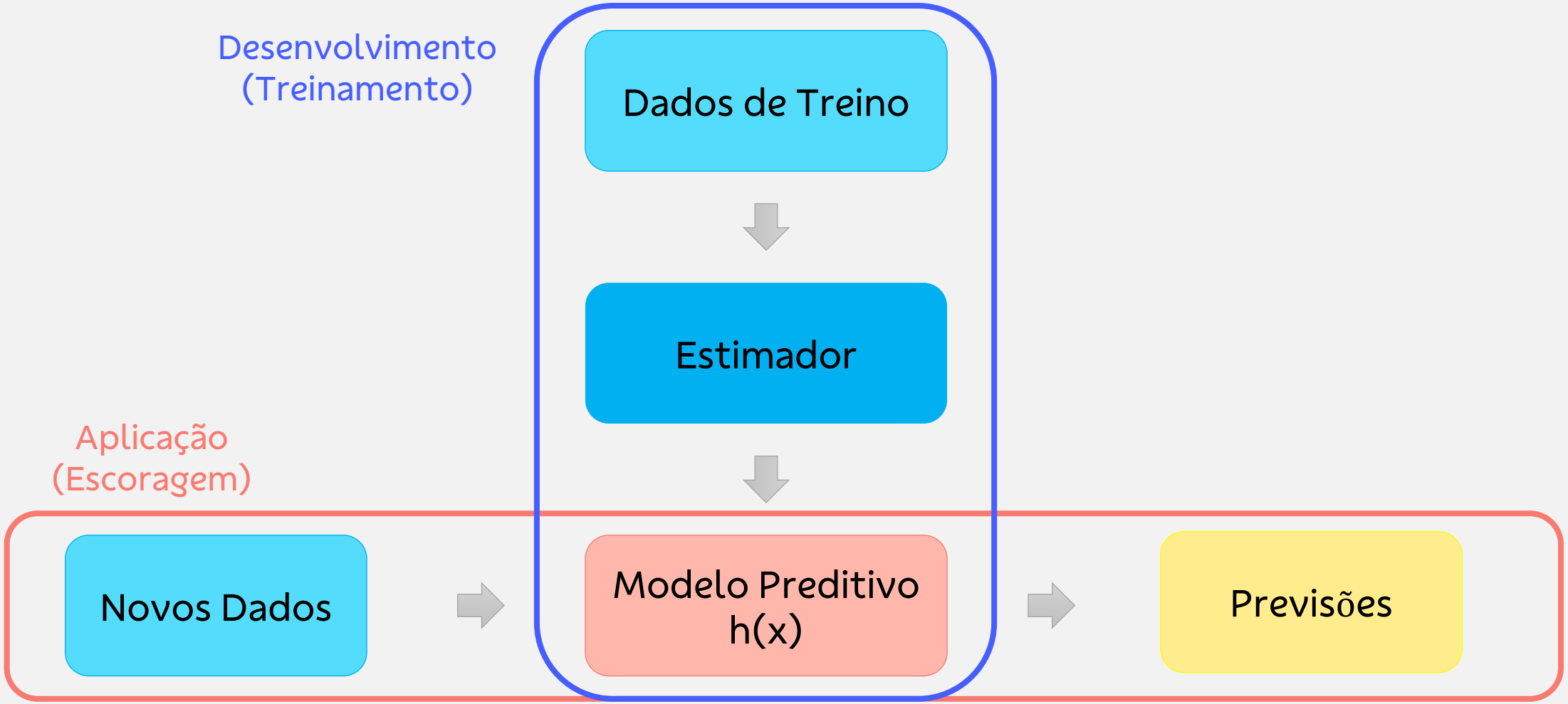
Fontes das imagens: https://assets.website-files.com/5e6f9b297ef3941db2593ba1/5f3a44b271f39266dd9861e0_zoYKG.png

Fonte da imagem: https://makeameme.org/meme/quando-a-magica


Entendimento do Negócio
Entendimento dos Dados
Preparação dos Dados
Modelagem
Avaliação
Aplicação
Dados

Aplicação

Uso do modelo para tomada de decisão



Expectativa x Realidade

Fonte da imagem: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQ8PoMALfEzaTwKtZPZG_GYJxgK07H4kuhSmg&usqp=CAU

Fonte da imagem: https://memegenerator.net/img/instances/58853736/s-teoriae-desconhece-a-situao-na-prtica-sabe-de-nadainocente.jpg

Indicações de Conteúdo

Comunidade Data Hackers

https://datahackers.com.br/

- Fórum (Slack)
- Podcasts (Google, Spotify)
- Blog (Medium)
- Newsletter semanal

Kaggle

https://www.kaggle.com/

- Bases de dados reais e fictícias
- Competições com premiações
- Compartilhamento de soluções entre usuários
- Cursos online gratuitos

Coursera

- Algoritmos de Machine Learning: https://www.coursera.org/learn/machine-learning

Pra tudo
que você
imaginar