

AULA 6: APRENDIZADO DE MÁQUINAS

INTRODUÇÃO A CIÊNCIA DE DADOS NA ENGENHARIA DE PETRÓLEO

Calendário

DATA	ATIVIDADE
26/08	Introdução
02/09	Tipos de dados/ Pré-processamento
09/09	Aula Prática 1
16/09	Aula Prática 2
23/09	Aula Prática 3
30/09	Introdução ML
07/10	ML Classificação
14/10	Aula Prática 4
21/10	ML Regressão/ML Agrupamento
28/10	Feriado
04/11	Aula Prática 5
11/11	Entrega dos Trabalhos

Tópicos

3

- Ciência dos dados: Etapas do Processo
- Modelo de dados: *Artificial Intelligence, Machine Learning e Deep Learning.*
- Tipos de Modelos de Machine Learning
 - ▣ Aprendizado Supervisionados
 - ▣ Aprendizado Não Supervisionados
 - ▣ Aprendizado por reforço
- Aprendizado
- Etapas da Modelagem de Dados
 - ▣ Série treino/teste
 - ▣ Bias e Variância
 - ▣ Validação Cruzada

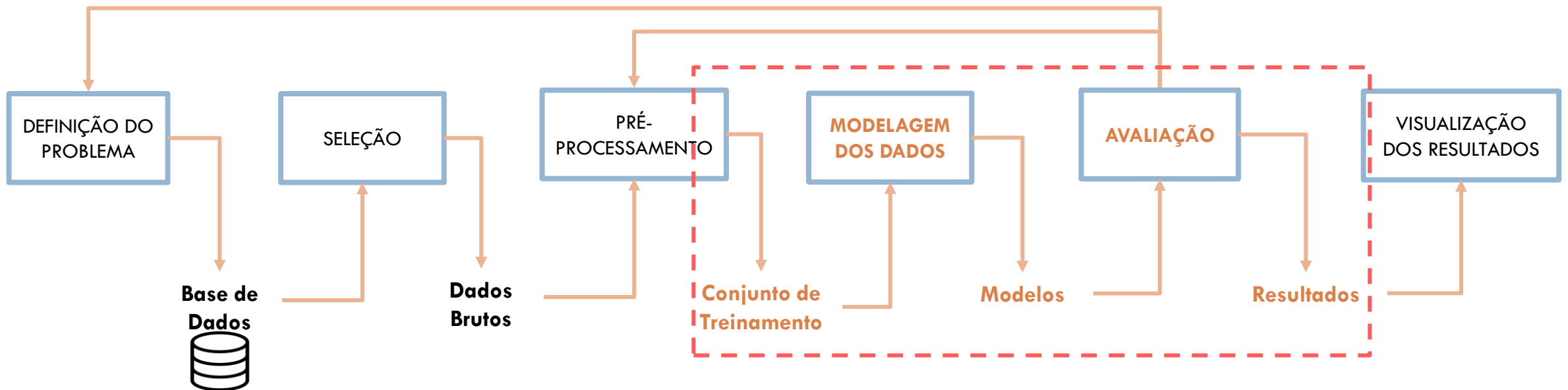
Ciência dos dados

Etapas do Processo

4

□ Revendo Etapas do Processo

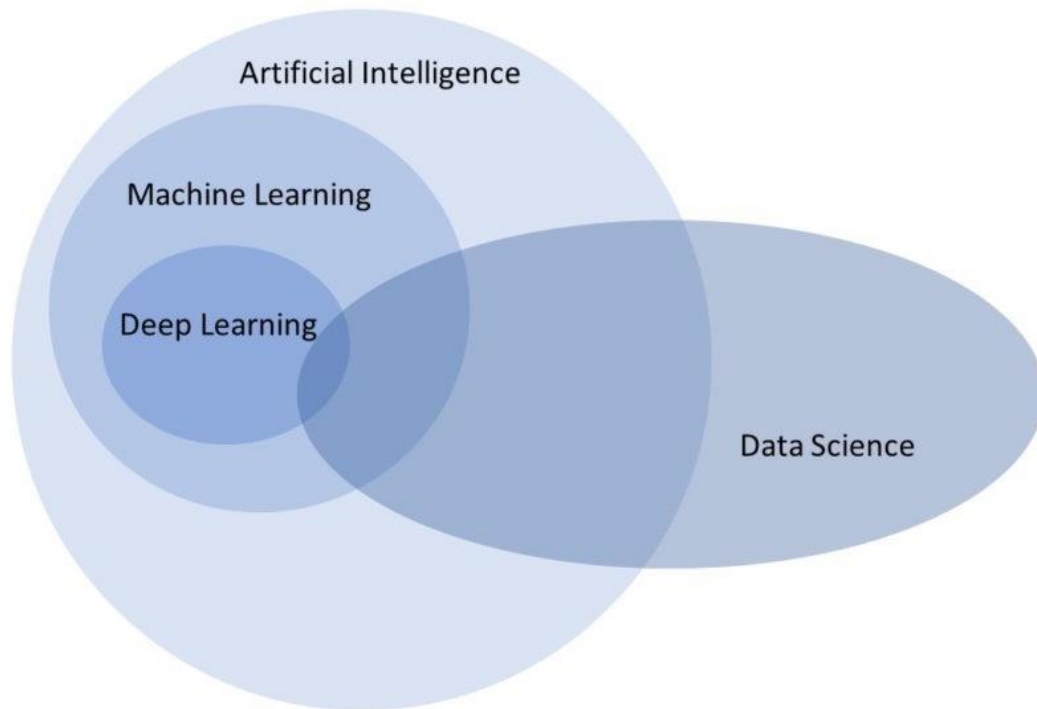
1. Definição do Problema.
2. Seleção do Conjunto de dados.
3. Análise Exploratória e limpeza dos dados no pré-processamento.
4. **Modelagem e Avaliação dos resultados.**



MODELO DE DADOS

Inteligência Artificial, Machine Learning e Deep Learning

5



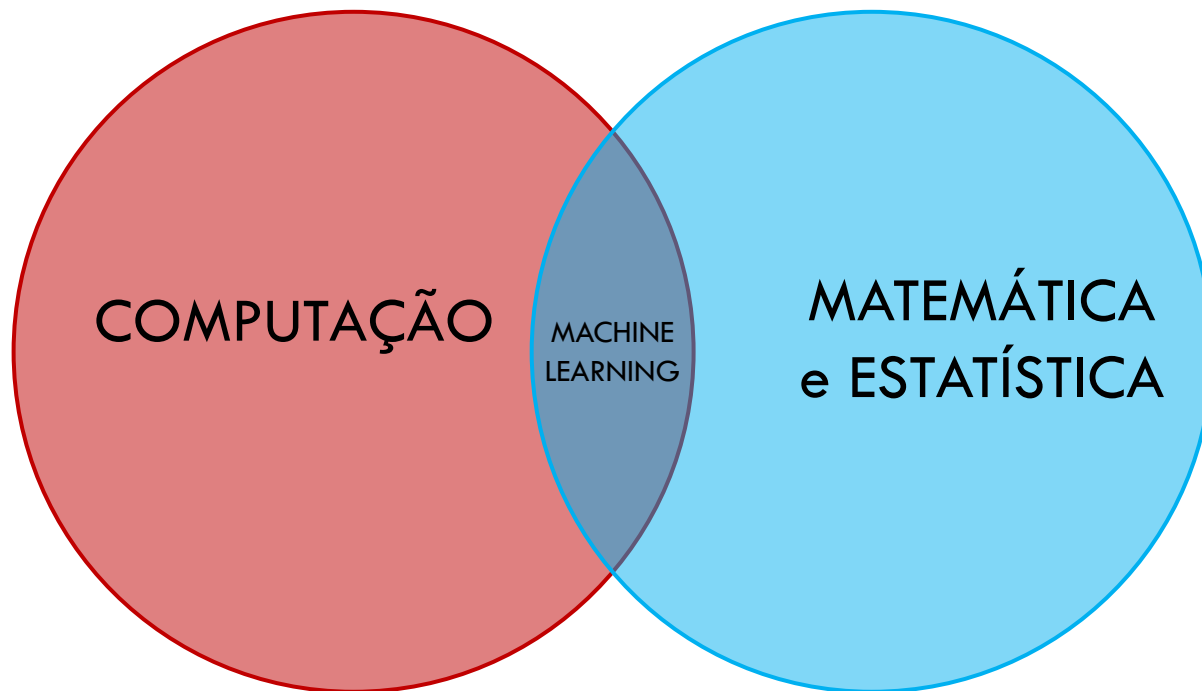
<https://blog.finxter.com/artificial-intelligence-machine-learning-deep-learning-and-data-science-whats-the-difference/>

- **Inteligência Artificial** (*Artificial Intelligence*):
 - ▣ Técnicas que capacitam máquinas a imitar a inteligência humana.
 - ▣ Lógicas, regras de associação, NLP, ML, DL.
- **Aprendizado de Máquinas** (*Machine Learning*):
 - ▣ Métodos estatísticos que permitem máquinas aprenderem a partir de dados de programação.
- **Aprendizado Profundo** (*Deep Learning*):
 - ▣ Tipo de ML que utilizam modelos mais complexos, de várias camadas para obtenção de resultados mais acurados.
 - ▣ Redes Neurais Profundas.

Aprendizado de Máquinas

(Machine Learning)

6



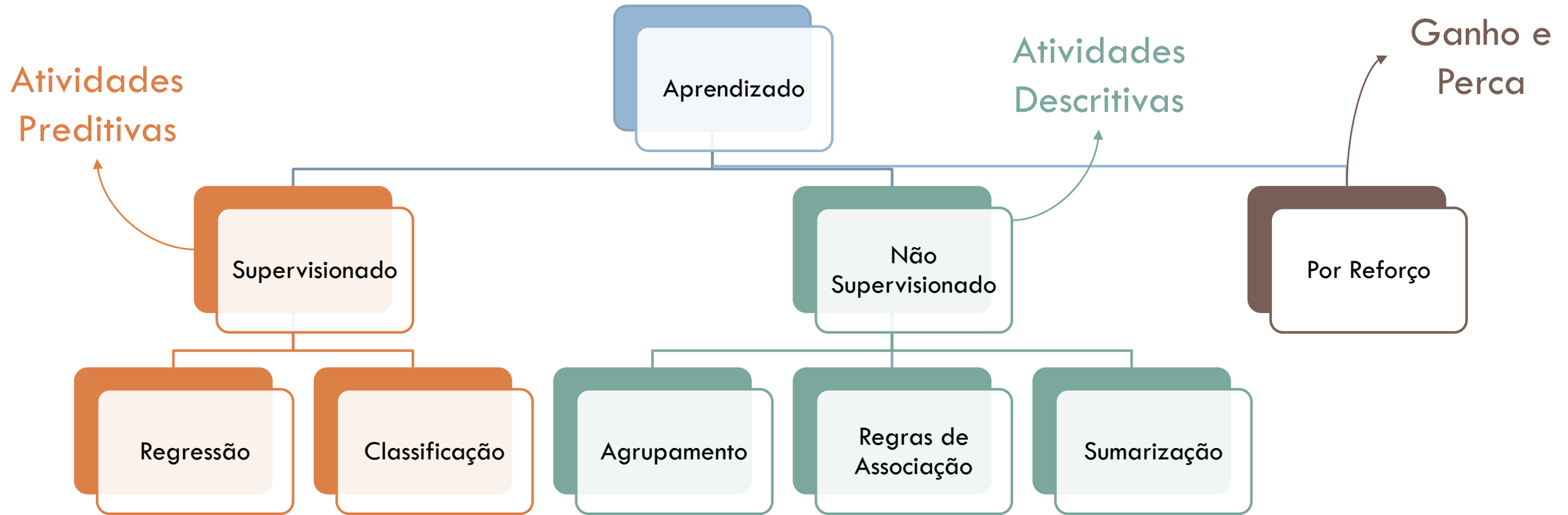
MACHINE LEARNING:

- ❑ Utiliza uma variedade de algoritmos que aprendem iterativamente a partir dos dados de treinamento para melhorar, descrever dados e prever resultados.
- ❑ Busca de padrões no conjunto de dados.

Aprendizado de Máquinas

Tipos de Modelos

7



Modelos Supervisionados

8

□ Regressão:

Idade	Tipo	Despesa
30	E	R\$ 500,00
50	C	R\$ 600,00
25	E	R\$ 200,00
20	V	R\$ 300,00
35	C	R\$ 500,00
20	E	R\$ 150,00
34	C	?

Treinamento

Novo Registro

X

Y

Variáveis Dependentes:
Contínuas

□ Classificação:

Idade	Tipo	Classe
30	E	Sim
50	C	Não
25	E	Sim
20	V	Sim
35	C	Não
20	E	Sim
34	C	?

Treinamento

Novo Registro

X

Y

Variáveis Dependentes:
Discretas

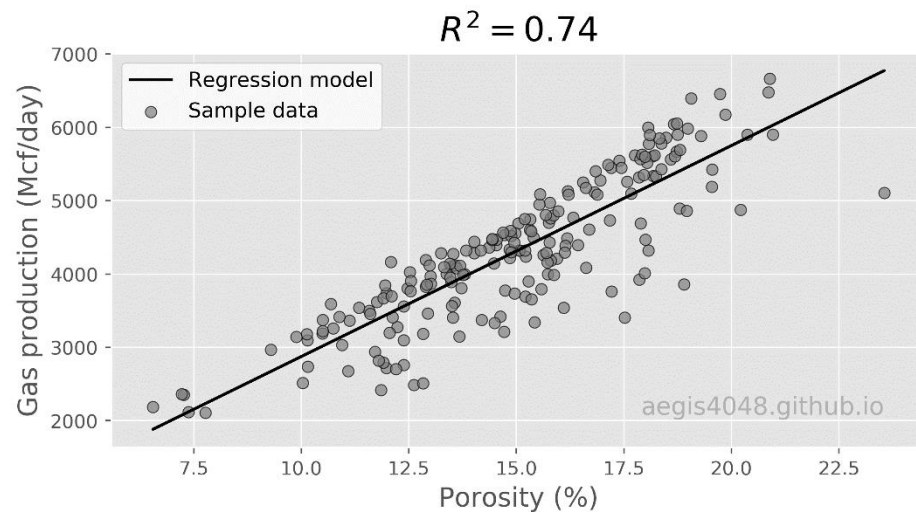
Modelos Supervisionados

Regressão

9

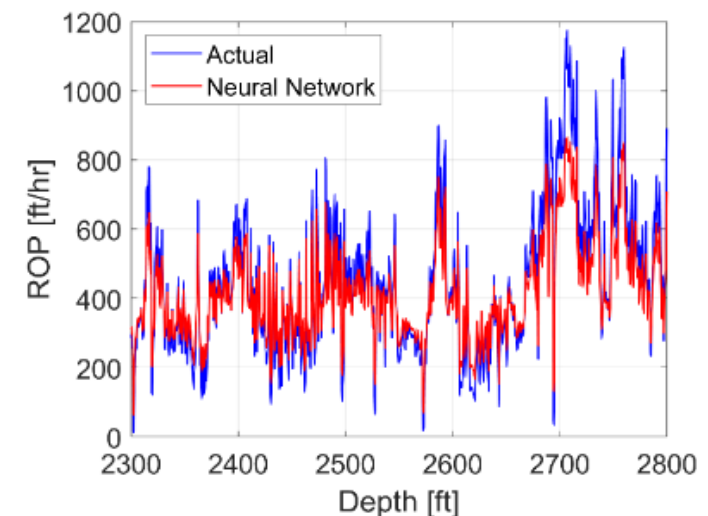
- Realizar estimativa do valor da variável de saída numérica a partir das variáveis de entrada.
- Avaliação da predição:
 - Função do erro de predição (Valor Predito - Valor Correto).

Modelos Lineares



https://aegis4048.github.io/multiple_linear_regression_and_visualization_in_python

Modelos não lineares



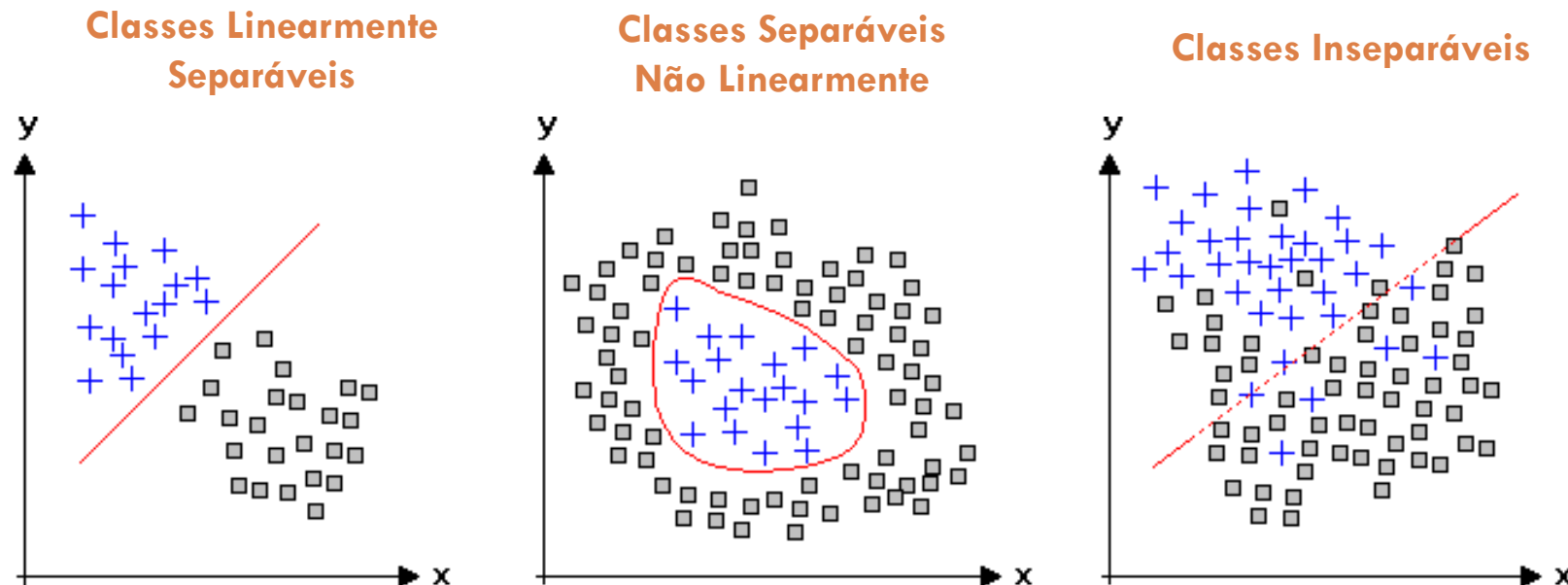
Chandrasekaran et al. (2020).

Modelos Supervisionados

Classificação

10

- Realizar estimativa do valor da variável de saída discreta a partir das variáveis de entrada.
- Avaliação da predição:
 - ▣ Precisão na predição da classe correta



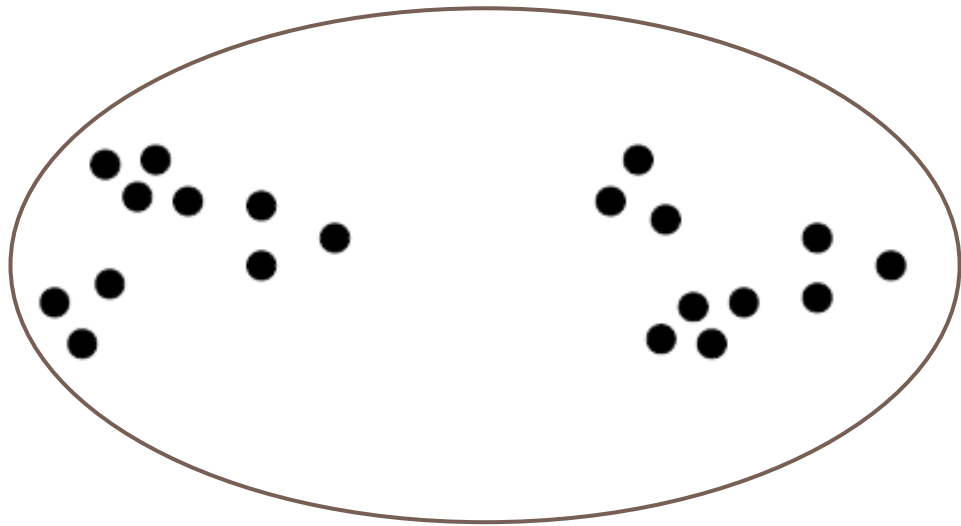
- Por melhor que seja o modelo, o erro de classificação não pode ser nulo.
- Estabelece um limite teórico de erro de um classificador

Modelos Não Supervisionados

Agrupamento

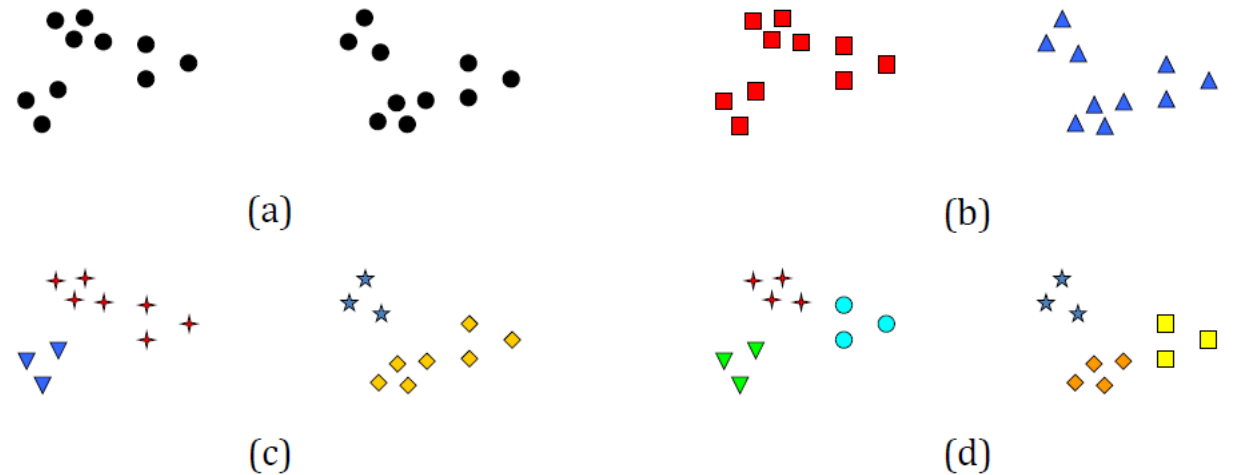
11

- Encontrar grupos de registros similares.



Quantos grupos
dividir?

Como avaliar a qualidade do
resultado do agrupamento?



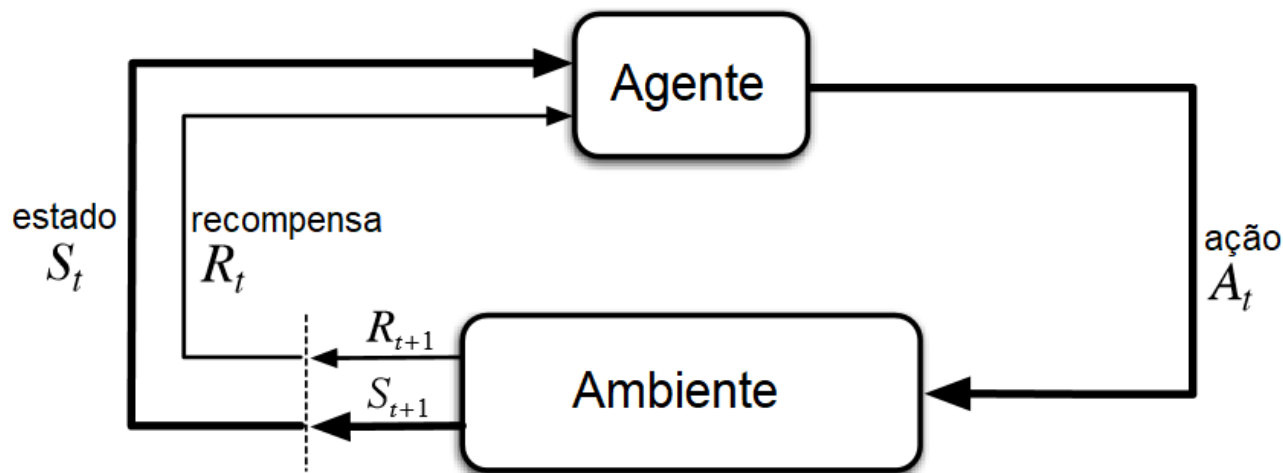
Evsukoff (2020)

- Dificuldade na definição de uma estatística de desempenho.

Modelos por Reforço

12

- Programa um agente que aprende sozinho a realizar uma tarefa com base em **tentativa e erro**, ou seja, *feedback* que recebe de suas ações.

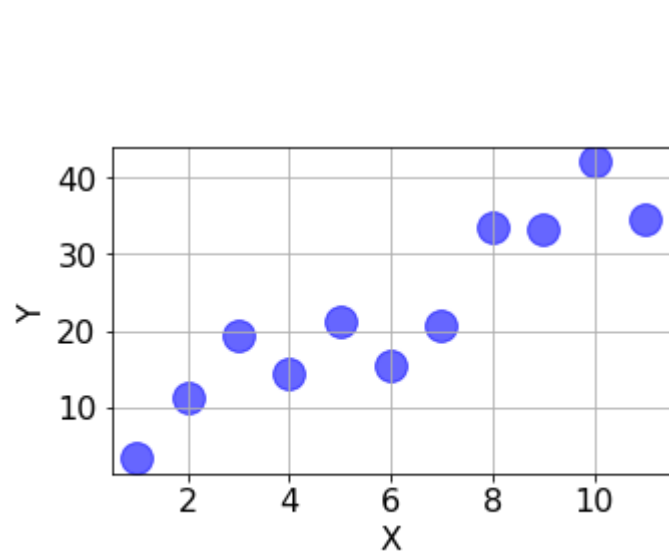


Maximizar a
Recompensa total

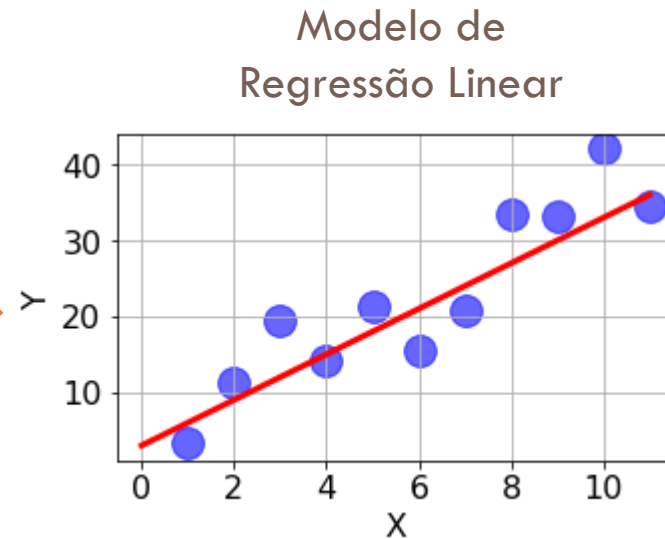
Aprendizado

13

□ Aprendizado = Representação + Avaliação + Otimização



Conjunto de registros
de treinamento



Representação:

$$h_{\theta} = \theta_1 x + \theta_0$$

Representação:
Seleção do tipo e da
estrutura do modelo.

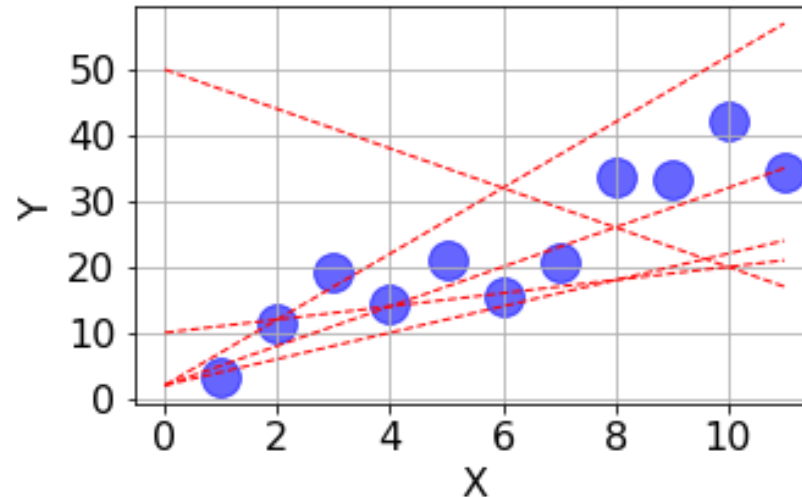
Aprendizado

14

□ Aprendizado = Representação + Avaliação + Otimização

Avaliação:

Medida para avaliar a
qualidade da instância.



Diferentes combinações
de θ_1 e θ_0

Como determinar
os parâmetros do
modelo?

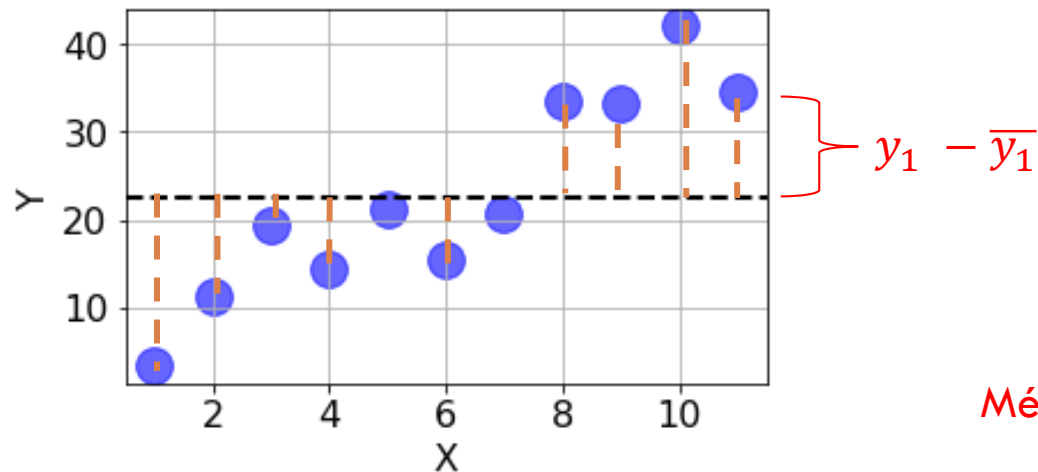
Aprendizado

Avaliação

15

□ Aprendizado = Representação + Avaliação + Otimização

Caso $\theta_1 = 0$ e $\theta_0 = 22,6$.



$$h_{\theta} = \theta_1 x + \theta_0$$

Soma do quadrado dos resíduos:

$$\text{SQR} = (y_1 - \theta_1 x_1 - \theta_0)^2 + (y_2 - \theta_1 x_2 - \theta_0)^2 + \dots + (y_n - \theta_1 x_n - \theta_0)^2$$

$$\text{SQR} = \sum_{i=1}^n (y_i - \theta_1 x_i - \theta_0)^2$$

Método dos mínimos
quadrados

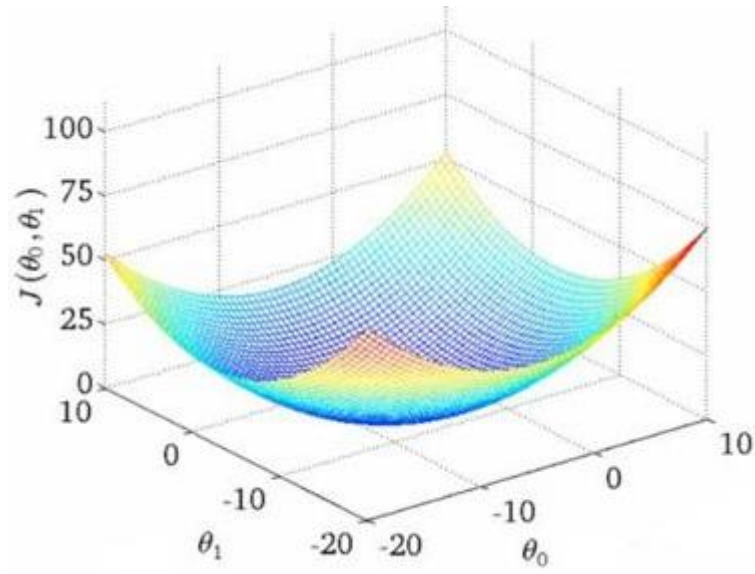
Métrica de Avaliação:

$$\begin{aligned} \text{Mínimo}(\text{SQR}) = \\ \text{Mínimo}(\sum_{i=1}^n (y_i - \theta_1 x_i - \theta_0)^2) \end{aligned}$$

Aprendizado

16

□ Aprendizado = Representação + Avaliação + Otimização



<https://slideplayer.com/slide/15834222/>

Otimização:

Método que irá encontrar o melhor conjunto de parâmetros.

Como iremos encontrar a resposta da equação?

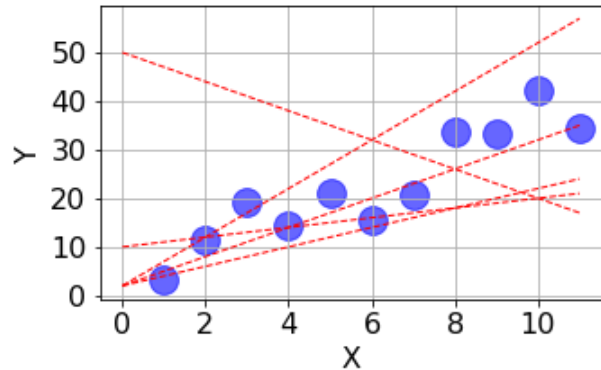
$$\text{Mínimo}(\sum_{i=1}^n (y_i - \theta_1 x_i - \theta_0)^2)$$

Aprendizado

17

□ Aprendizado = Representação + Avaliação + Otimização

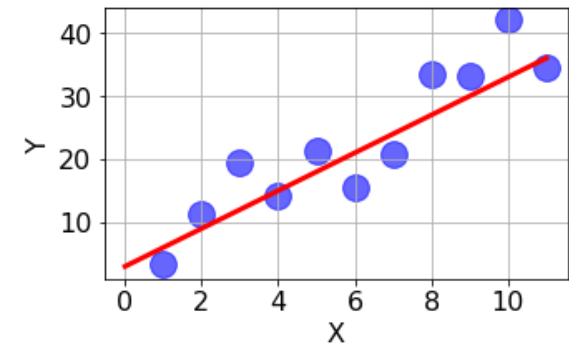
Regressão Linear



Diferentes parâmetros de ajustes

Métrica de Avaliação:
Mínimo(SQR)

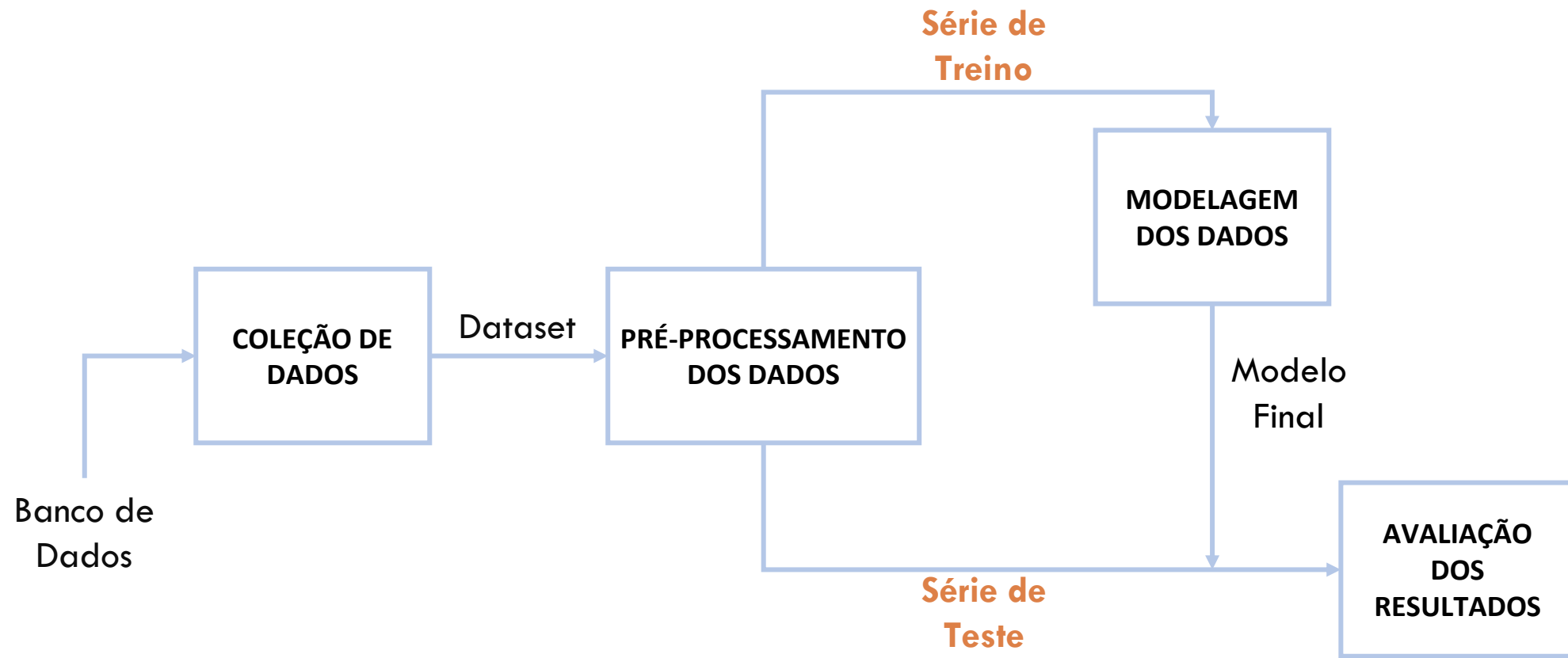
Método de
Otimização



Melhor reta com o
mínimo valor da soma
do quadrado dos
resíduos

ETAPAS DA MODELAGEM DOS DADOS

18

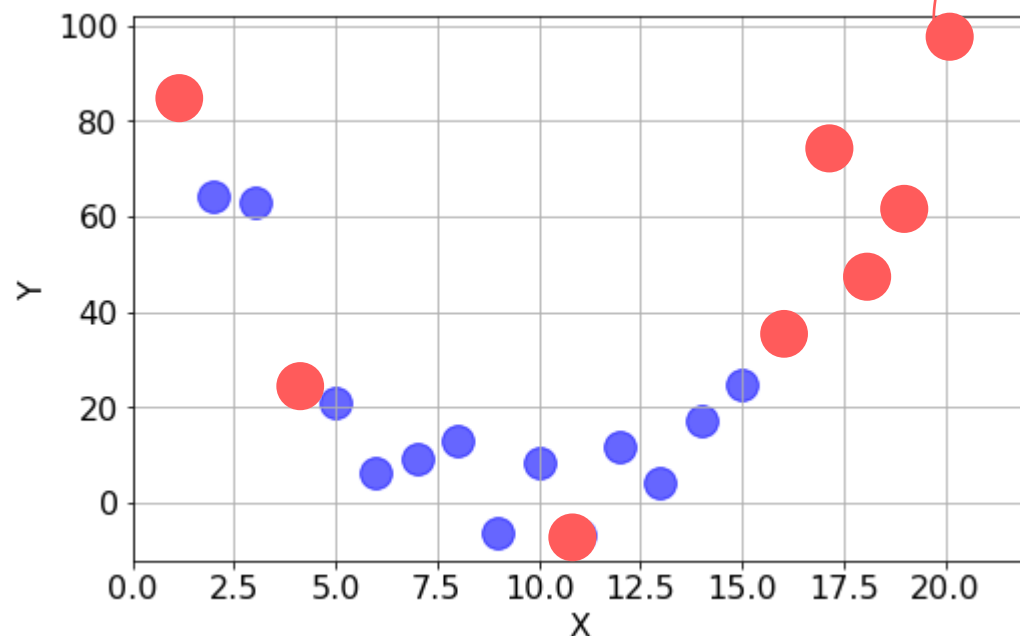


ETAPAS DA MODELAGEM DOS DADOS

1. Divisão em Série de Treino e Teste

19

Conjunto de Dados

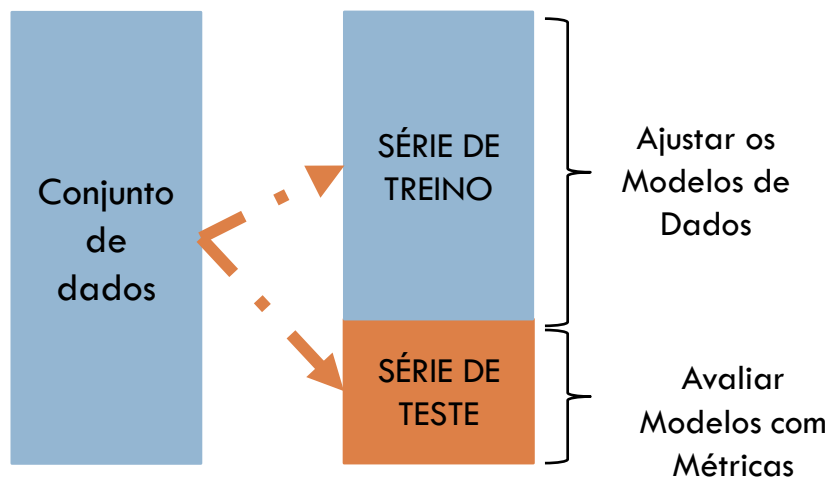


Série de Teste: Conjunto de dados para ser avaliado no final do processo.

ETAPAS DA MODELAGEM DOS DADOS

1. Divisão em Série de Treino e Teste

20



Idade	Tipo	Classe
30	E	Sim
50	C	Não
25	E	Sim
20	V	Sim
35	C	Não
20	E	Sim
34	C	?

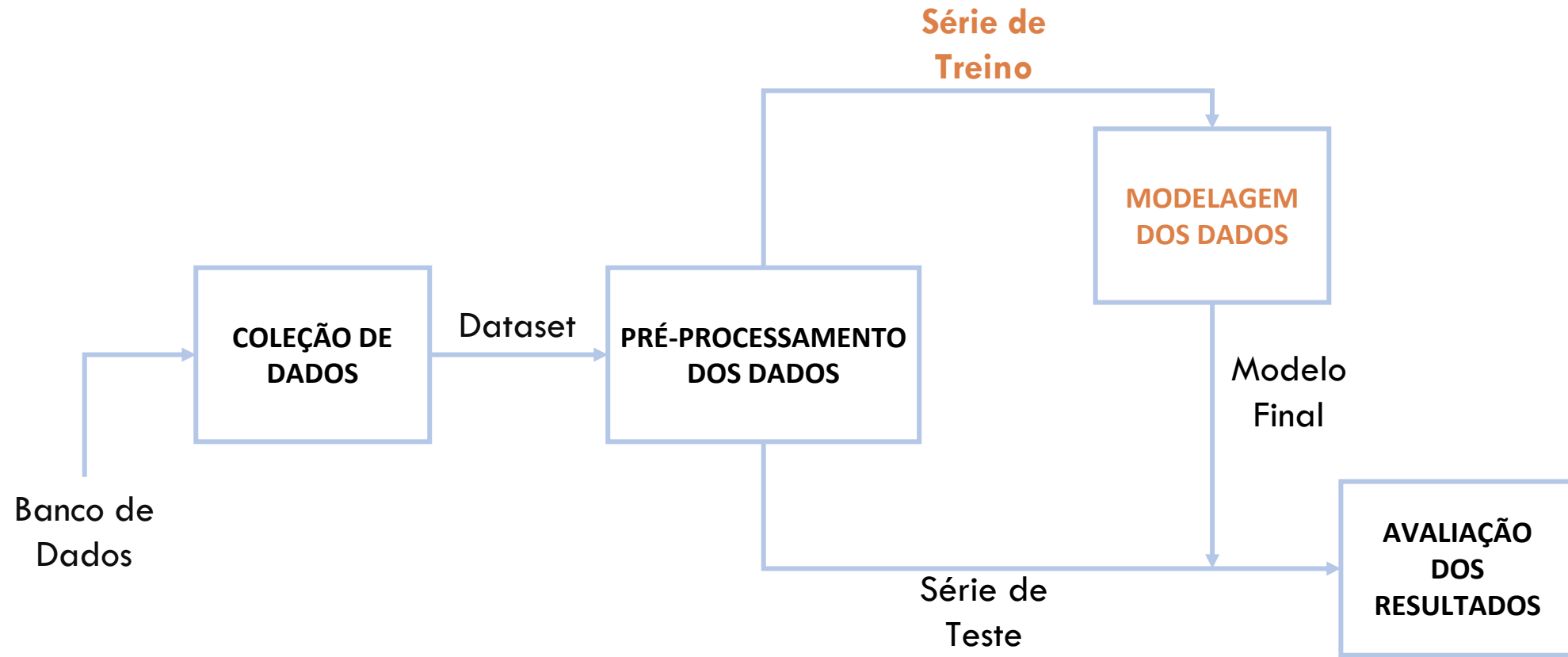
Ex2: Esta divisão é boa?

Idade	Tipo	Classe
30	E	Sim
50	C	Não
25	E	Sim
20	V	Sim
35	C	Não
20	E	Sim
34	C	?

- Em problemas de classificação, atentar para ter as **diferentes classes** no treino e no teste.

ETAPAS DA MODELAGEM DOS DADOS

21

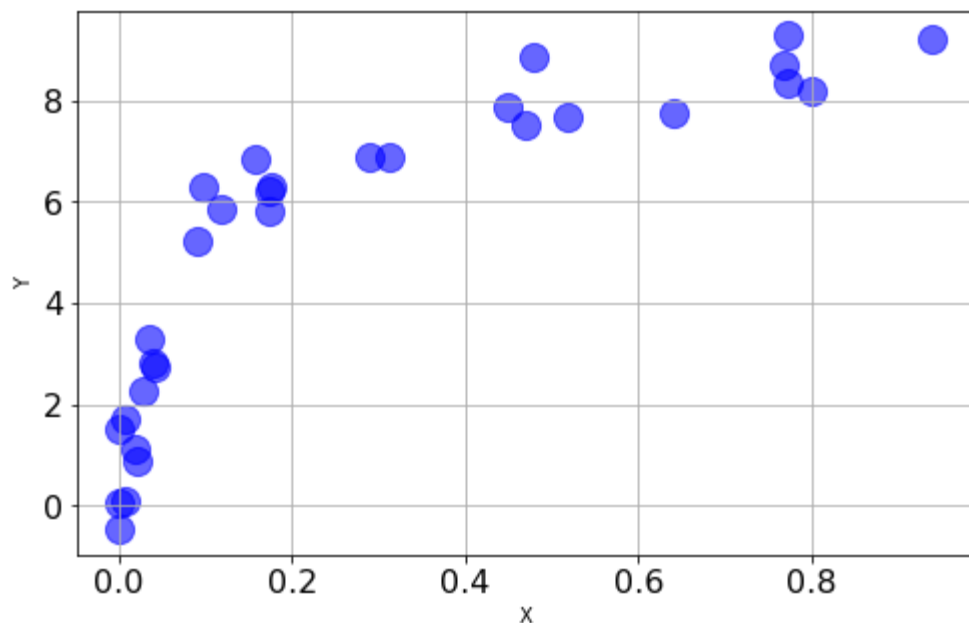


ETAPAS DA MODELAGEM DOS DADOS

2. Modelagem dos Dados

22

□ Como escolher o melhor modelo?



Conjunto de Treino

- Testar diferentes modelos.
- Ajuste dos parâmetros do modelo.
- Avaliação dos resultados por métricas.

Modelo: Quanto mais próximo possível da relação real, melhor será o desempenho do modelo para predição de novos dados.

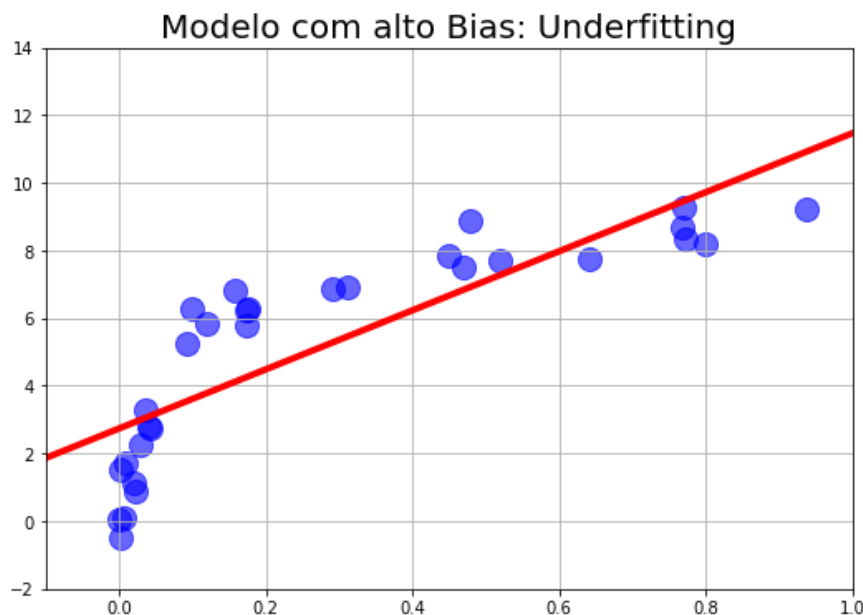
ETAPAS DA MODELAGEM DOS DADOS

Bias x Variância

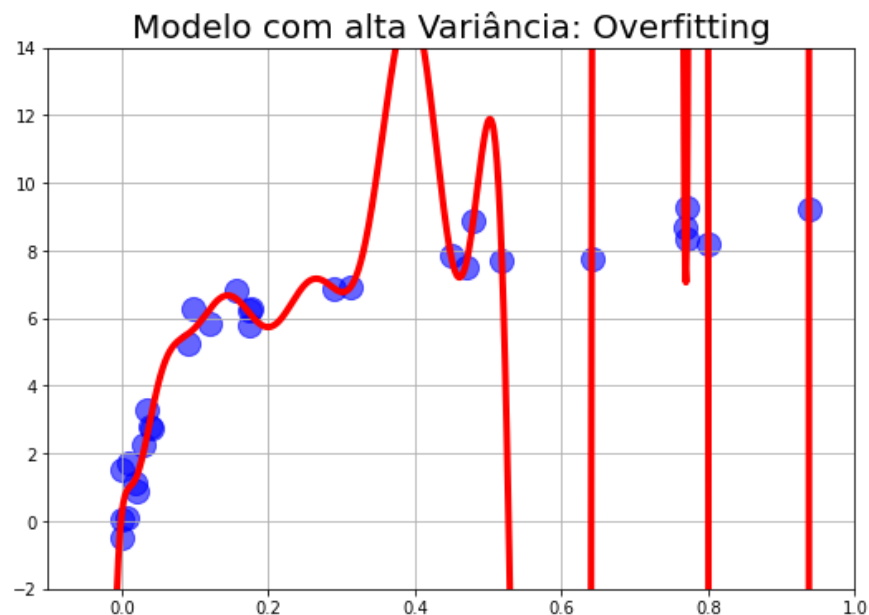
23

□ Como escolher o melhor modelo?

Conjunto de Treino



Regressão Linear



Polinômio de Grau 20

Qual o melhor modelo?

Nenhum!

Primeiro : sub-ajuste do comportamento real dos dados.

Segundo: Super ajuste do comportamento real dos dados.

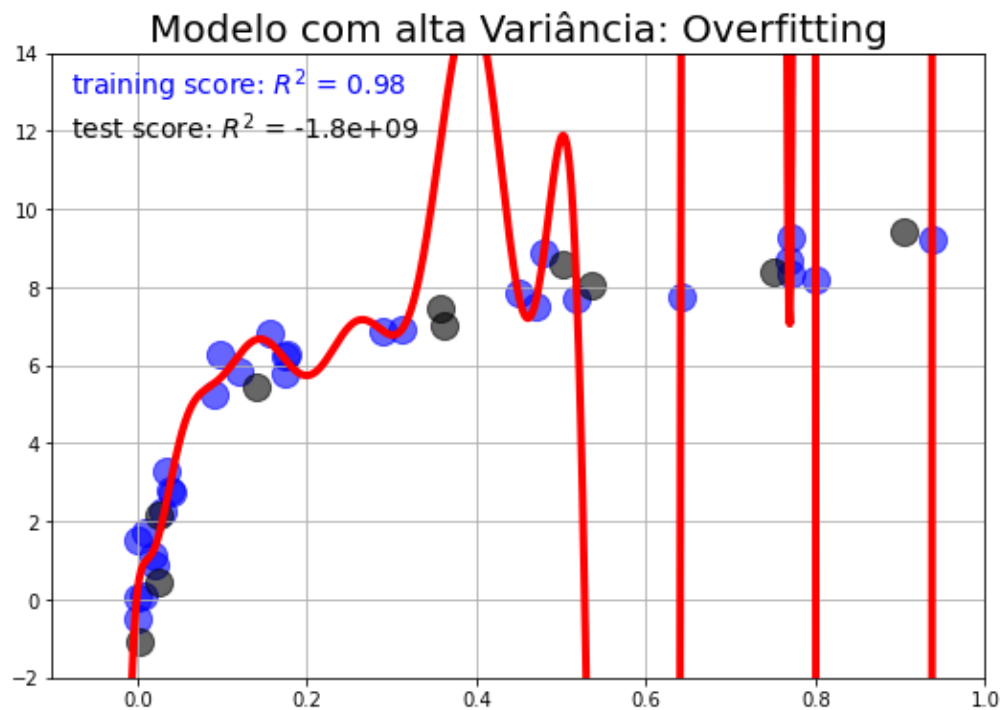
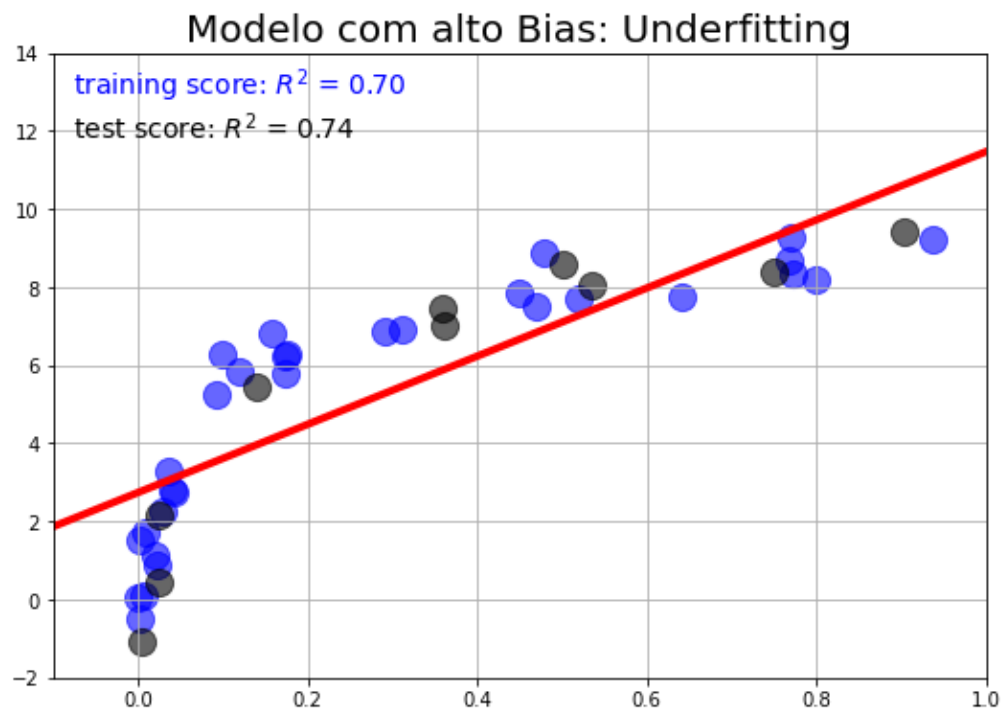
ETAPAS DA MODELAGEM DOS DADOS

Bias x Variância

24

□ Como escolher o melhor modelo?

Conjunto de Treino e Teste



ETAPAS DA MODELAGEM DOS DADOS

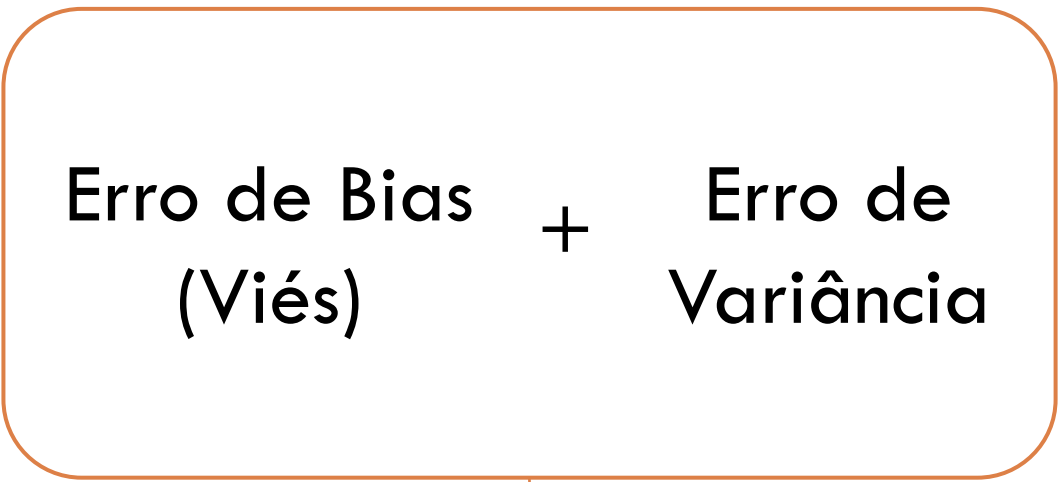
Bias x Variância

25

- Como escolher o melhor modelo?

$$\text{Erro de Previsão} = \text{Erro de Bias (Viés)} + \text{Erro de Variância} + \text{Erro Irredutível (ruído)}$$

Definem um bom modelo de dados

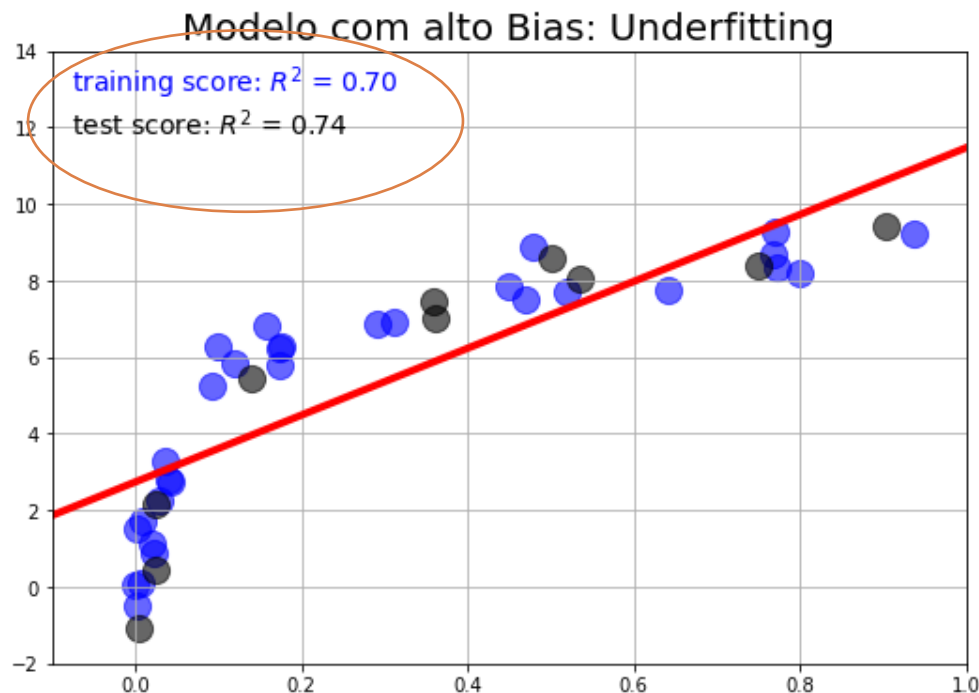


ETAPAS DA MODELAGEM DOS DADOS

Bias x Variância

26

□ Como escolher o melhor modelo?



○ Erro de viés:

- Algoritmo não é capaz de expressar o fenômeno.
- Mesmo com amostras diferentes ou maiores, nunca irá se aproximar do valor real, pois o modelo não é capaz.

**ALTO BIAS
BAIXA VARIÂNCIA**



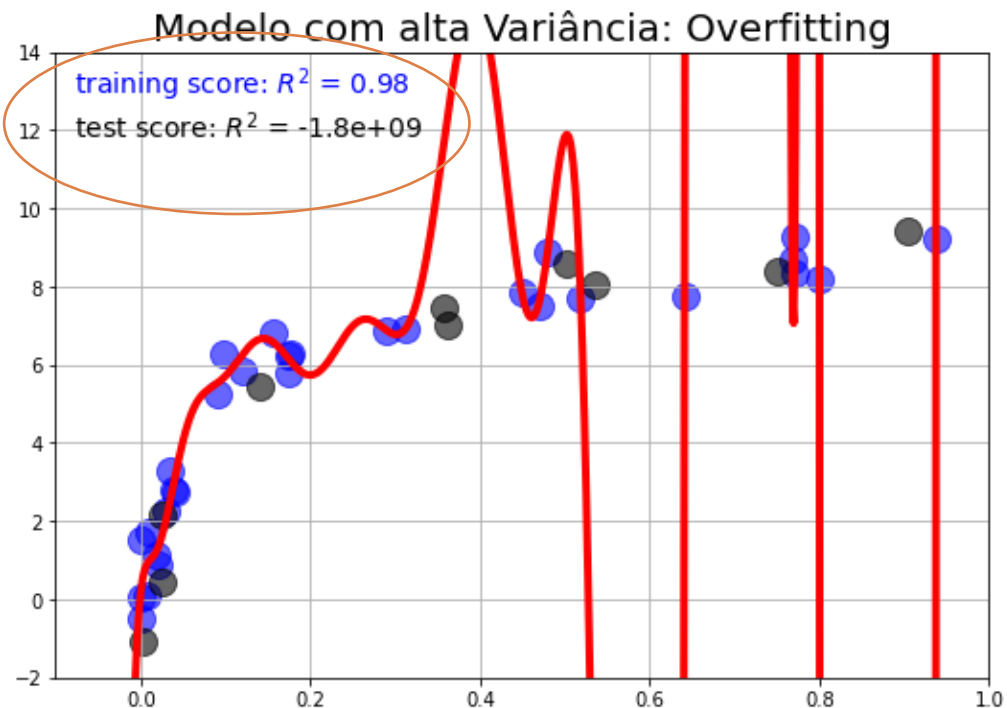
UNDERFITTING

ETAPAS DA MODELAGEM DOS DADOS

Bias x Variância

27

□ Como escolher o melhor modelo?



○ Erro de variância:

- Erro de sensibilidade.
- Não tem como garantir que a estimativa será boa no conjunto de teste.
- Modelo tem complexidade maior que a complexidade da relação.

**BAIXO BIAS
ALTA VARIÂNCIA**



OVERFITTING

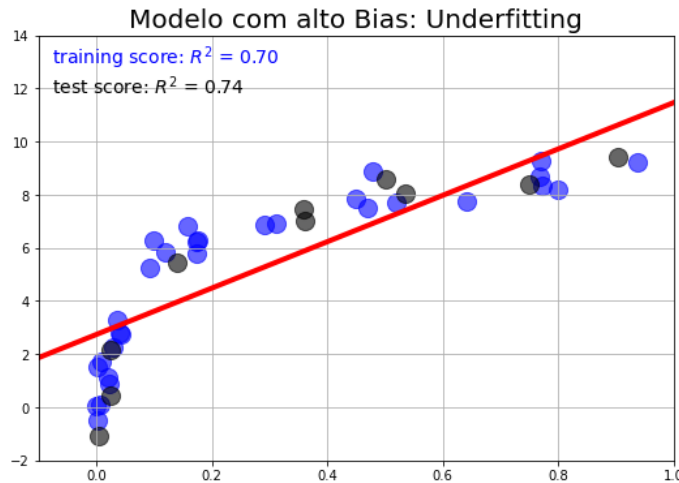
ETAPAS DA MODELAGEM DOS DADOS

Bias x Variância

28

□ Como escolher o melhor modelo?

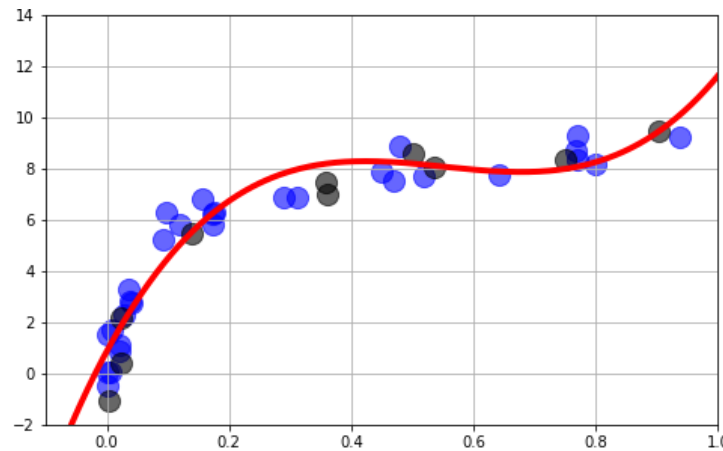
Underfitting



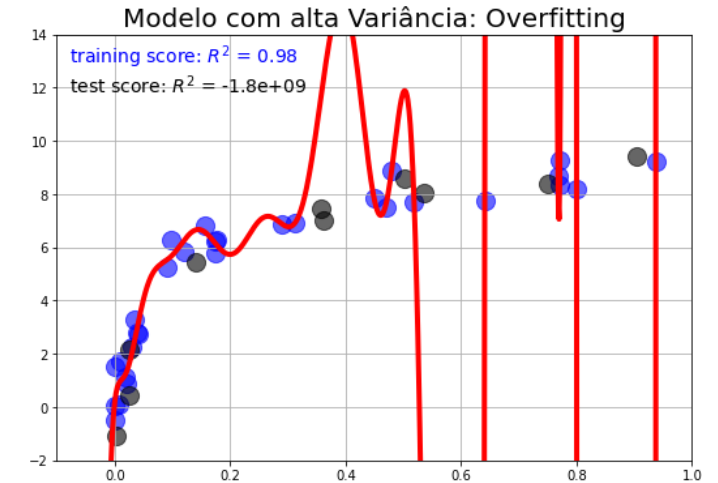
Underfitting. O que fazer?

- Aumentar número de variáveis.
- Aumentar a complexidade dos modelo.
- Diminuir a regularização.

Bom Modelo



Overfitting



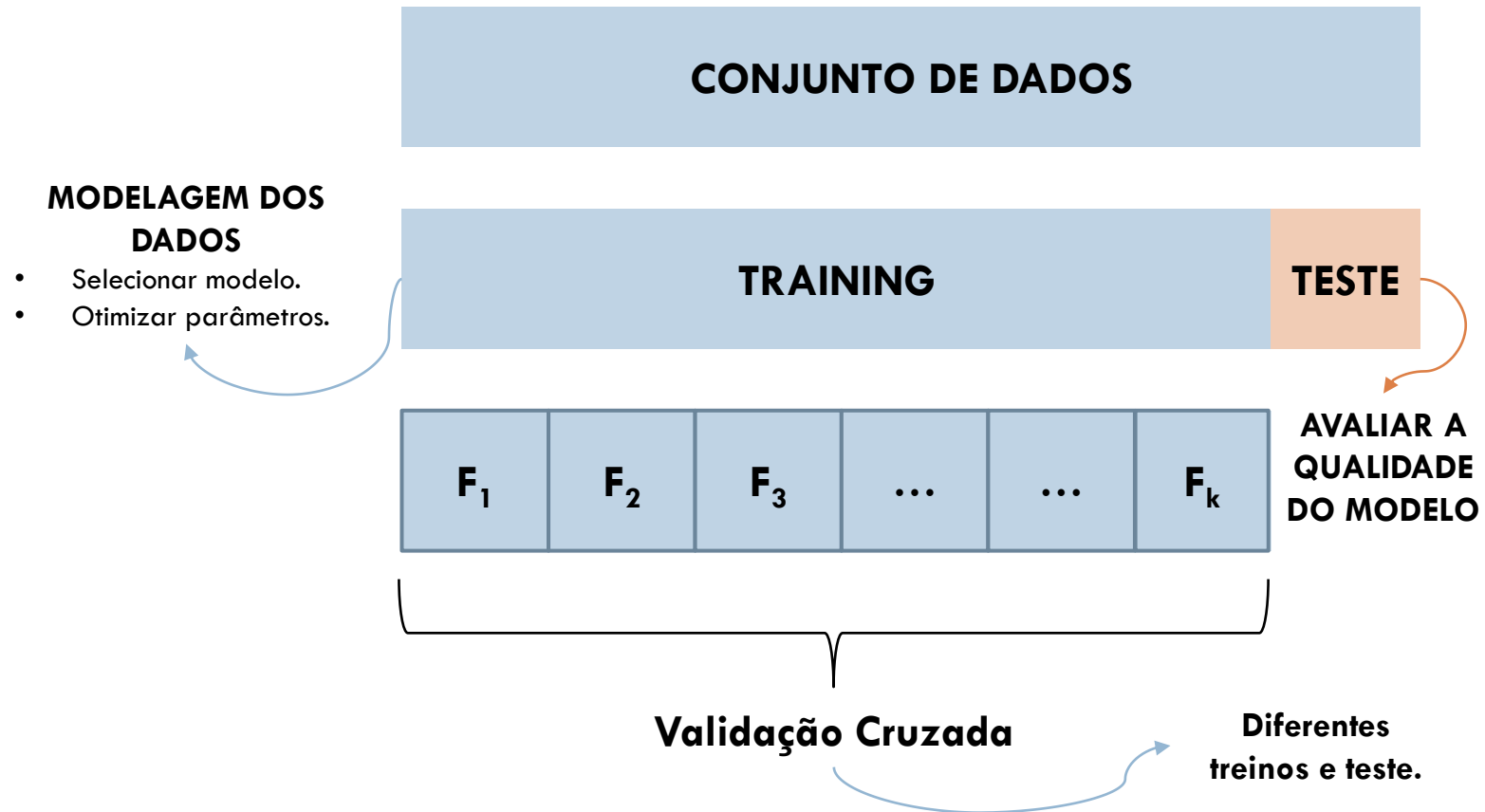
Overfitting. O que fazer?

- Aumentar número de observações.
- Diminuir número de variáveis (menor complexidade).
- Aumentar regularização.

VALIDAÇÃO CRUZADA

29

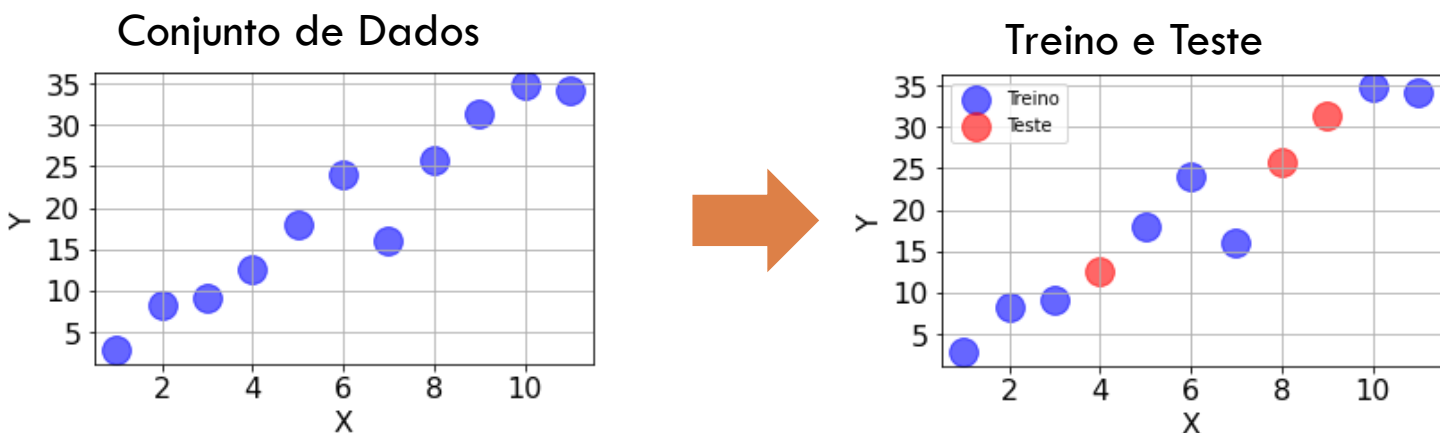
- Técnica que consegue controlar o erro de variância.
- Garante diferentes treinos/testes, de forma a avaliar diferentes partes do conjunto de dados.



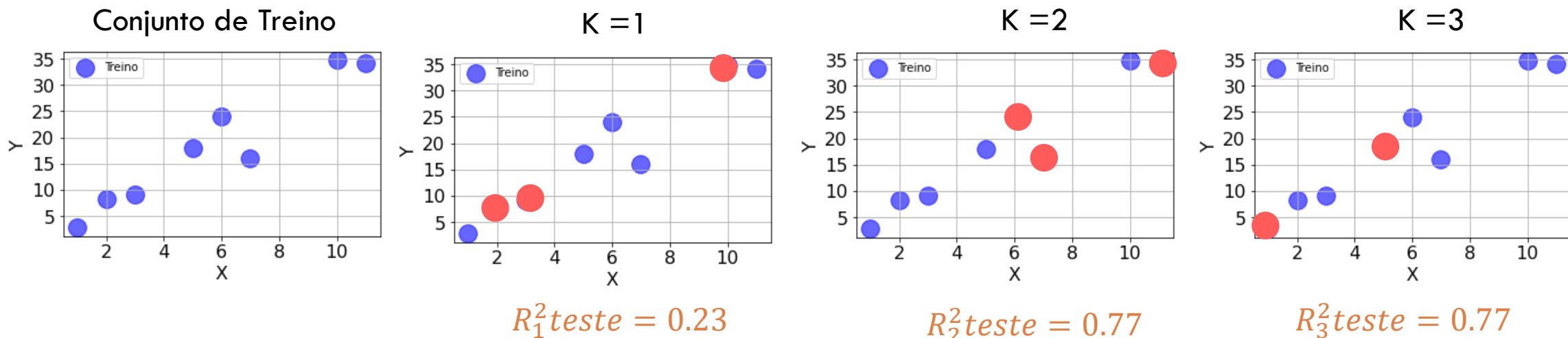
VALIDAÇÃO CRUZADA

30

Parte 1 Divisão Treino e Teste



Parte 2 Validação Cruzada k = 3



Referências Bibliográficas

- Chandrasekaran, Sridharan & Govindarajan, Suresh Kumar. (2020). Drilling Efficiency Improvement and Rate of Penetration Optimization by Machine Learning and Data Analytics. International Journal of Mathematical, Engineering and Management Sciences.
- Evsukoff, A G. **INTELIGÊNCIA COMPUTACIONAL Fundamentos e aplicações**. 2020.
- Li, H., Yu, H., Cao, N. *et al.* Applications of Artificial Intelligence in Oil and Gas Development. *Arch Computat Methods Eng* (2020). <https://doi.org/10.1007/s11831-020-09402-8>.
- VanderPlas, J. **Python Data Science Handbook**. 2016.