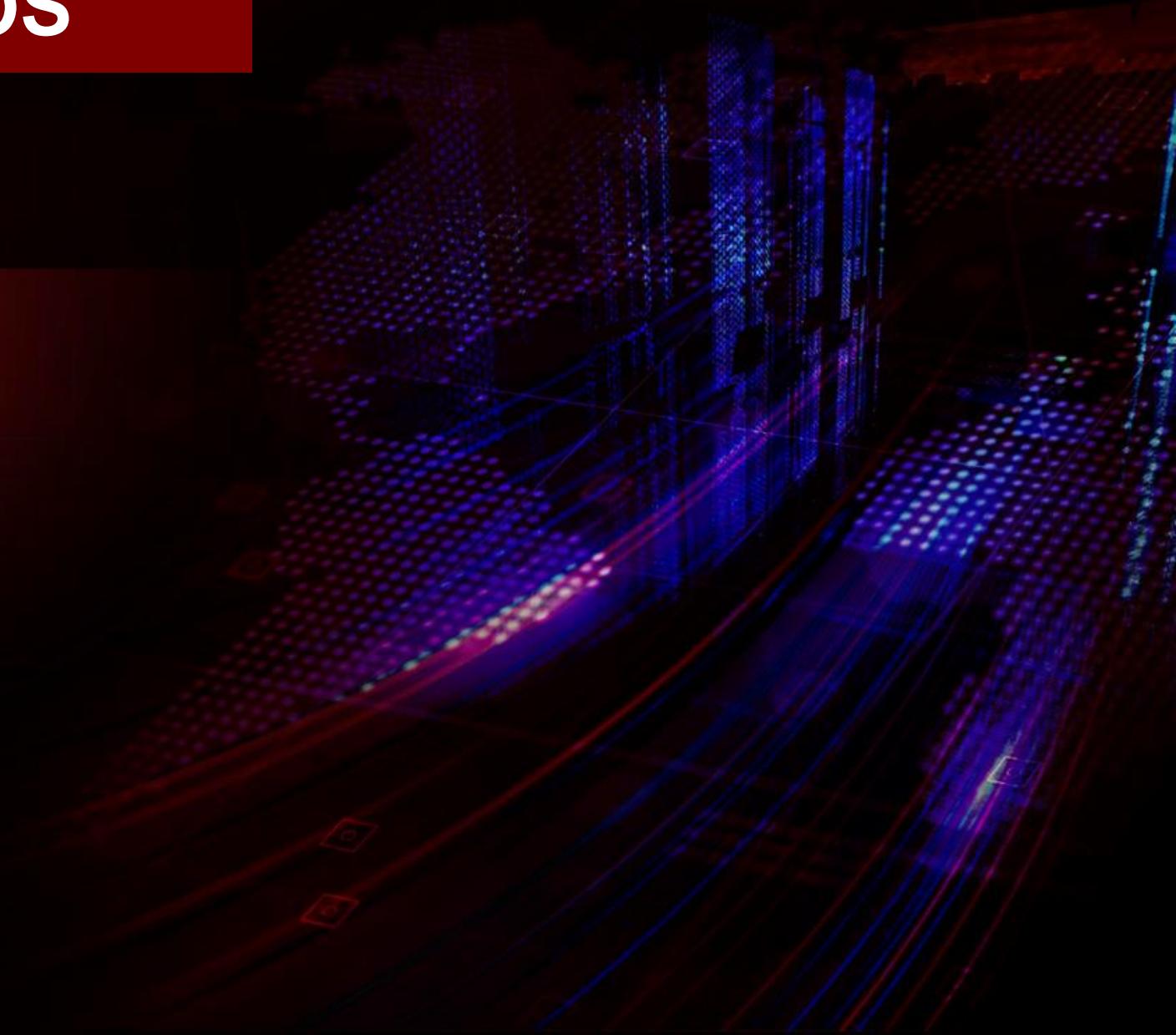


MINERAÇÃO DE DADOS

Classificação de dados



APRENDIZADO SUPERVISIONADO

Registros históricos com resultados

Cadastro de clientes:

- Idade
- Renda
- Tipo de financiamento

Resultado: Adimplente / Inadimplente



APRENDIZADO SUPERVISIONADO

Resultado



Atributo alvo

Rótulo de classe

Valor de saída

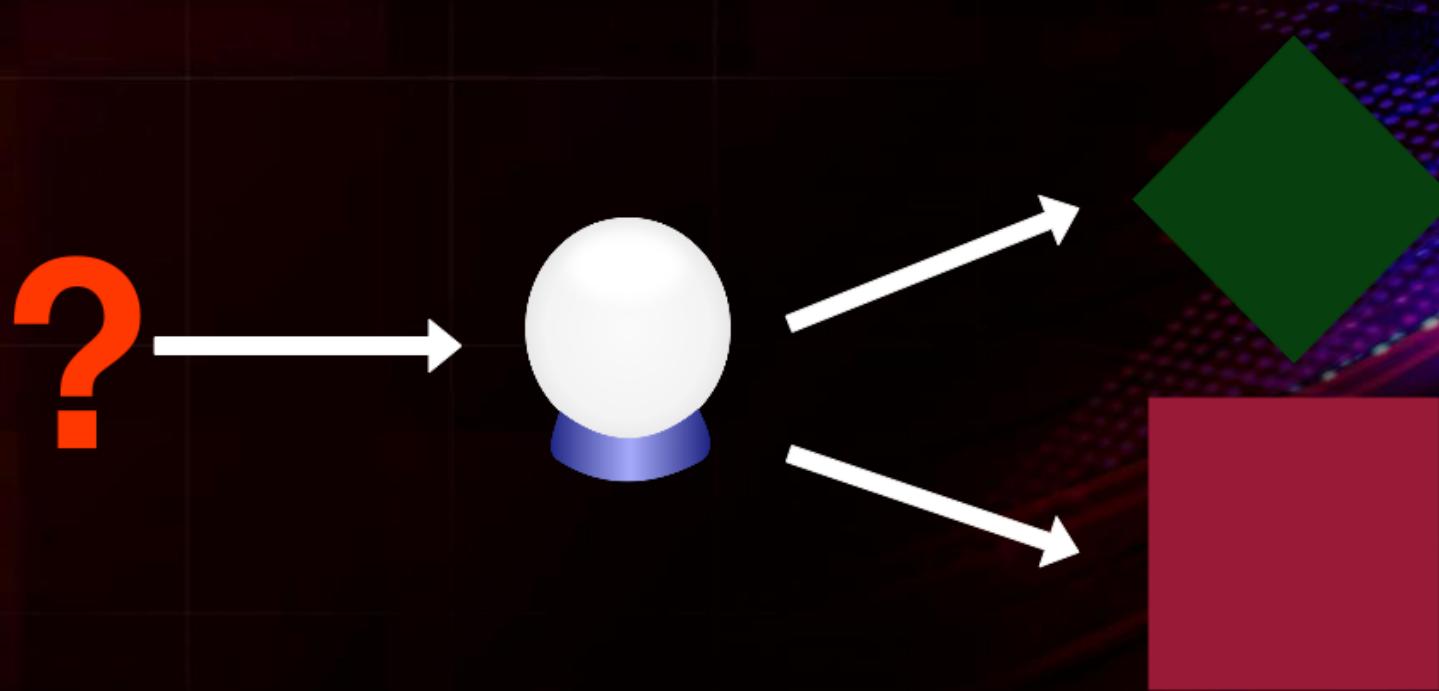
Podemos usar valores de saída conhecidos para identificar novos registros?



PREDIÇÃO

Ato ou efeito de predizer ou de afirmar o que se acredita que vai acontecer no futuro.

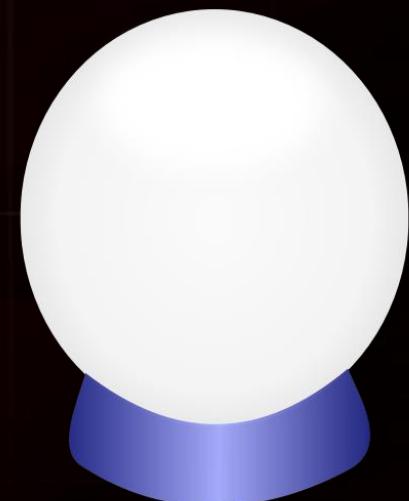
- Michaelis



MODELO PREDITIVO

Indica qual é a saída para novos registros

Construídos a partir de algoritmos de mineração



TIPOS DE PREDIÇÃO

- Classificação (discreta)

Ex: definir se um cliente pode receber crédito

- Estimação (contínua)

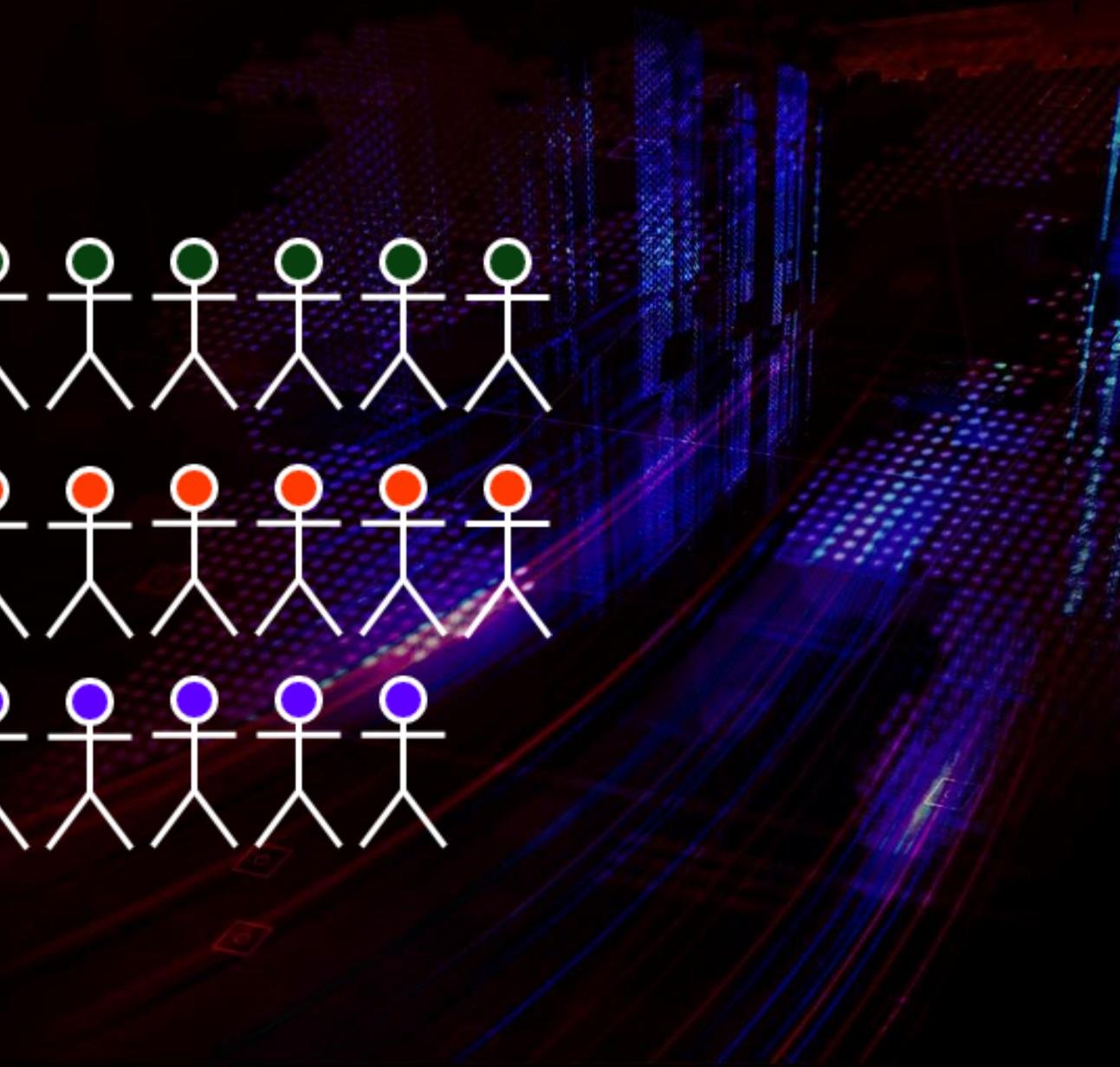
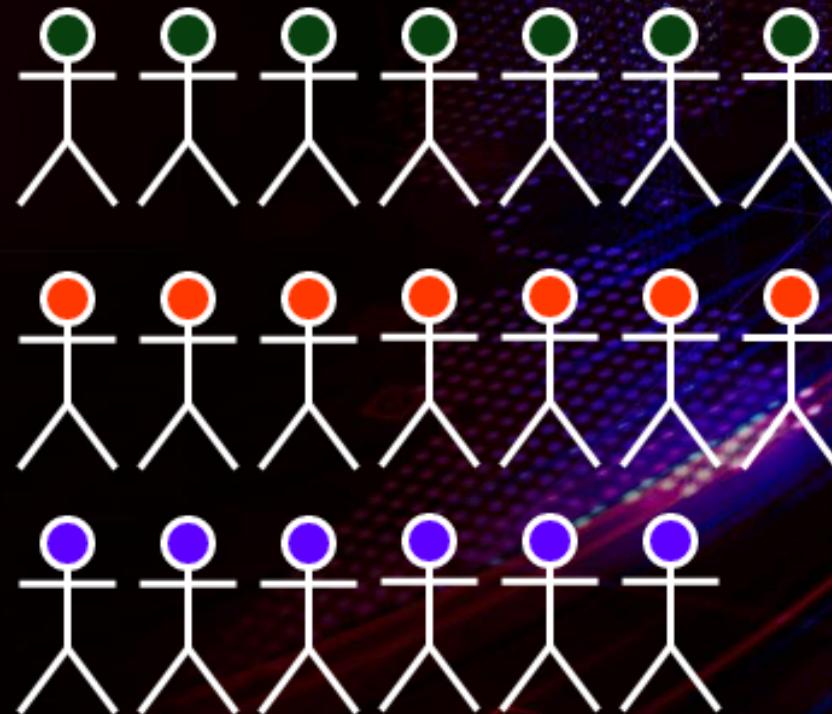
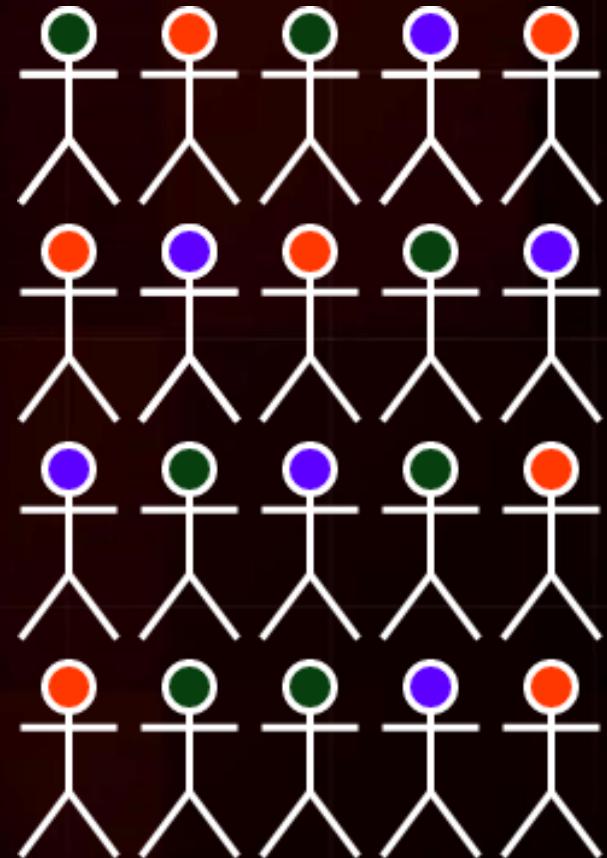
Ex: definir o valor do crédito concedido

ESTIMAÇÃO

- Regressão linear / polinomial
- Redes neurais artificiais



CLASSIFICAÇÃO



ETAPAS DE CONSTRUÇÃO DO MODELO

➤ Treinamento:

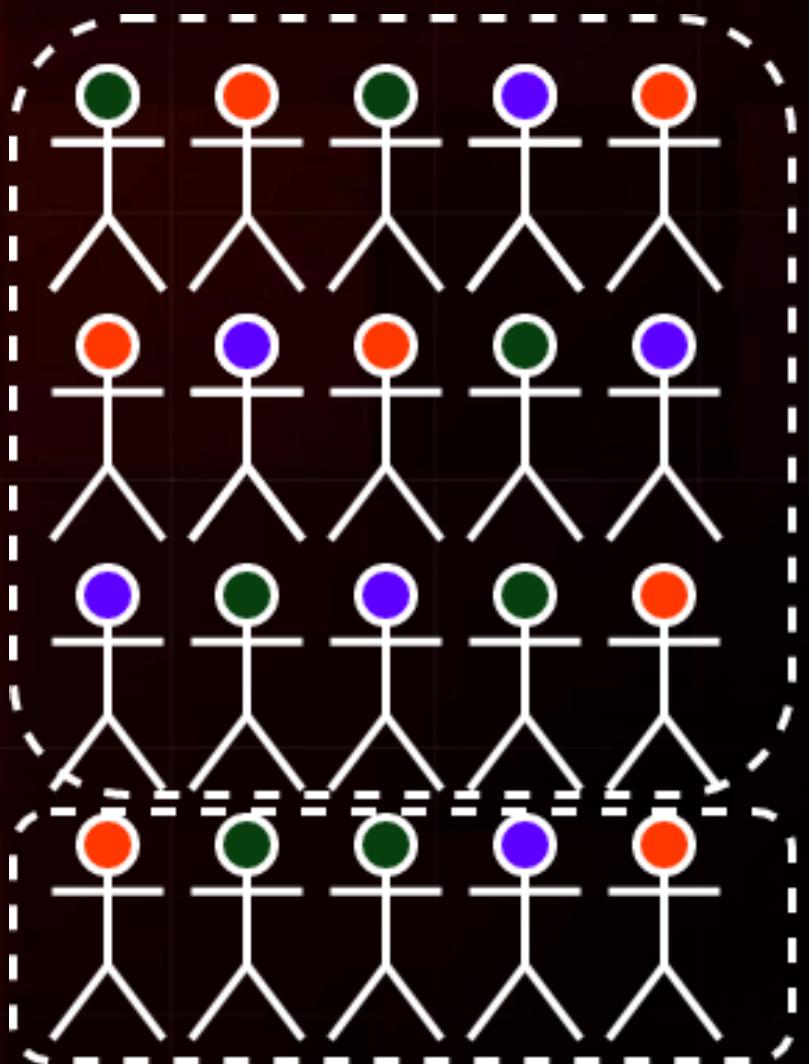
- Geração do modelo
- Uso de dados rotulados

➤ Teste:

- Avaliação do modelo
- Dados não usados no treinamento



ETAPAS DE CONSTRUÇÃO DO MODELO



Treinamento

Teste

ERROS

➤ **Erro de representação (efeito *bias*):**

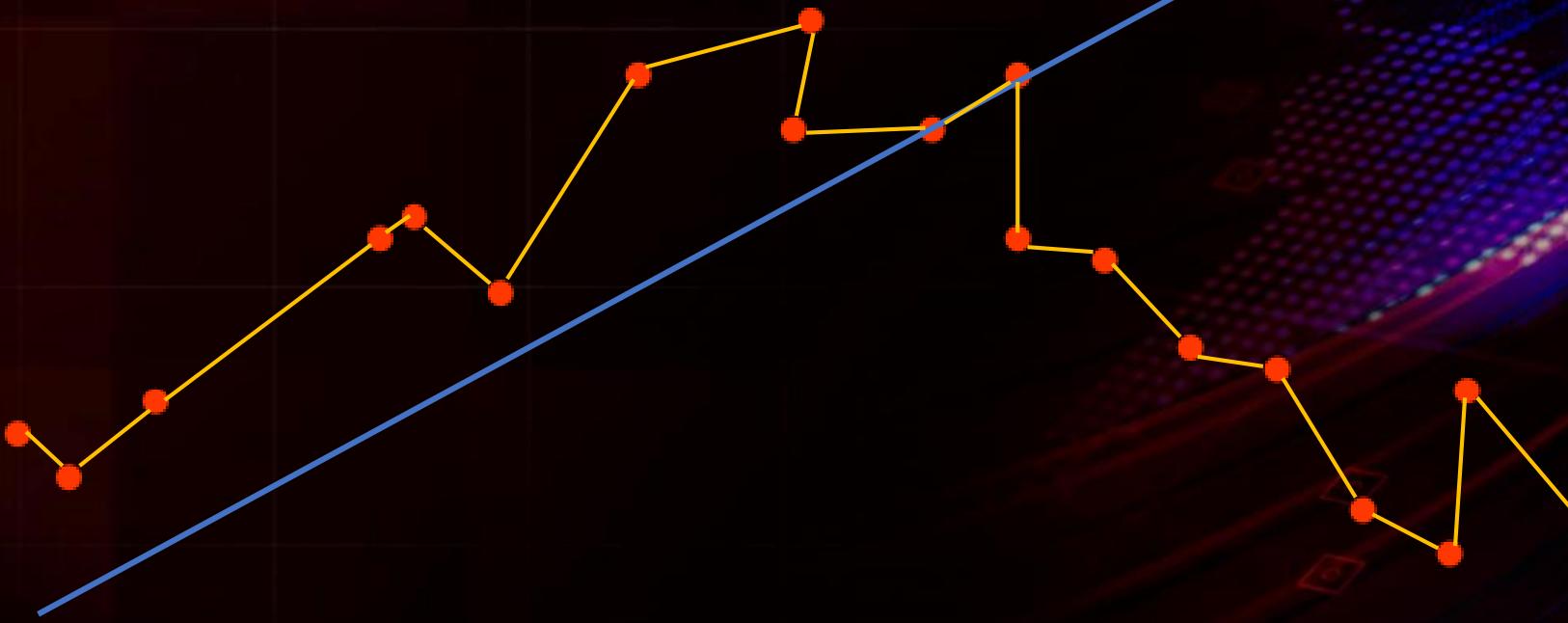
- Dados completos disponíveis
- Inadequação do modelo

➤ **Erro de generalização (variância):**

- Apenas uma amostra dos dados disponível
- Sobregeneralização

ERROS

- Dados de treinamento
- Subajustado (bias)
- Sobreajustado (variância)
- Bom ajuste bias-variância



DILEMA BIAS-VARIÂNCIA

➤ Validação cruzada

➤ Critério de parada

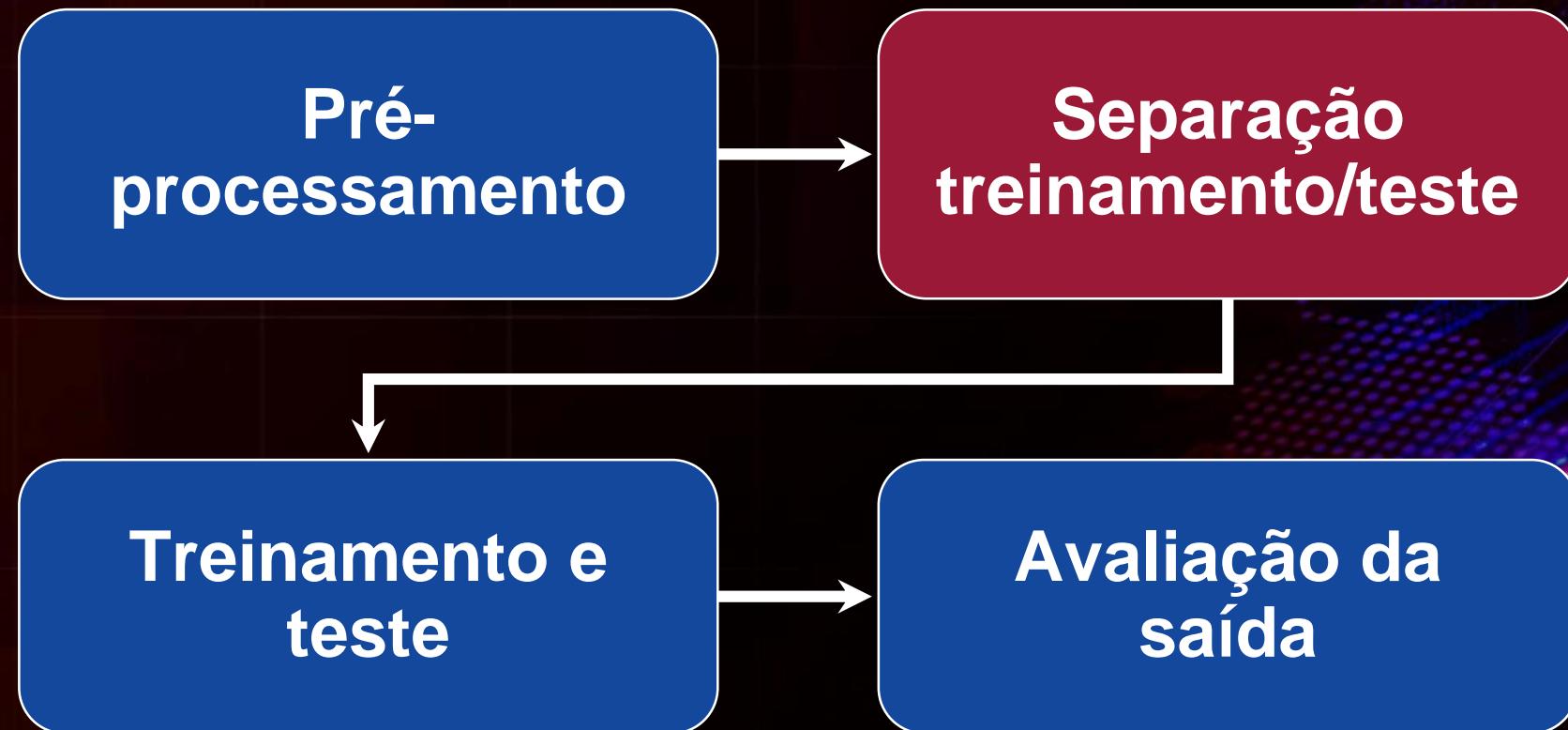
Erro do modelo começa a aumentar consecutivamente



PROCESSO DE PREDIÇÃO DE DADOS



PROCESSO DE PREDIÇÃO DE DADOS



SEPARAÇÃO TREINAMENTO / TESTE

- Critério personalizado (*ad-hoc*)
- Validação cruzada



VALIDAÇÃO CRUZADA

- Particionamento sistemático
- Reduzir a variabilidade dos resultados
- Todos os dados são usados para treinamento e teste
- Validação cruzada em k-pastas

VALIDAÇÃO CRUZADA K-PASTAS

➤ k = 10 pastas

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

...

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

VALIDAÇÃO CRUZADA K-PASTAS

➤ $k = 10$ pastas

1 vez = 10 iterações

10 vezes = 100 iterações

➤ $k = n = n$ iterações (leave-one-out)

n é o número de objetos da base

PROCESSO DE PREDIÇÃO DE DADOS



TREINAMENTO E TESTE

Uso de diferentes algoritmos

Ajustes nos parâmetros do algoritmo

Execução



PROCESSO DE PREDIÇÃO DE DADOS



AVALIAÇÃO DA SAÍDA

- Verificação de acertos e erros
 - Medidas de desempenho
- Acertos e erros podem ter pesos distintos
- Número de classes:
 - Binárias / multiclassses

MATRIZ DE CONFUSÃO

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

- VP: Verdadeiros Positivos
- VN: Verdadeiros Negativos
- FP: Falsos Positivos
- FN: Falsos Negativos

MATRIZ DE CONFUSÃO

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

Ex: classificação de e-mails

- VP: spam classificado como spam
- VN: mensagem normal classificada como normal
- FP: mensagem normal classificada como spam
- FN: spam classificado como mensagem normal

MEDIDAS DE DESEMPENHO

Taxa de verdadeiros positivos (TVP) $\frac{VP}{VP + FN}$

Taxa de falsos positivos (TFP) $\frac{FP}{FP + VN}$

MEDIDAS DE DESEMPENHO

Acurácia (ACC)

% global de objetos corretos

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

Erro (E)

% global de objetos errados

$$E = 1 - ACC$$

MEDIDAS DE DESEMPENHO

Relevância

Precisão (Pr)

% objetos recuperados relevantes

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

Revocação (Re)

% recuperação de objetos relevantes

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

Medida-F (F-score)

Acurácia de precisão e revogação

$$\frac{2 \times Pr \times Re}{(Pr + Re)}$$

REFERÊNCIAS

Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações: Cap. 5: Classificação de dados.

Leandro Nunes de Castro e Daniel Gomes Ferrari. Editora Saraiva, 2016.