Curso 2 – CD, AM e DM

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligência Artificial e BigData

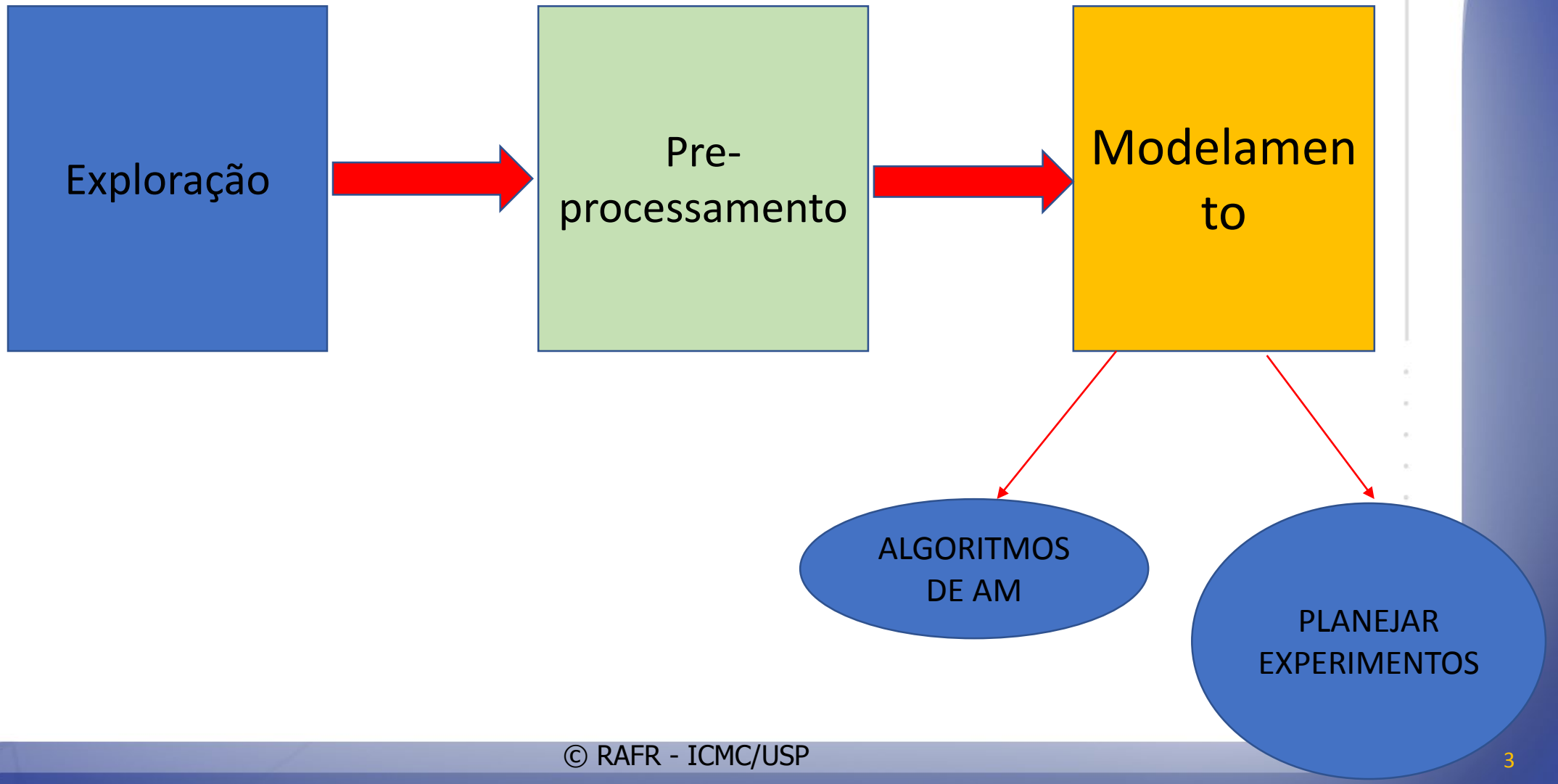
Depto. de Ciências de Computação

ICMC - USP

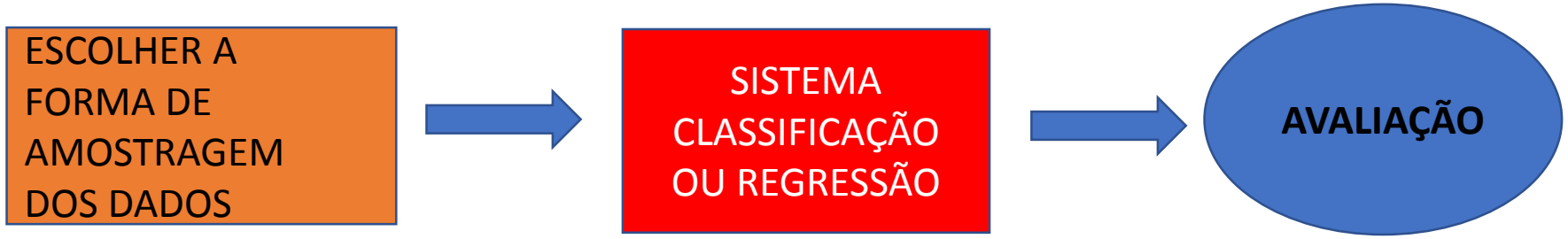
Semana 3

- Paradigmas de Aprendizado.
- Modelos preditivos.
- Partição dos dados.
- Reamostragem (Holdout, bootstrap, K-fold cross validation).
- Modelamento de Dados

3º.ETAPA


Exploração
Pre-processamento
Modelamento
ALGORITMOS DE AM
PLANEJAR EXPERIMENTOS

PLANEJAMENTO DE EXPERIMENTOS e AVALIAÇÃO DE ALGORITMOS



Desempenho Preditivo

- Depende da tarefa a ser resolvida:
  - Classificação: considera taxa de exemplos incorretamente classificados
    - Acurácia
  - Regressão: considera diferença entre valor previsto e valor correto: ERRO
  - Agrupamento: diferentes critérios
- Média dos erros obtidos em diferentes execuções de um experimento

Desempenho Preditivo

- Pode ser avaliado para:

  1 - Buscar o melhor modelo(s) de classificação gerados pelo mesmo algoritmo, variando

    - Valores de hiperparâmetros
    - Partições/atributos nos dados de treinamento

Desempenho Preditivo

2 - Busca encontrar melhor algoritmo de classificação

- Avaliar vários modelos gerados (funções, hipóteses)
  - Hiper-parâmetros de cada algoritmo com valores default ou otimizados
  - Conjuntos de dados com mesmas partições e atributos preditivos

Desempenho preditivo

Principal objetivo:

- Classificação correta de novos exemplos
  - Errar o mínimo possível
    - Minimizar taxa de erro para novos exemplos

Desempenho preditivo

- Taxa de erro

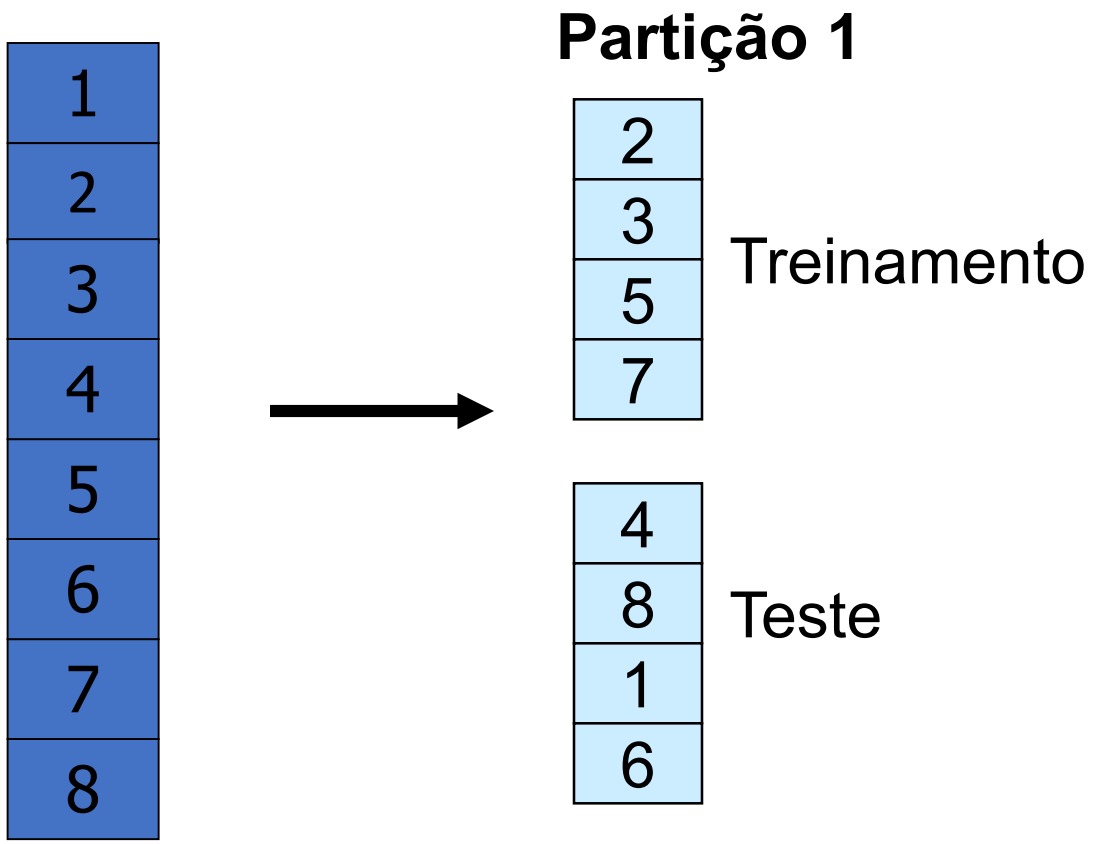
  Geralmente não é possível medir com exatidão; ELA DEVE ser estimada

  - Amostra de teste (simula novos exemplos) do conjunto de dados disponível
  - Utilizando modelo induzido com uma amostra de treinamento do conjunto de dados disponível

Amostragem de dados

- Ajuda a obter a melhor estimativa do desempenho de um modelo ou algoritmo
  - Treinamento (validação) e teste
- Procedimentos
  - Amostragem única
    - *Hold-out*
  - Re-amostragem

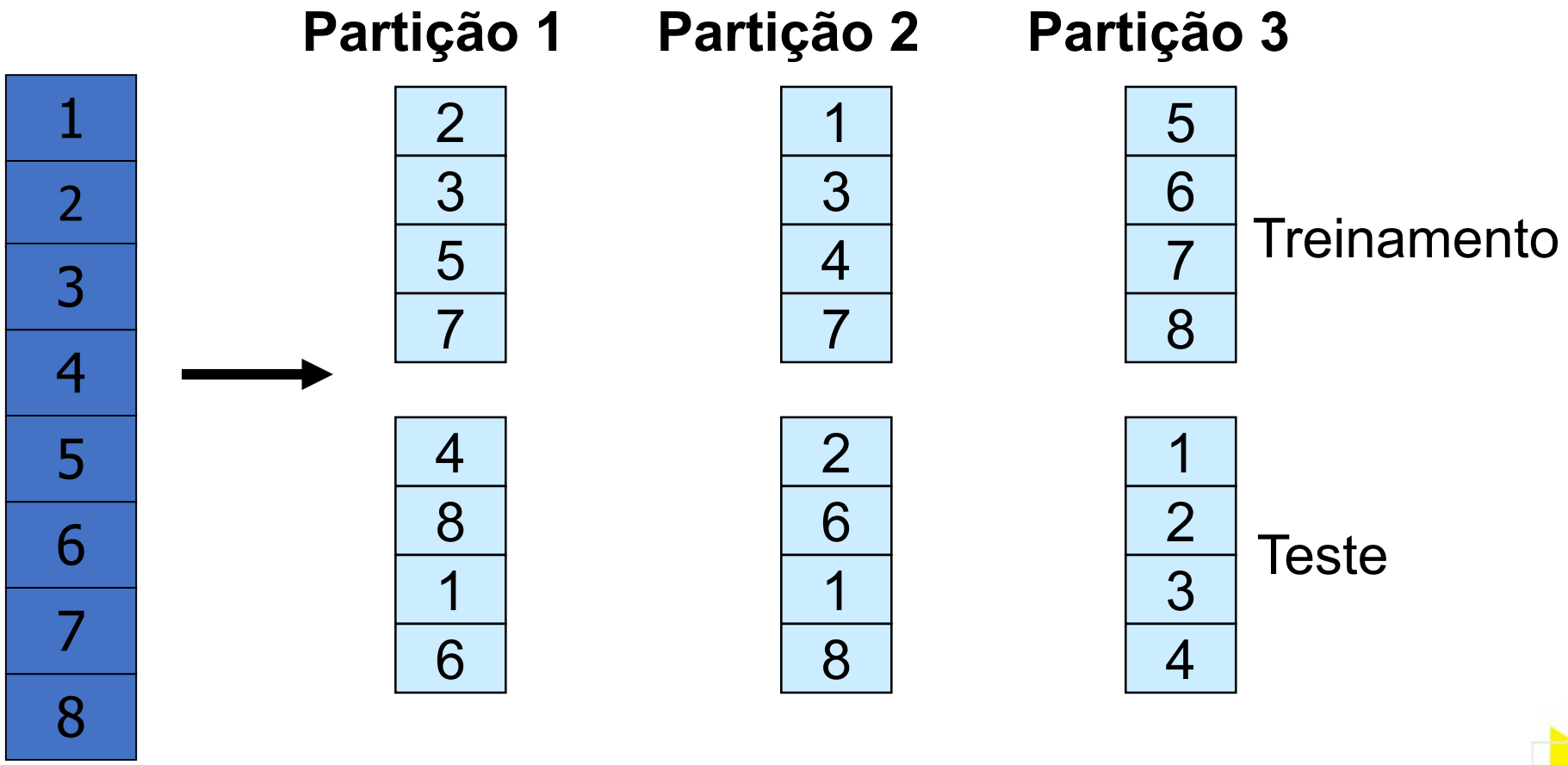
Hold-out



Métodos de Reamostragem

- Amostragem única é pouco confiável
- Reamostragem: geram várias partições para conjuntos de treinamento e teste (validação)
  - *Random subsampling*
  - *K-fold Cross-validation*
    - *Leave-one-out*
  - *Bootstrap (ou Bootstrapping)*

Random subsampling



Bootstrap

- Estocástico, com diversas variações
  - Alguns exemplos podem não participar do treinamento
- Variação mais simples:
  - Amostragem com reposição
    - Cada partição é uma amostra aleatória com reposição do conjunto total de exemplos
      - Conjunto de treinamento têm o mesmo número de exemplos do conjunto total
  - Esta reamostragem é feita muitas vezes (de 1000 a 10000 vezes) para criar uma estimativa da função de distribuição acumulada.

Bootstrap

- Se conjunto original tem N exemplos

escolhe-se um exemplo N vezes

  - A probabilidade de um exemplo não ser amostrado é de: (1-1/n)^n ~1/e ~0.368.
  - O mesmo exemplo pode ser escolhido várias vezes
  - Amostra de tamanho N tem ≈ 63,2% dos exemplos originais
  - O restante: é deixado para teste: 36.8%

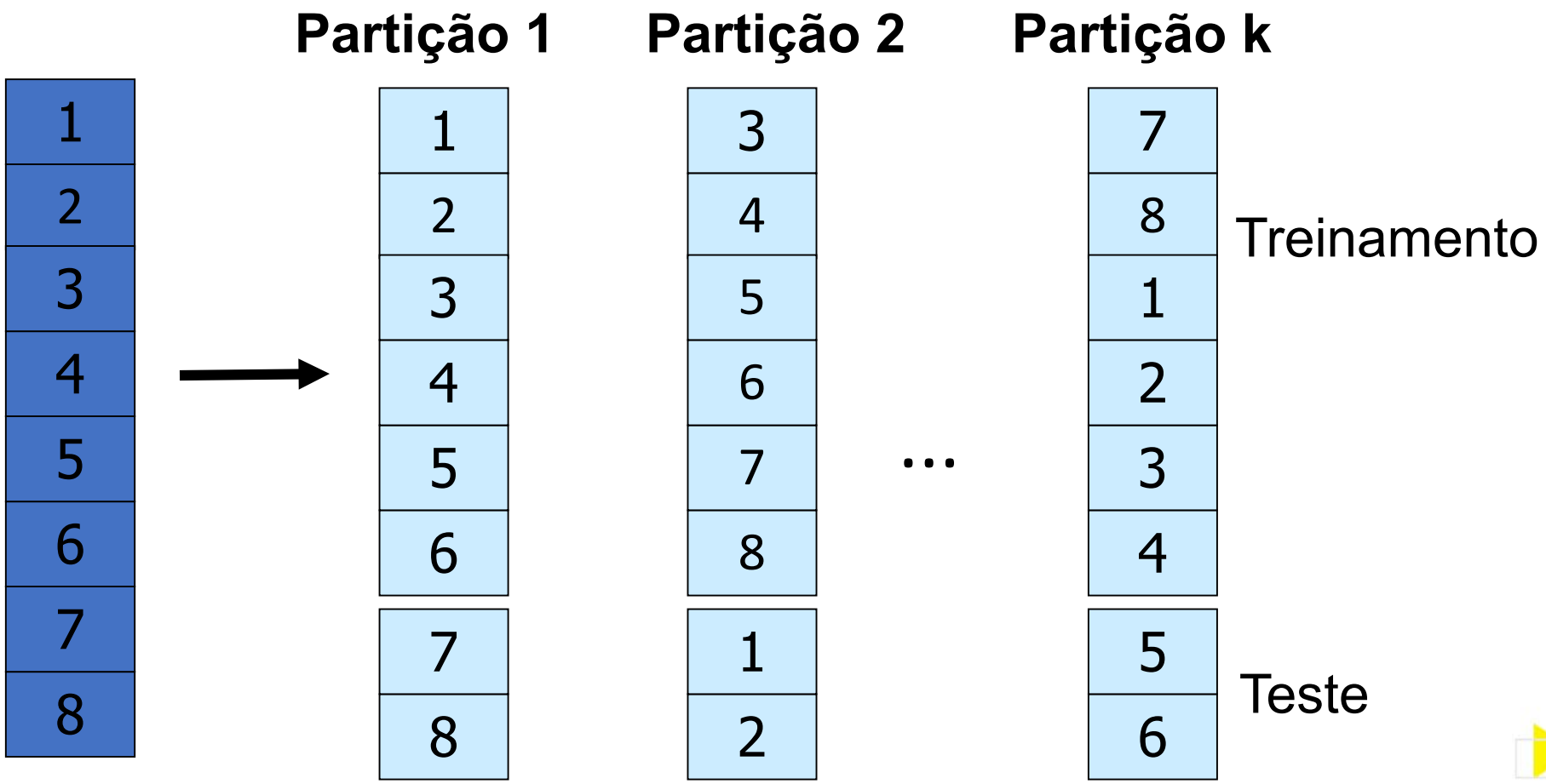
Boostrap

- Processo é repetido k vezes
  - Resultado final é a média dos k experimentos
  - Adequado para conjuntos pequenos.

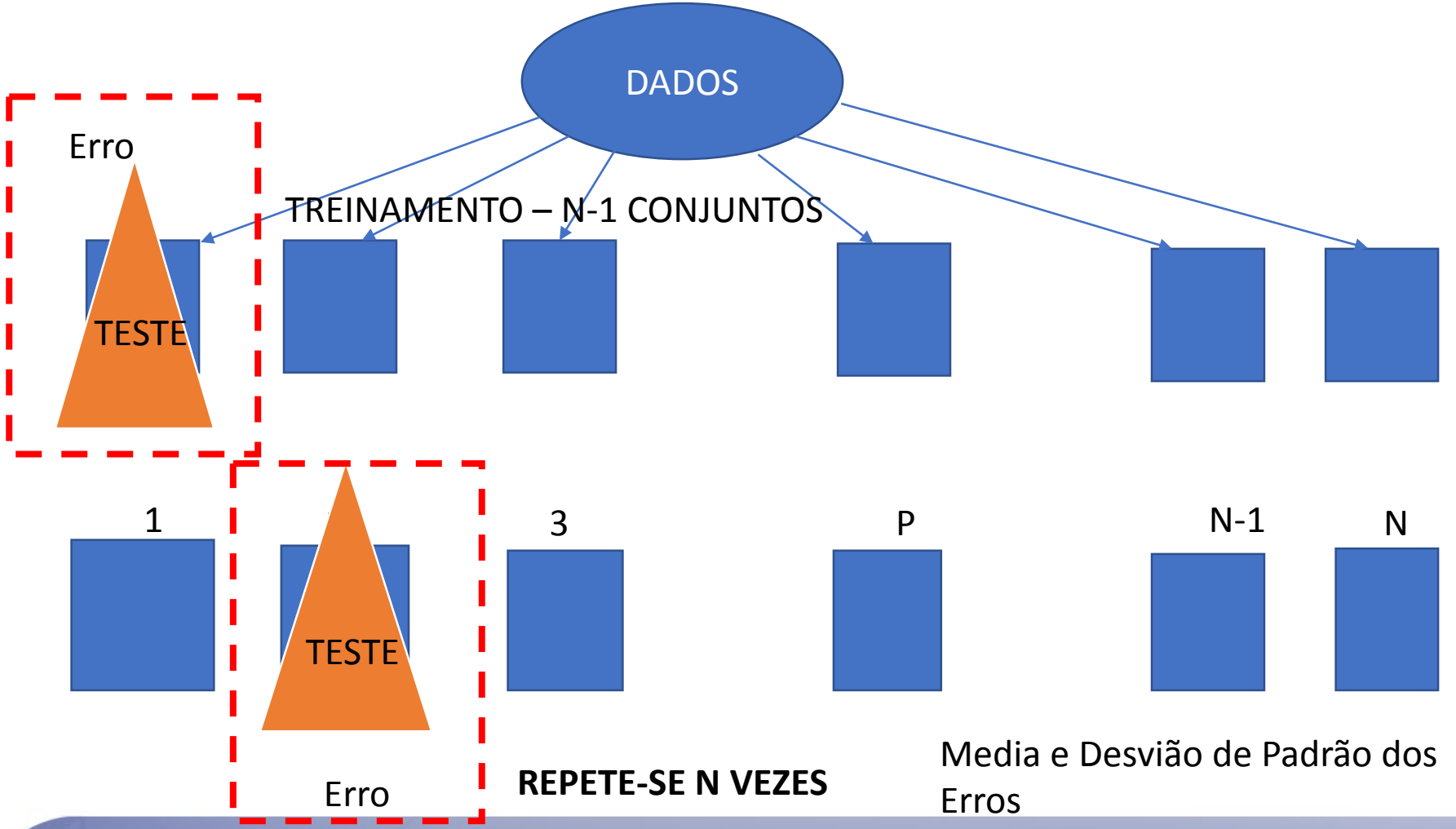
Bootstrap

- Estima incerteza de um algoritmo
  - Tende a ter menor variância e ser mais pessimista que *k-fold cross-validation*

K-fold cross-validation



K – FOLD – VALIDAÇÃO CRUZADA



K – FOLD – VALIDAÇÃO CRUZADA

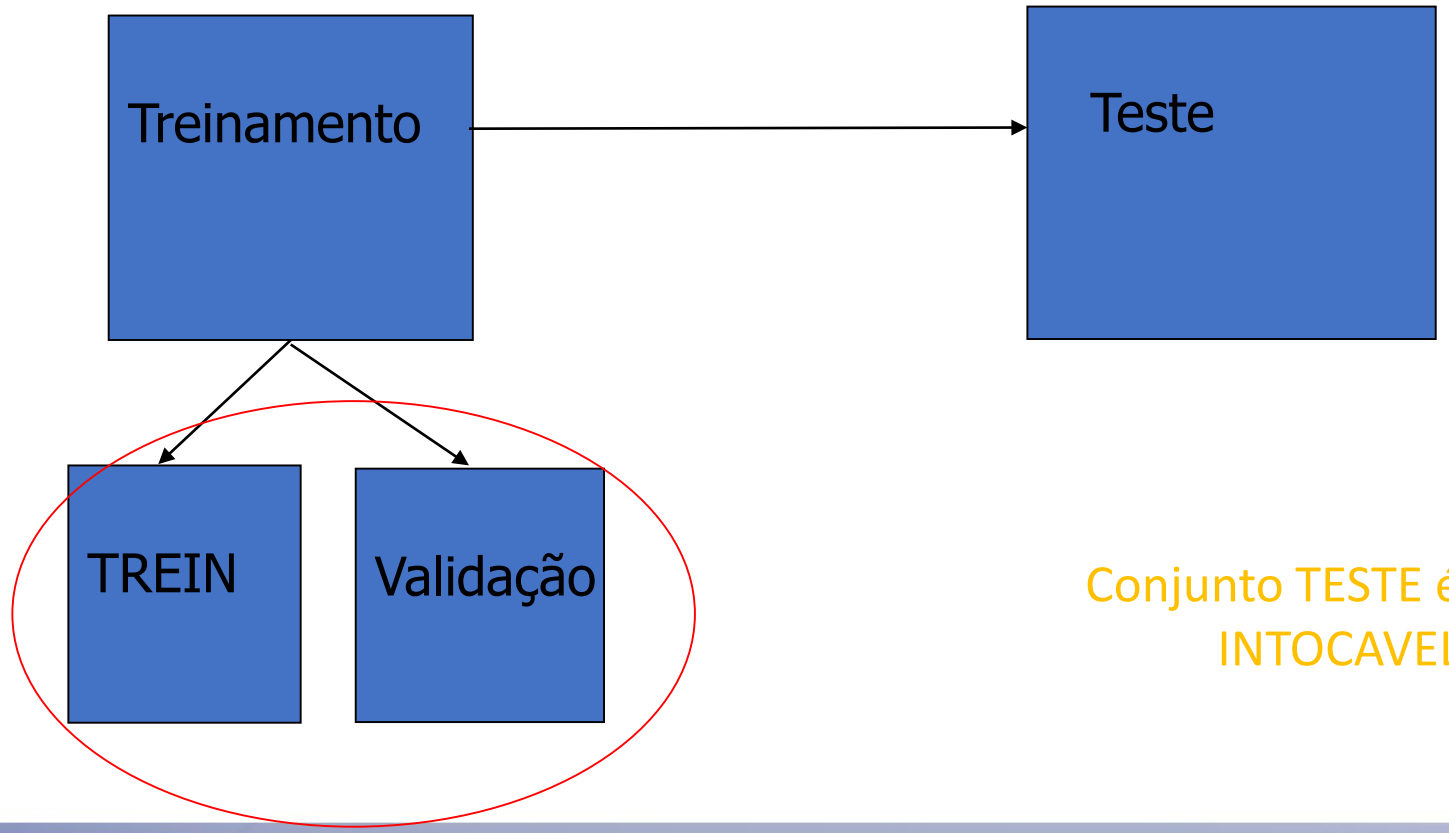
- K-1 folds são usados para treinamento
- 1 fold para teste
- Repete-se o processo k vezes
- Toma-se a **MEDIA E DESVIO PADRÃO** da predição em cada um dos k conjuntos de testes

Leave-one-out

- Consiste em deixar um exemplo de fora (para testar) e treinar com os demais.
- Estimativa de erro praticamente não tendenciosa
  - Tende a taxa de erro verdadeiro
- Computacionalmente caro para conjuntos grandes
  - Geralmente utilizado para pequenos conjuntos de dados
- Variância tende a ser elevada

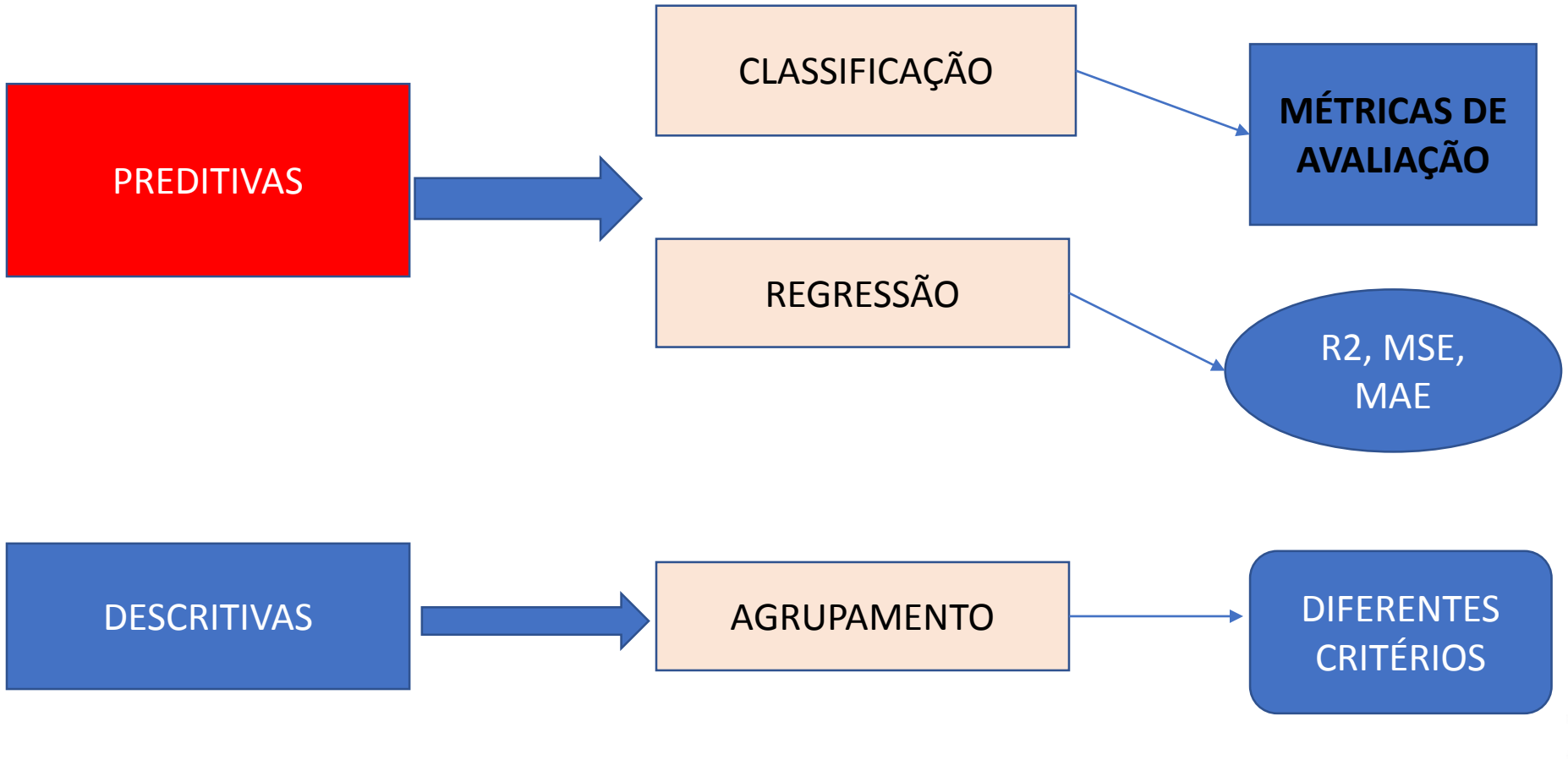
PARTICIONAMENTO DOS DADOS

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
random_state=0)
```



Conjunto TESTE é INTOCAVEL

TAREFAS PREDITIVAS E DESCRITIVAS



Desempenho preditivo

- Medir o Desempenho de AlgoritMos Preditivos:

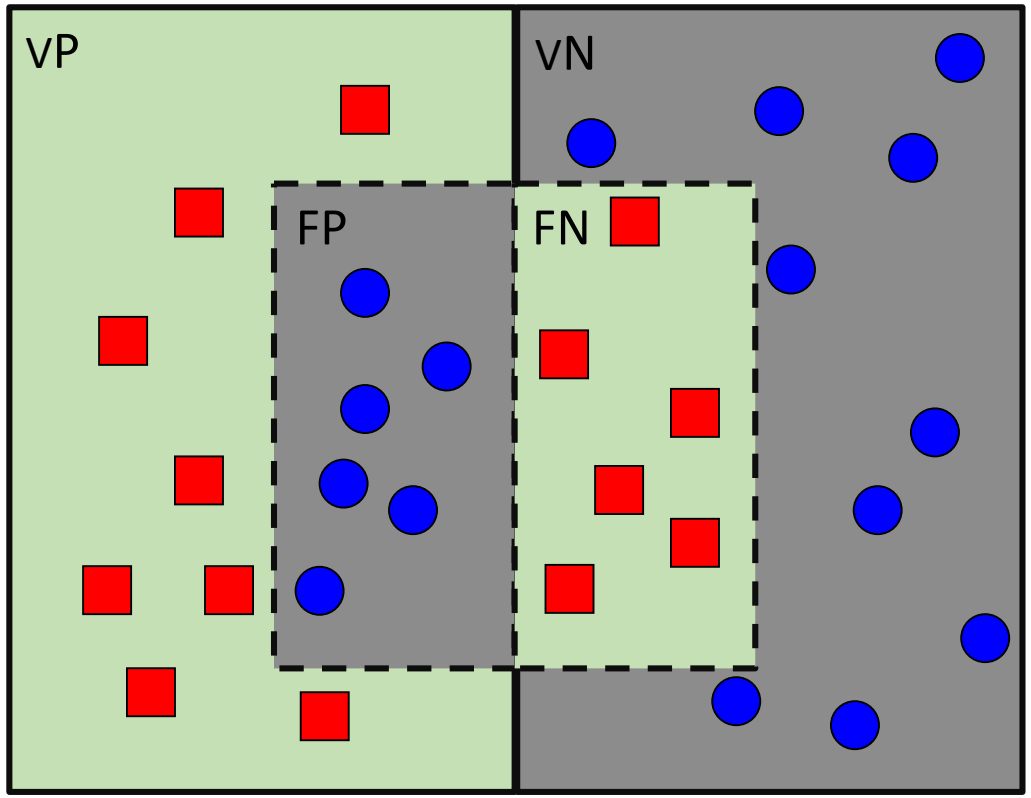
  Tipos de erro: Tipo I, Tipo II

  Medidas: acurácia, precisão, recall, F1

Desempenho preditivo
Classificação binária

- Classe de interesse é a classe positiva
- Dois tipos de erro:
  - Classificação de um exemplo N como P
    - Falso positivo (alarme falso)
      - Ex.: Diagnosticado como doente, mas está saudável
  - Classificação de um exemplo P como N
    - Falso negativo
      - Ex.: Diagnosticado como saudável, mas está doente

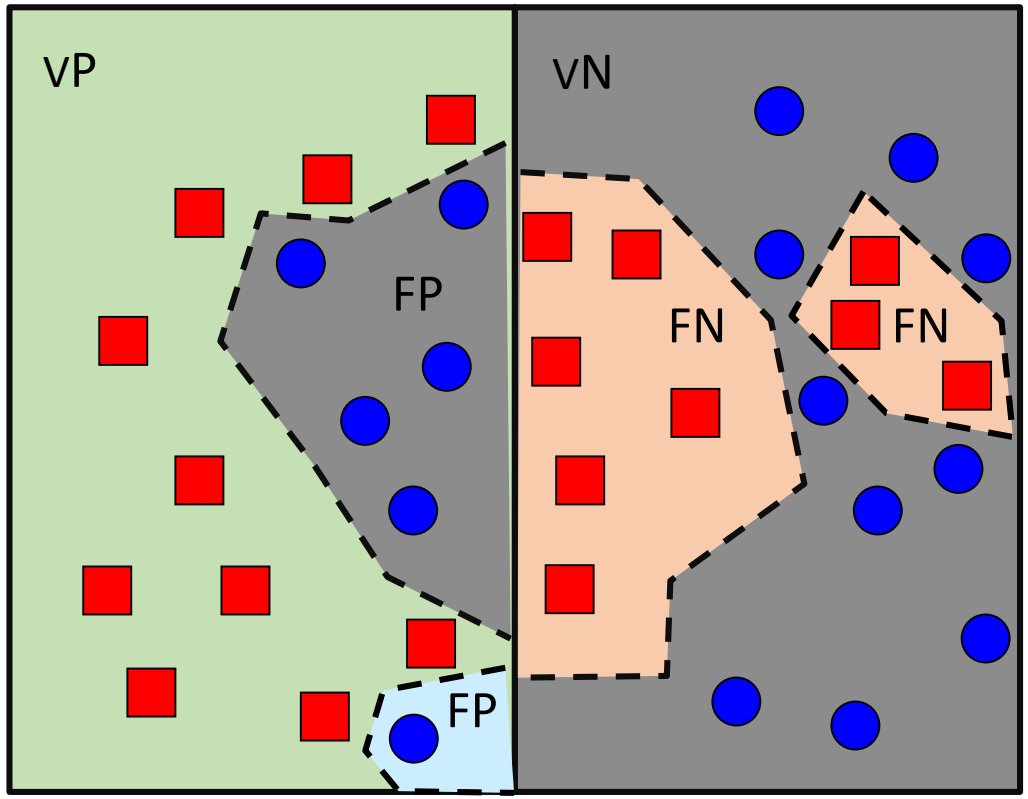
Classificação binária



Reais Positivos

Reais negativos

Classificação binária



Reais positivos

Reais negativos

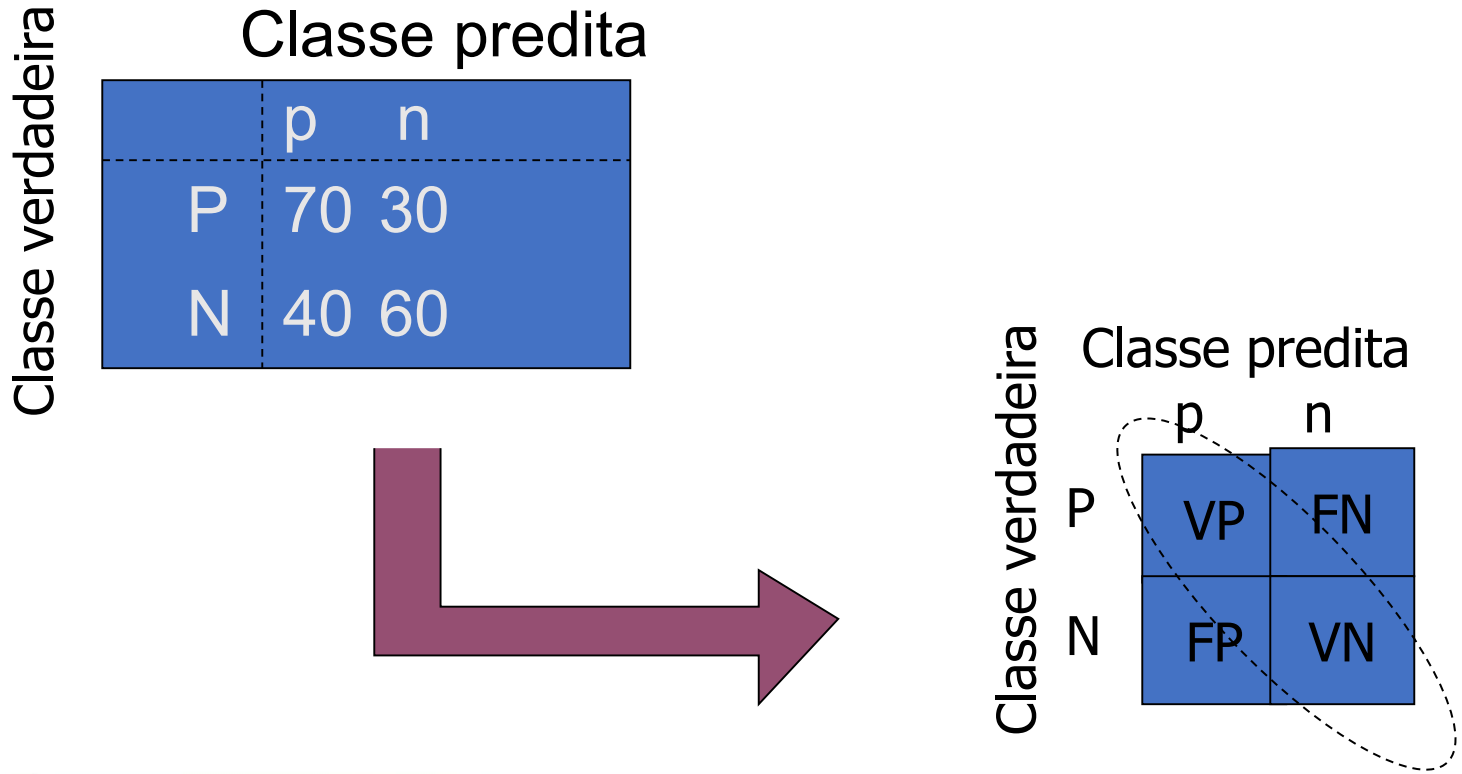
Desempenho preditivo

- **Matriz de confusão (tabela de contingência) pode ser utilizada para distinguir os erros**
  - **Base de várias medidas**
  - **Pode ser utilizada com 2 ou mais classes**

Classe verdadeira \ Classe predita	1	2	3
1	25	0	5
2	10	40	0
3	0	0	20

Exemplo

- Matriz de confusão para 200 exemplos divididos em 2 classes

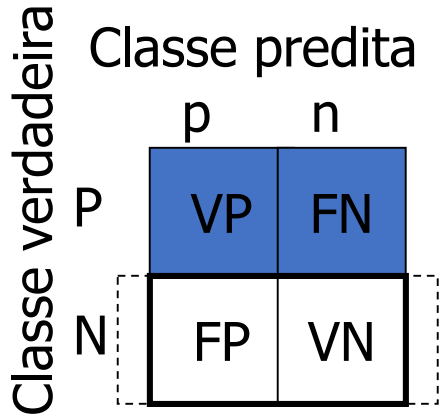


Medidas de avaliação

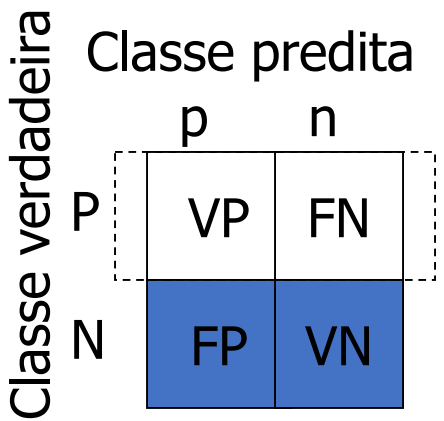
Taxa de FP (TFP) = (Alarmes falsos) $\frac{FP}{FP+VN}$

Taxa de FN (TFN) = $\frac{FN}{VP+FN}$

Erro do tipo I



Erro do tipo II



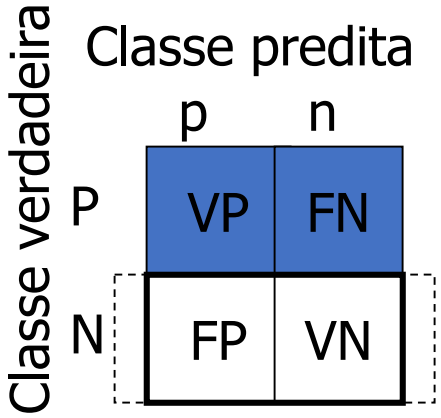
Medidas de avaliação

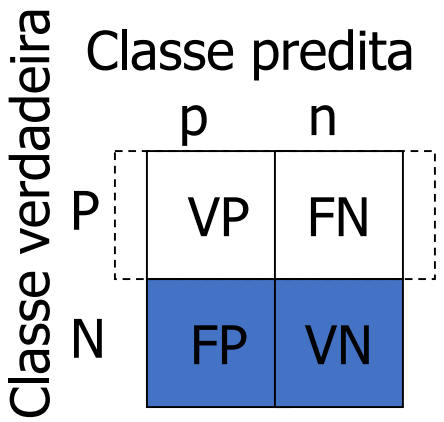
Taxa de FP (TFP) = $\frac{FP}{FP+VN}$
(Alarmes falsos)

Custo

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

Benefício





Exemplo

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

- Avaliação de 3 classificadores

Classe verdadeira \ Classe predita	p	n
P	20	30
N	15	35

Classificador 1
TVP =
TFP =

Classe verdadeira \ Classe predita	p	n
P	70	30
N	50	50

Classificador 2
TVP =
TFP =

Classe verdadeira \ Classe predita	p	n
P	60	40
N	20	80

Classificador 3
TVP =
TFP =

Exemplo

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

- Avaliação de 3 classificadores

Classe verdadeira \ Classe predita	p	n
P	20	30
N	15	35

Classificador 1
TVP = 0.4
TFP = 0.3

Classe verdadeira \ Classe predita	p	n
P	70	30
N	50	50

Classificador 2
TVP = 0.7
TFP = 0.5

Classe verdadeira \ Classe predita	p	n
P	60	40
N	20	80

Classificador 3
TVP = 0.6
TFP = 0.2

Medidas de avaliação

$$\frac{FP}{FP+TN}$$

False positive rate (FP) = 1-TN

$$\frac{FN}{TP+FN}$$

False negative rate (FN) = 1-TP

$$\frac{TP}{TP+FN}$$

True positive rate (TP), also known as recall or sensitivity

$$\frac{TN}{TN+FP}$$

True negative rate (TN), also known as specificity

$$\frac{TP}{TP+FP}$$

Positive predictive value (PPV), also known as Precision

$$\frac{TN}{TN+FN}$$

Negative predictive value (NPV)

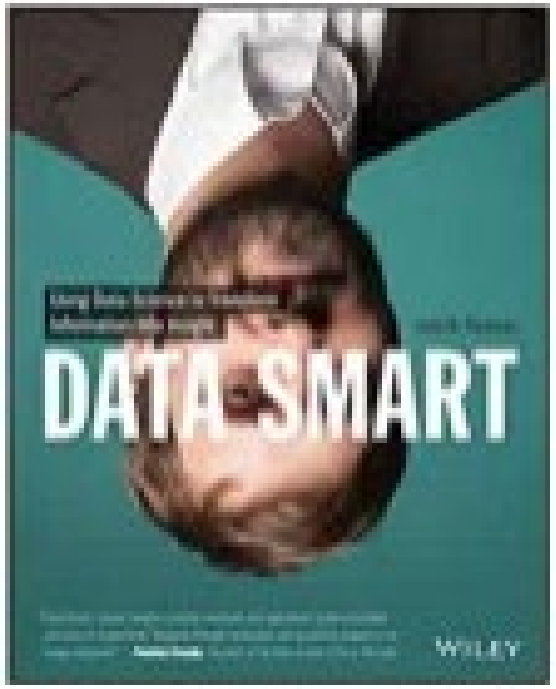
$$\frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy

$$\frac{2}{1/\,precision+1/\,recall}$$

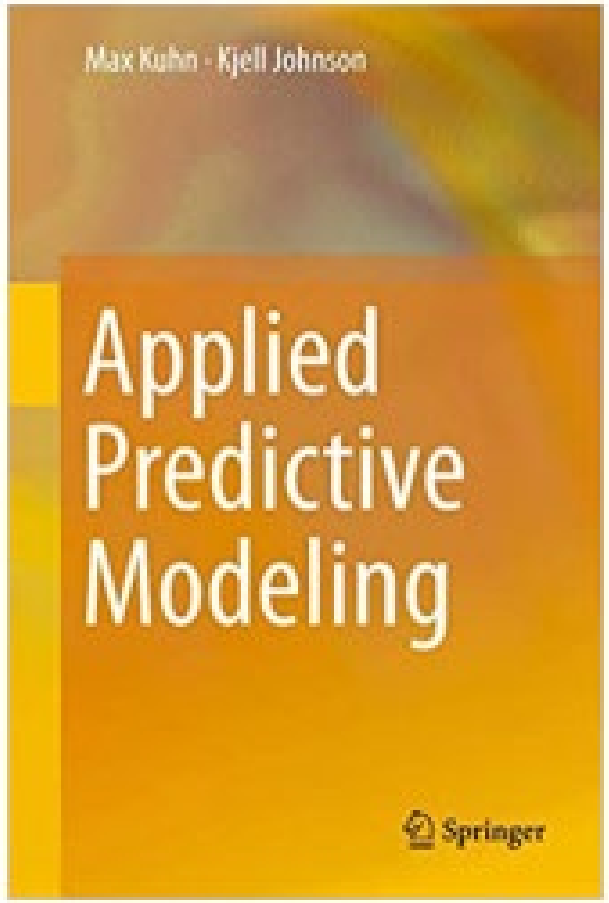
F1-measure

Data Smart: Using Data Science to Transform Information into Insight



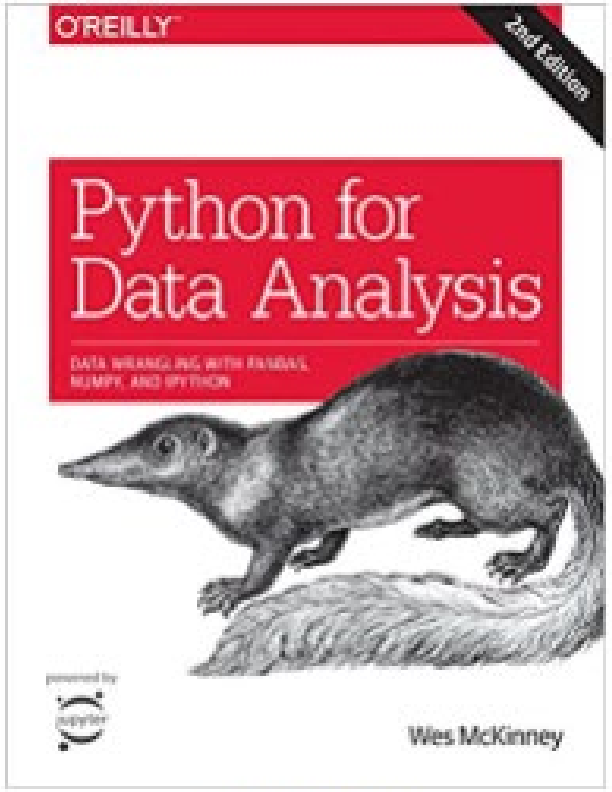
Source: Amazon.com

Applied Predictive Modeling



Source: Amazon.com

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython



Source: Amazon.com