Seleção de Variáveis e Validação Cruzada

Prof. Gustavo Willam Pereira

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Seleção de Variáveis

- A seleção de features (recursos) é uma das etapas mais importantes do aprendizado de máquina.
- É o processo de restringir um subconjunto de recursos a serem usados na modelagem sem perder as informações.
- Às vezes, a seleção de recursos é confundida com a redução de dimensionalidade.
- Ambos os métodos tendem a reduzir o número de recursos no conjunto de dados, mas de maneira diferente.
- A redução de dimensionalidade reduz o número de recursos criando novos recursos como combinações dos existentes.
- Todos os recursos são combinados para criar alguns recursos exclusivos.
- A seleção de recursos, por outro lado, funciona eliminando os recursos irrelevantes e mantendo apenas os relevantes.

Seleção de Variáveis

- Na aula de Regressão Linear Múltipla vimos uma forma de fazer a seleção de variáveis através dos métodos:
  - Backward Elimination
    - O procedimento começa com um conjunto vazio de features (variáveis). O melhor dos atributos é determinado e adicionado ao conjunto reduzido. A cada iteração subsequente, o melhor dos atributos originais restantes é adicionado ao conjunto.
  - Forward Elimination
    - O procedimento começa com o conjunto completo de atributos. A cada passo, ele remove o pior atributo restante no conjunto.
  - Bidirecional Elimination (Stepwise Elimination)
    - É a combinação dos 2 métodos acimas, a cada passo, o procedimento seleciona o melhor atributo e remove o pior dentre os demais atributos.

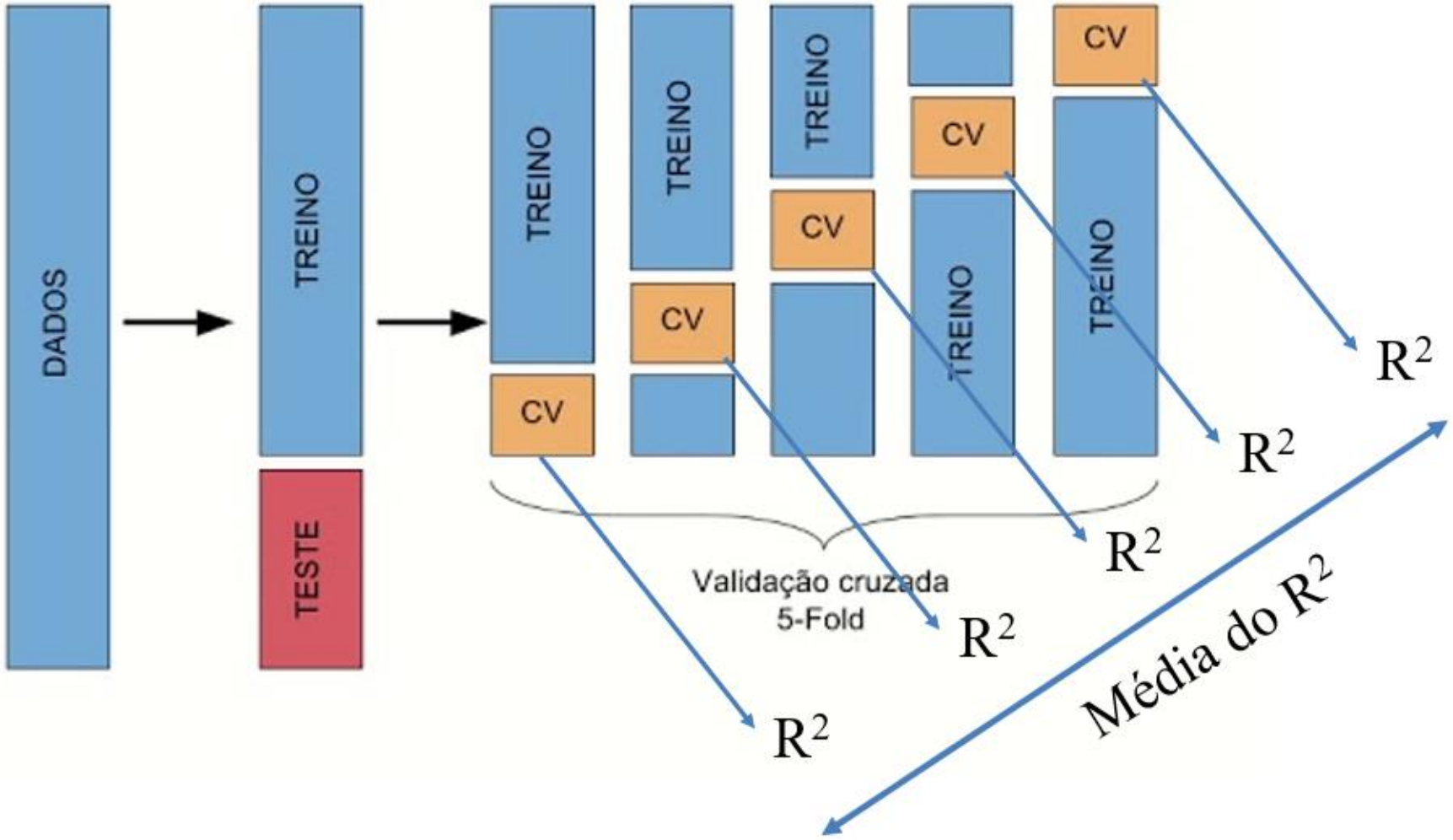
Seleção de Variáveis

- As principais vantagens da seleção de recursos são:
  - Melhora o desempenho do modelo: quando você tem recursos irrelevantes em seus dados, esses recursos agem como um ruído, o que faz com que os modelos de aprendizado de máquina tenham um desempenho ruim.
  - Modelos de aprendizado de máquina mais rápidos.
  - Evita overfitting, o que aumenta a generalização do modelo.

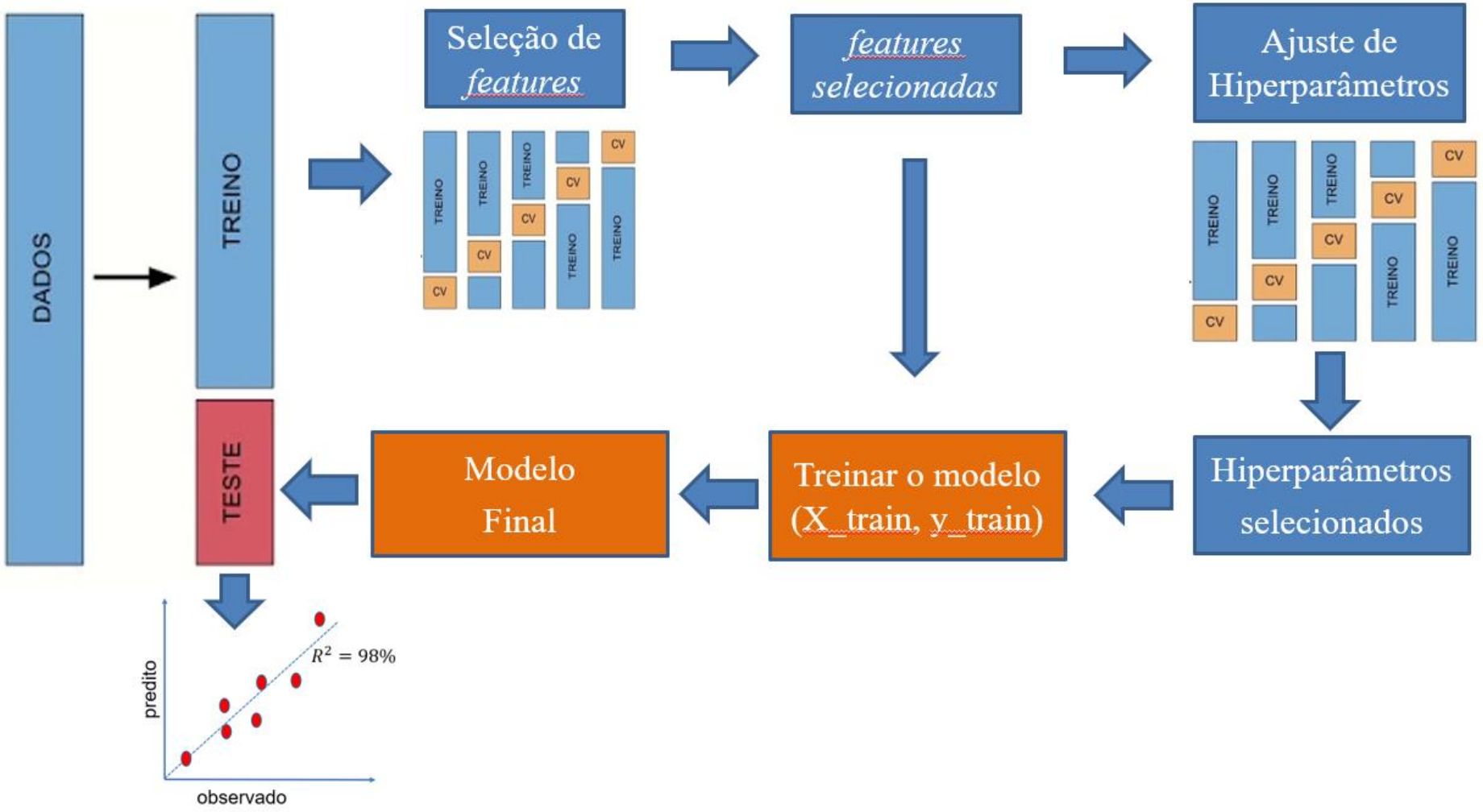
- A seleção de features pode ser feita de várias maneiras e veremos alguns dos métodos de seleção de features do Scikit-learn: sklearn.feature_selection.
- Iremos utilizar o dataset: **Breast Cancer Wisconsin (Diagnostic)**

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Validação Cruzada K-fold



Seleção de *features (variáveis)*



INSTITUTO FEDERAL
Sudeste de Minas Gerais