MÓDULO III

Regressão
Outros modelos

Alexandre Loureiros Rodrigues
Professor do departamento de Estatística - UFES

ESPECIALIZAÇÃO
INTELIGÊNCIA ARTIFICIAL
& CIÊNCIA DE DADOS

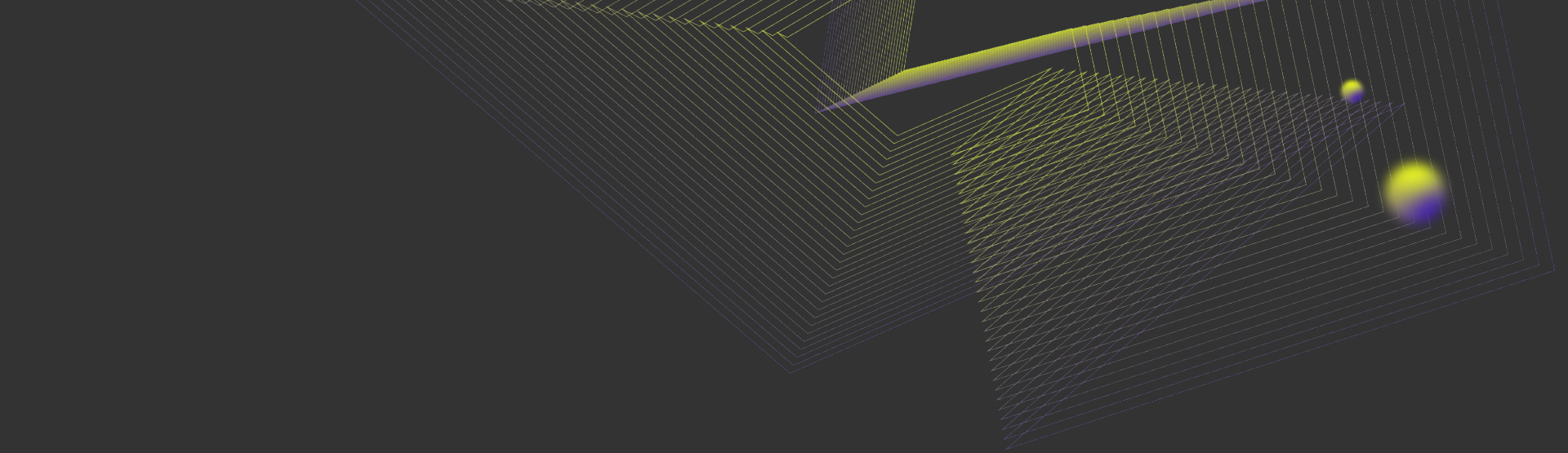
SEAD UFES | Superintendência de Educação a Distância

ÍNDICE

1. Introdução

Métodos não paramétricos

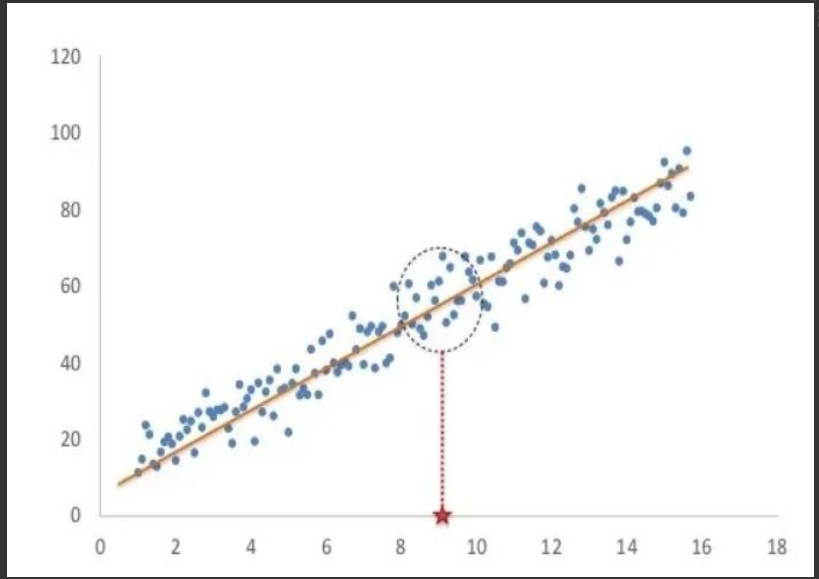
- Até agora nos restringimos a métodos paramétricos para realizar previsões
  - Muitas vezes este métodos são muito restritivos
- Forma linear pode não ser adequada para vários problemas
- Vamos explorar métodos menos interpretativos, mas com maior poder de previsão.



2. K vizinhos mais próximos

KNN

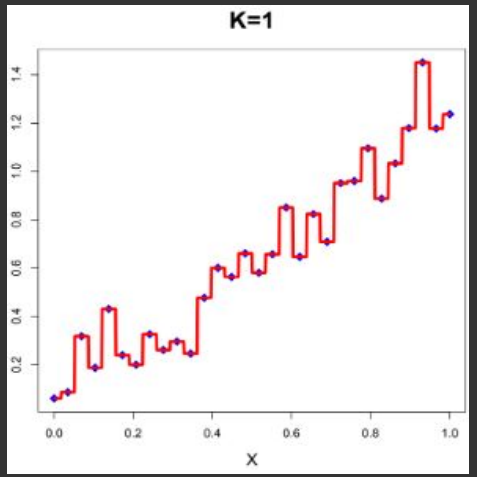
- Proposto por Stone (1977), métodos muito popular em machine learning
- Motivação: uso de médias locais
- Previsão dada pela média dos **K** vizinhos mais próximos nos espaço de características

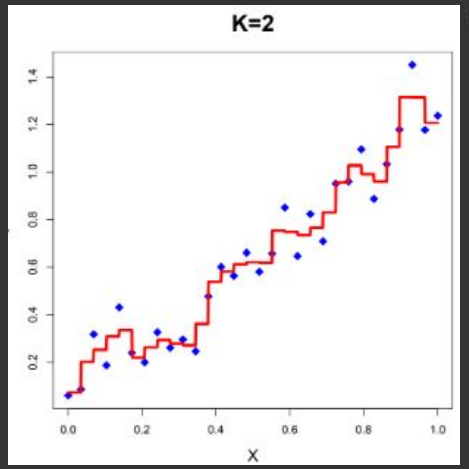


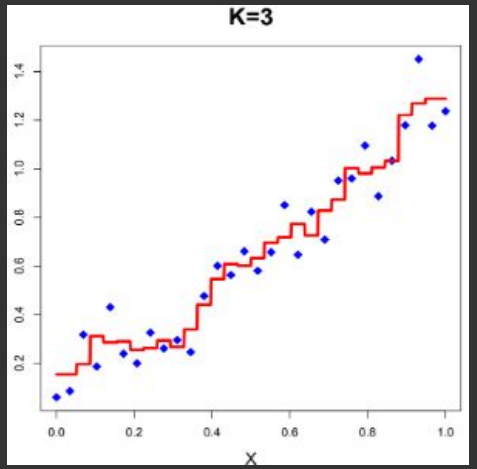
Efeito do K

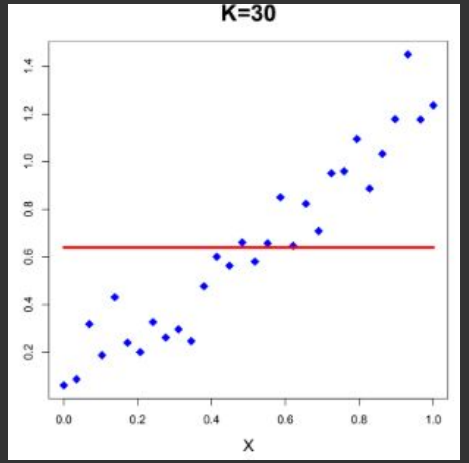
- Como escolher o melhor K ?

- Validação cruzada





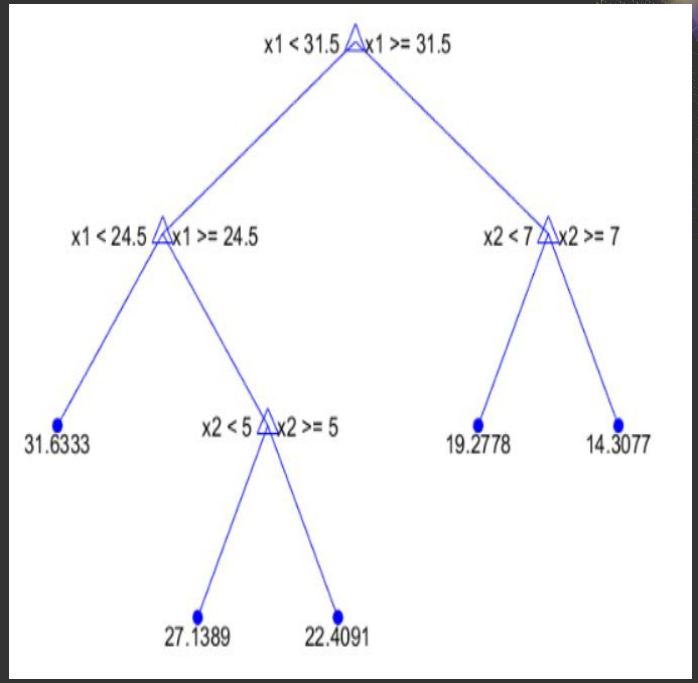


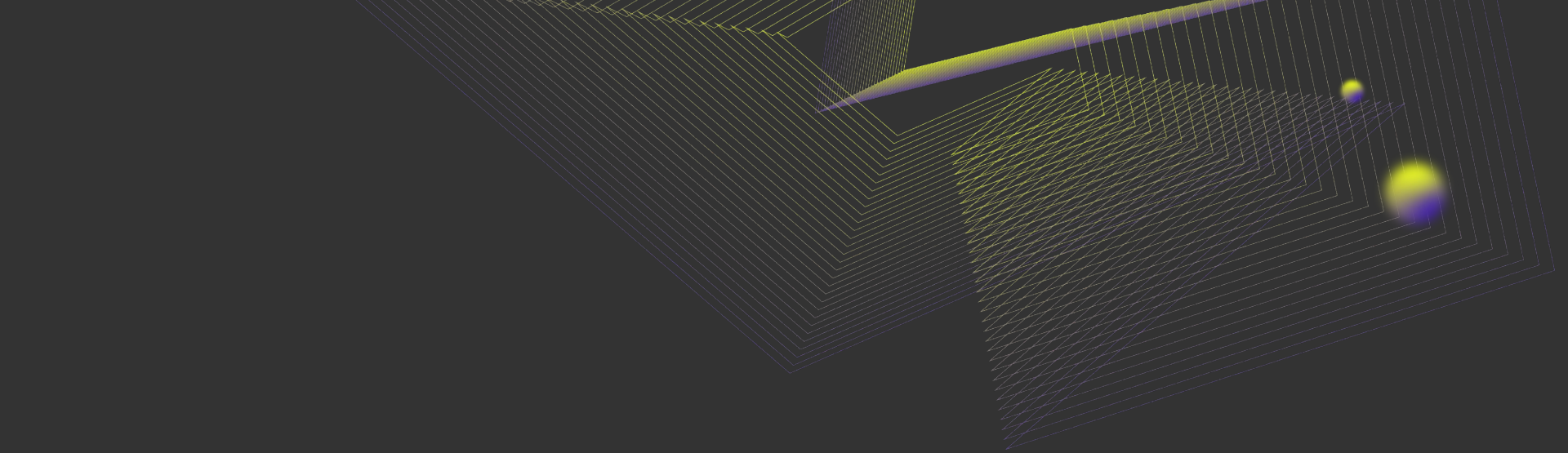


3. Árvore de decisão

Árvore de decisão

- Estrutura Modelo hierárquico com nós internos (condições) e folhas (previsões).
- **Critérios de Divisão**: **Gini** (impureza), **Entropia** (ganho de informação), **Redução de variância** (regressão).
- **Poda e Overfitting**: **Pré-poda** (limita profundidade) **Pós-poda** (remove ramos irrelevantes).





4. Florestas aleatórias

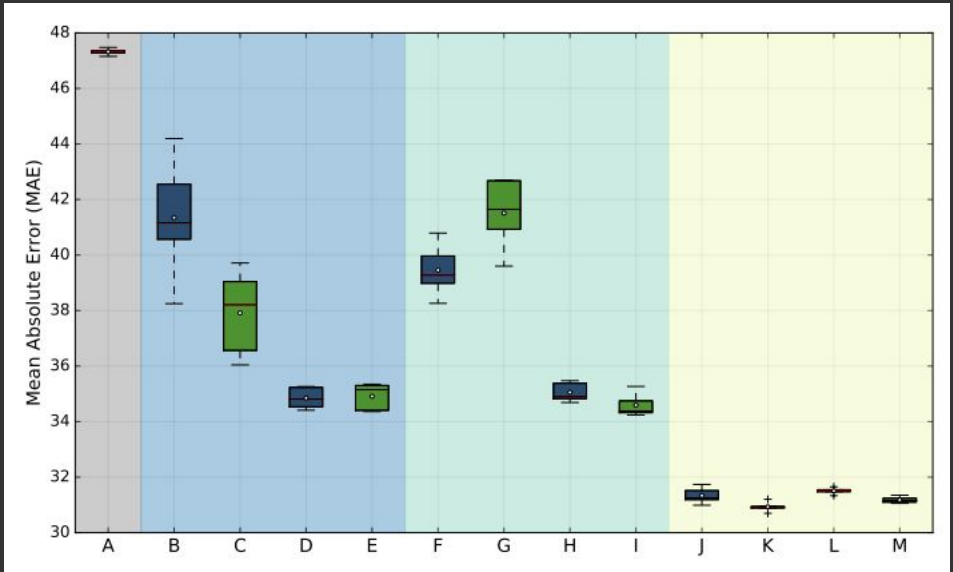
Florestas Aleatórias

- Conjunto de várias árvores de decisão treinadas com amostras aleatórias (bagging).
  - Vantagens: Reduz overfitting, funciona bem com grandes volumes de dados.
  - Desvantagens: Alto custo computacional, menos interpretável.
- Cada árvore recebe uma amostra aleatória do dataset e usa um subconjunto de variáveis para decisões. A predição final é feita por média (regressão)

5. Considerações finais

Pontos de atenção

- Em aplicações nas quais as observações são dependentes é ideal que pelo menos os conjuntos de teste e treino sejam independentes (evitar bias de similaridade).
- Usar validação cruzada pareada para comparar diferentes métodos de regressão em uma mesma base.
- Usar média, desvio padrão e boxplot para comparar performance dos diferentes métodos
    - Para maior rigor, usar testes estatísticos (t-pareado, Wilcoxon, Friedman)



Comparação de regressores

Problema: Previsão do consumo de energia

TABLE II

METRICS OF THE BEST CONFIGURATION OF EACH MODEL VARIATION

Model	MAE	MAPE	MdAPE
Baseline	47.32 ± 0.12	97.91 ± 0.92	16.71 ± 0.03
FC + CW[a]	34.85 ± 0.39	62.67 ± 2.12	12.43 ± 0.13
CNN + CW[a]+ Meta[b]	34.59 ± 0.42	62.90 ± 0.53	12.78 ± 0.30
LSTM + STD[c]+ Meta[b]	**30.93 ± 0.18**	**46.53 ± 0.28**	**10.68 ± 0.03**