

# Projeto de aprendizado de máquina

Prof. Wesin Ribeiro



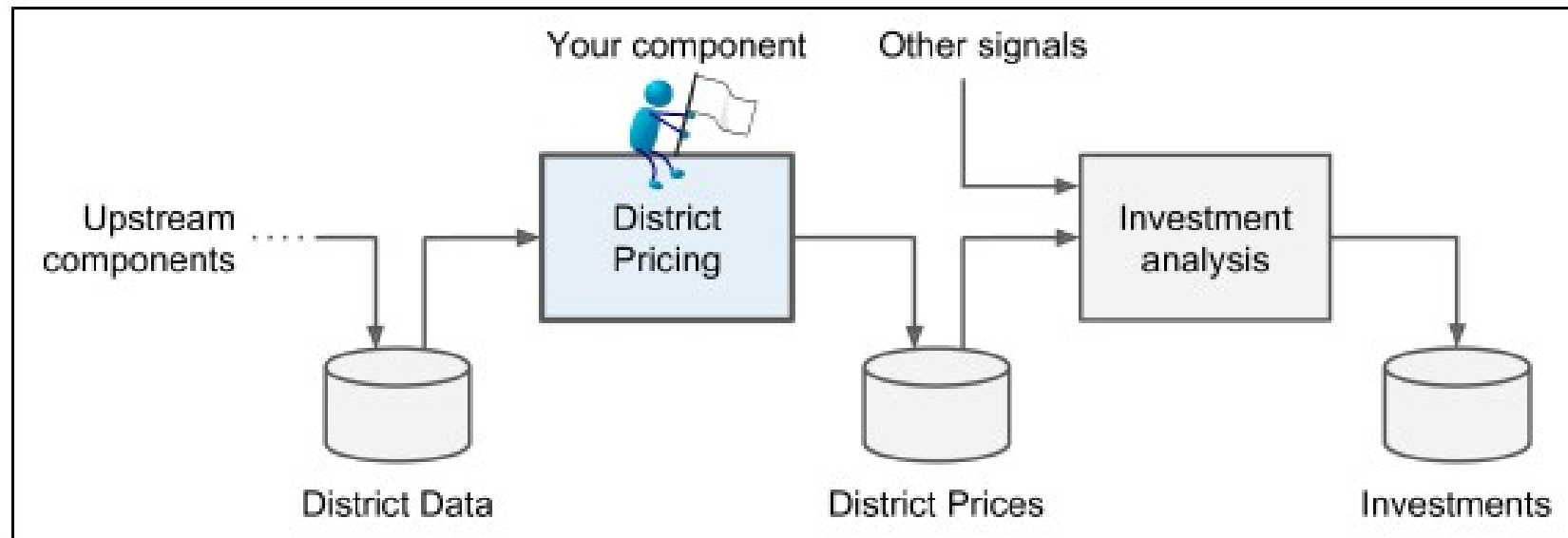
# ETAPAS

---

- **Análise do problema**
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

# A sua tarefa

- Você trabalha para um empresa americana de investimentos e precisa criar um modelo para calcular o preço das casas usando os dados de censo da California.



# Abordagem utilizada

- Conhecer solução atual se houver
- Aprendizado supervisionado
- Tarefa de regressão
- Regressão univariada
- Aprendizado por lote

# Como avaliar o desempenho do seu modelo?

- Erro médio quadrático
- Erro médio absoluto

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left( h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^m \left| h(\mathbf{x}^{(i)}) - y^{(i)} \right|$$



# Verifique as hipóteses

- E se o sistema que coleta os dados dos preços das casas fornecesse o valor em três categorias: barato, médio, caro?







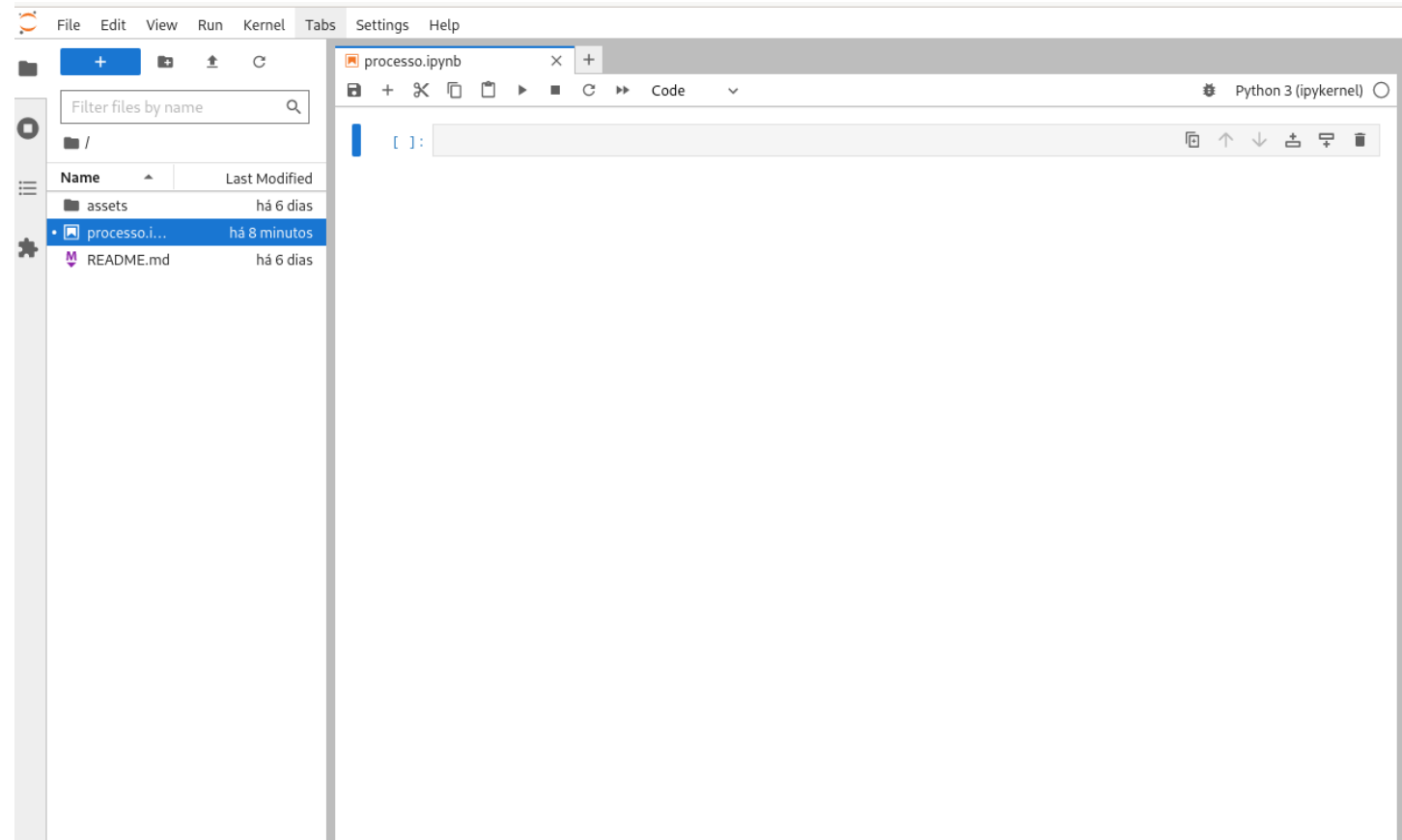
# ETAPAS

---

- Análise do problema
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

# Prepare seu ambiente de trabalho

- Organize o diretório
- Configure as variáveis de ambiente (se for o caso)
- Instale as ferramentas





# Separar conjunto de treino e teste

- De forma aleatória
- De forma estratificada (a população é dividida de forma homogênea em subgrupos)



# ETAPAS

---

- Análise do problema
- Coleta de dados
- **Análise exploratória de dados**
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

# Explorar e visualizar dados

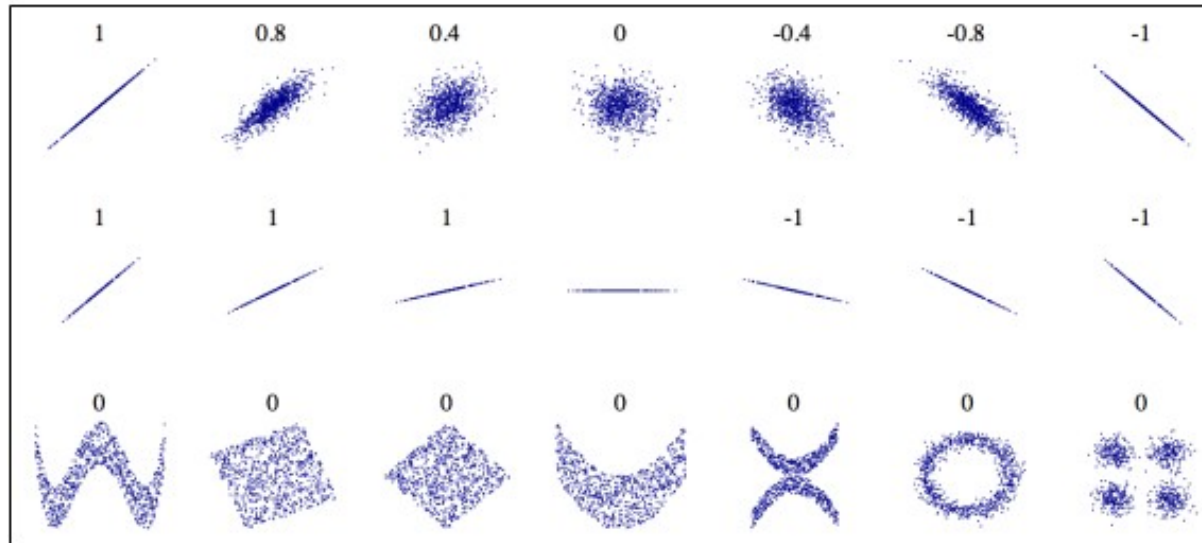
- Obter informações úteis
- Colher uma pequena amostra para fazer a análise
- Fazer cópia dos dados para não alterar acidentalmente





# Buscar correlações

- O coeficiente de correlação varia de  $-1$  a  $1$
- Perto de  $1$ , indica uma forte correlação positiva
- Perto de  $-1$ , indica uma forte correlação negativa





# ETAPAS

---

- Análise do problema
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

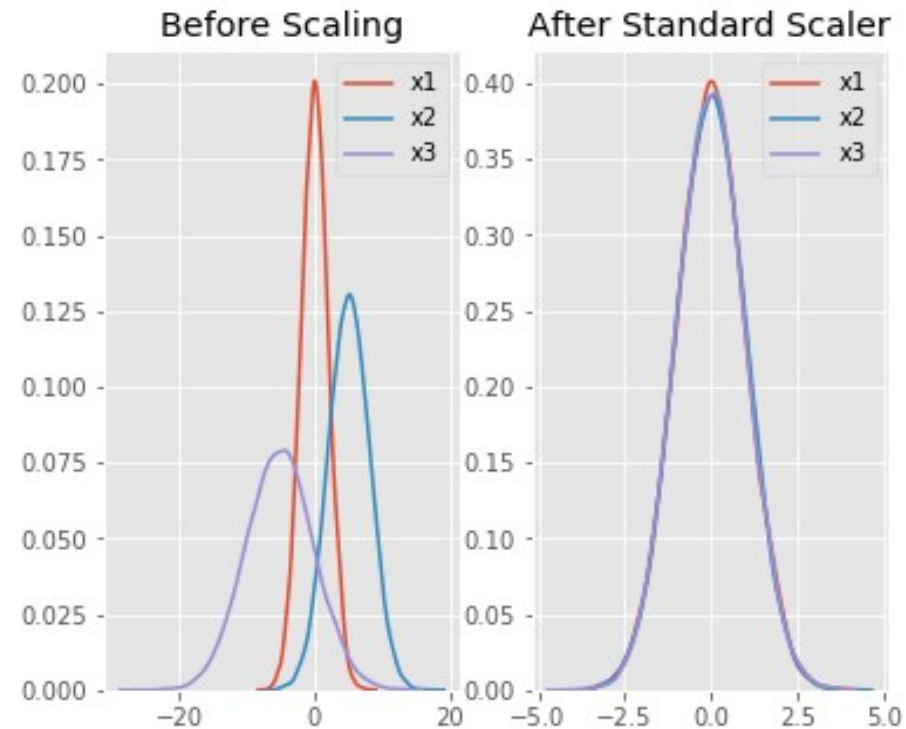
# Preparar os dados para os algoritmos

- Limpeza dos dados
- Manipular texto e atributos categóricos
- Transformação customizada
- Escalonamento de atributos
- Esteira de transformação



# Escalonamento de atributos

- Min-max ou normalização mantem os valores escalonados entre 0 e 1.
- A padronização transforma os valores em torno da média 0.





# ETAPAS

---

- Análise do problema
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- **Seleção e treinamento de modelos**
- Ajuste fino
- Lançamento

# Treinamento e avaliação do modelo

- Você já fez o todo o trabalho sujo!
- Comece selecionando o modelo mais simples possível
- Use a validação cruzada para obter um resultado mais confiável





# ETAPAS

---

- Análise do problema
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

# Ajuste fino do seu modelo

- Busca manual
- Busca em grade
- Busca aleatória

# Com o melhor modelo em mãos

- Avalie seu sistema sobre o conjunto de teste
- Lembre-se de executar a esteira de transformação sobre os dados de teste
- Utilize intervalos estatísticos de confiança para conferir os resultados





# ETAPAS

---

- Análise do problema
- Coleta de dados
- Análise exploratória de dados
- Engenharia de atributos
- Seleção e treinamento de modelos
- Ajuste fino
- Lançamento

# Lance, monitore e dê manutenção

- Colete dados do ambiente de produção
- Monitore o desempenho do sistema em intervalos regulares
- Avalie a qualidade dos dados de entrada
- Retreine seu modelo com uma base de dados mais atualizada

# Obrigado pela atenção!

Desafio: repita o processo para outra base de dados