Analytics e Inteligência Artificial

Aula 13

Aprendizagem Supervisionada - Regressão

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

LABDATA FIA

NOSSOS DIFERENCIAIS | QUEM SOMOS

O **Laboratório de Análise de Dados** - LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.

Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Corpo Diretivo

COORDENADORES DO LABDATA | ATUAÇÃO ACADÊMCIA E PROFISSIONAL

Profª Dra.
Alessandra Montini

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Têm muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em estatística aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, Coordenadora de Grupos de Pesquisa no CNPQ, Parecerista da FAPESP e Colunista de grandes Portais de Tecnologia.

linkedin.com/in/alessandramontini/

Prof. Dr.
Adolpho Walter Canton

Diretor do LABDATA-FIA. Consultor em Projetos de *Analytics*, *Big Data* e Inteligência Artificial. Professor FEA – USP. PhD em Estatística Aplicada pela *University of North Carolina at Chapel Hill*, Estados Unidos.

Currículo – Prof. João Nogueira

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

- **(2019-Presente)** – Professor nos cursos de Extensão, Pós e MBA em Big Data e Data Mining na Fundação Instituto de Administração (FIA) – www.fia.com.br
- **(2018-Presente)** – Cientista de Dados na Via Varejo – https://viavarejo.com.br
- **(2016-Presente)** – Doutorando em Física Computacional e Estatística pelo Departamento de Física na Universidade Federal do Ceará – https://física.ufc.br
- **(2014-2016)** – Mestre em Física da Matéria Condensada pelo Departamento de Física na Universidade Federal do Ceará - https://física.ufc.br
- **(2012-2013)** – Estudante Intercambista na Universidade de Coimbra – Portugal – https://www.uc.pt
- **(2010-2014)** – Bacharel em Física pela Universidade Federal do Ceará – http://www.ufc.br
- **Contatos:**
  - E-mail: joaonogueira@fisica.ufc.br

Conteúdo Programático da Disciplina – Projeto de Inteligência Artificial

Data	Horário	Tema
09/03/2021	19:00	Aula 1 - Introdução ao Ambiente de Desenvolvimento
11/03/2021	19:00	Aula 2 - Revisão de Python
16/03/2021	19:00	Aula 3 - Manipulação de Dados
18/03/2021	19:00	Aula 4 - Análise Exploratória de Dados
23/03/2021	19:00	Aula 5 - Projeto da disciplina - Parte 1 - Análise Exploratória de Dados
25/03/2021	19:00	Aula 6 - Introdução, Motivação e Framework de Machine Learning
06/04/2021	19:00	Aula 7 - Analytical Base Table
08/04/2021	19:00	Aula 8 - Aprendizagem Supervisionada - Classificação
13/04/2021	19:00	Aula 9 - Aprendizagem Supervisionada - Classificação
15/04/2021	19:00	Aula 10 - Aprendizagem Supervisionada - Classificação
20/04/2021	19:00	Aula 11 - Projeto da disciplina - Parte 2 - Machine Learning - Classificação
22/04/2021	19:00	Aula 12 - Projeto da disciplina - Parte 2 - Machine Learning - Classificação
27/04/2021	19:00	Aula 13 - Aprendizagem Supervisionada - Regressão
29/04/2021	19:00	Aula 14 - Aprendizagem Supervisionada - Regressão
04/05/2021	19:00	Aula 15 - Projeto da disciplina - Parte 3 - Machine Learning - Regressão
06/05/2021	19:00	Aula 16 - Aprendizagem Não-Supervisionada
11/05/2021	19:00	Aula 17 - Aprendizagem Não-Supervisionada
13/05/2021	19:00	Aula 18 - Projeto da disciplina - Parte 4 - Machine Learning - Clusterização
18/05/2021	19:00	Aula 19 - AutoML
20/05/2021	19:00	Aula 20 - Demonstração de Deploy de Machine Learning

Conteúdo da Aula

1. Métricas para problemas de Regressão
   i. R2 e R2 Ajustado
   ii. MSE
   iii. RMSE
   iv. RMSLE
   v. MAE
   vi. MedAE
   vii. MAPE
2. Algoritmos para problemas de Regressão
   i. Modelos baseados em árvores
3. Exercícios

Material das aulas

- Iremos utilizar o Google Colab para desenvolver os códigos durante as aulas.

- Acesse https://bit.ly/tutorial-colab-projeto para realizar o tutorial de utilização do Google Colab.

1. Métricas para problemas de Regressão

1. Métricas para problemas de Regressão

MÉTRICAS

- Existem diversas métricas que podemos estudar para os problemas de regressão.
- Algumas que iremos estudar:
  - R2 e R2 Ajustado
  - MSE
  - RMSE
  - RMSLE
  - MAE
  - MedAE
  - MAPE

1. Métricas para problemas de Regressão

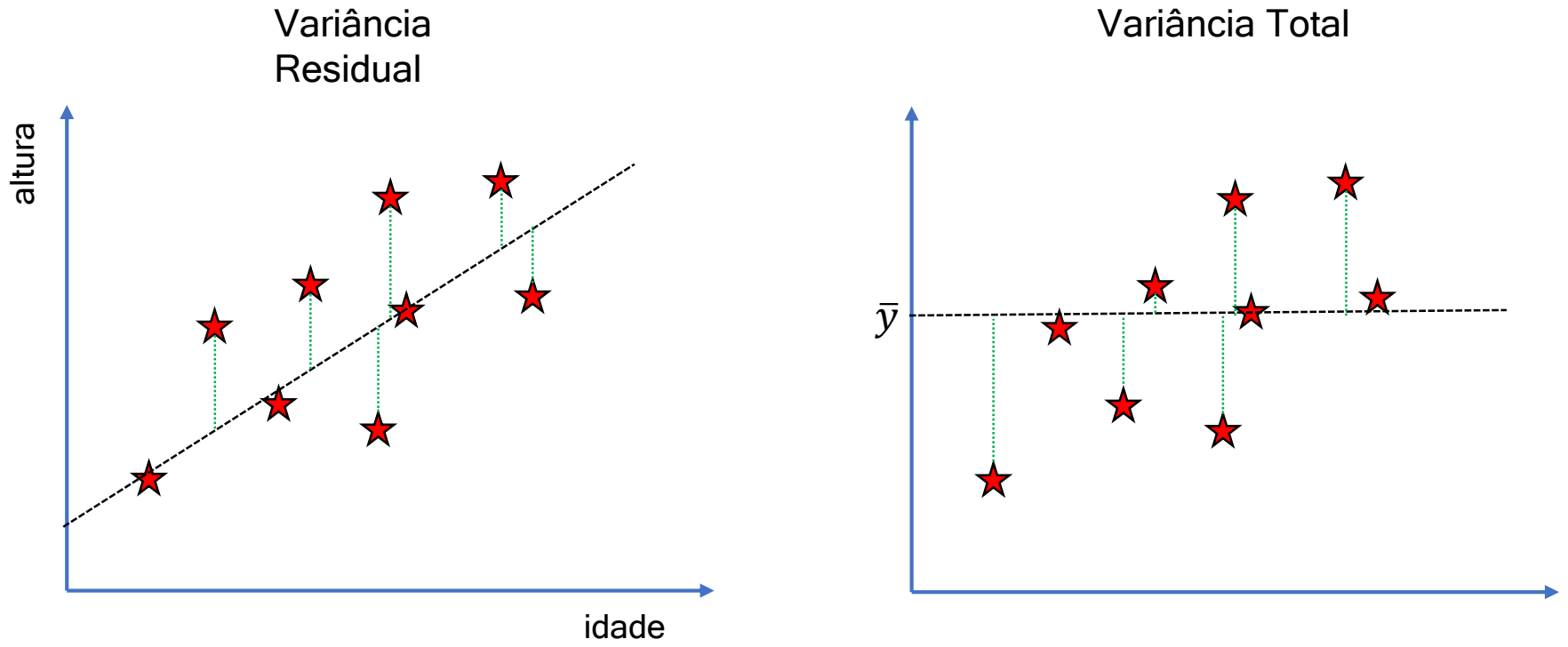
MÉTRICAS

- **R2** ou Coeficiente de determinação, é uma métrica que visa expressar a quantidade da variância dos dados que é explicado pelo modelo construído.
- É uma medida que varia de 0 a 1 e geralmente é representado em porcentagem.

$$R^2 = 1 - \frac{\text{Varianca Residual}}{\text{Varianca Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

1. Métricas para problemas de Regressão

MÉTRICAS

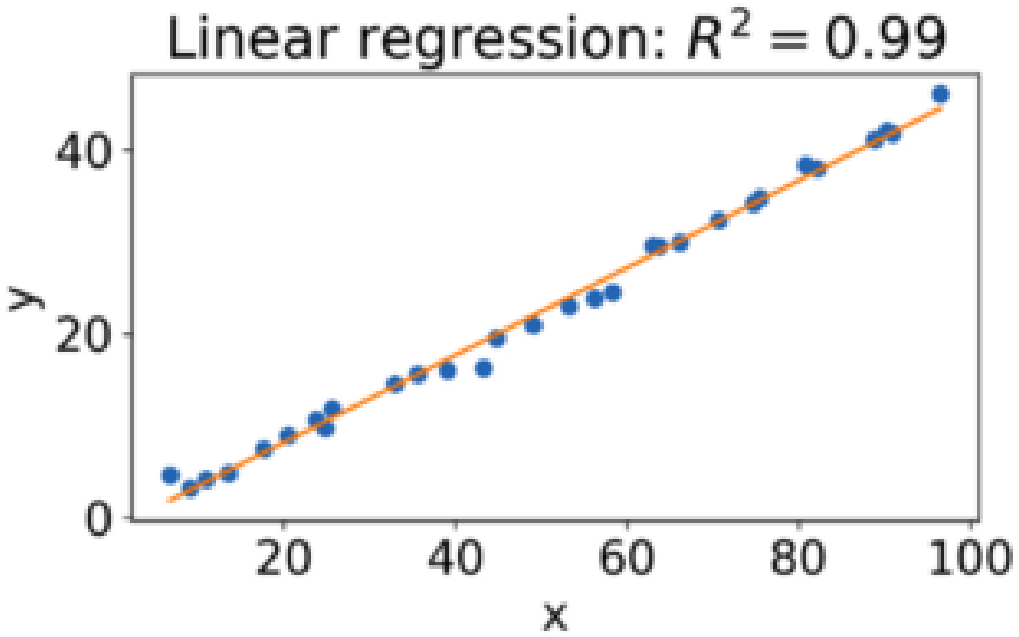


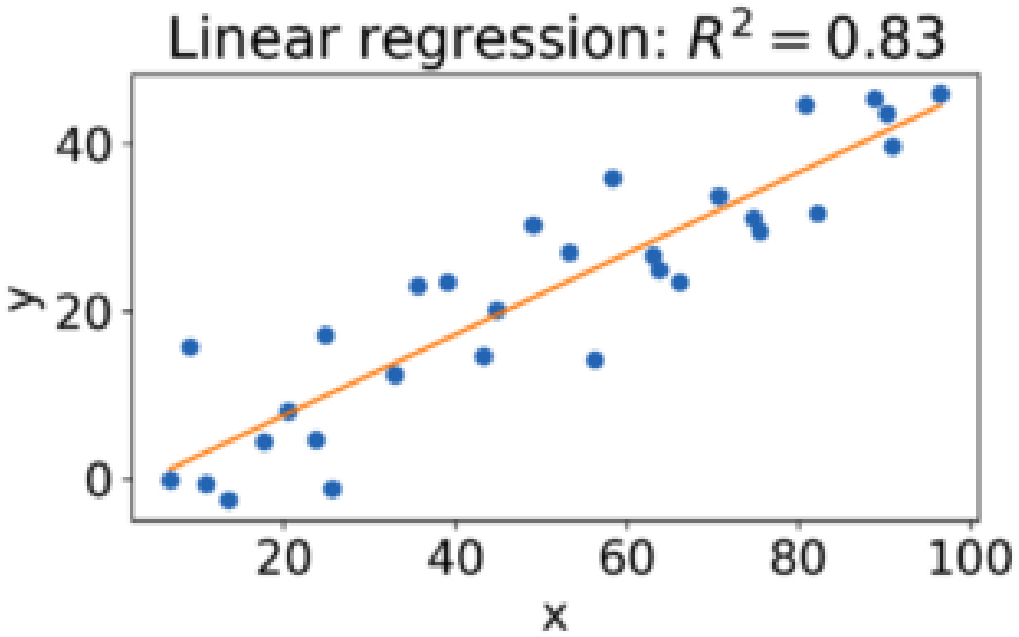
$$R^2 = 1 - \frac{\text{Varianca Residual}}{\text{Varianca Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

1. Métricas para problemas de Regressão

MÉTRICAS

- Podemos ver que quanto mais próximo as predições forem da reta (modelo ajustado) melhor será o R2 (imagem da esquerda).





Fonte: https://bit.ly/32N7tdb

1. Métricas para problemas de Regressão

MÉTRICAS

- Alguns pontos de atenção:
  - Só pode ser aplicado com confiança em modelos com apenas uma feature. À medida que aumentamos a quantidade de features, tendemos a diminuir o viés do modelo (diminuindo a variância residual) e o R2 pode aumentar sem necessariamente estarmos aumentando o poder preditivo do modelo, levando ao overfitting.
  - Em casos de overfitting, o valor dessa métrica continua alta.
  - R2 é enviesado pois os algoritmos de regressão utilizam a correlação dos dados de forma a incrementar o valor do R2 injustamente.

- **Apenas a métrica R2 não consegue indicar se um modelo de regressão é eficiente ou não.**

1. Métricas para problemas de Regressão

MÉTRICAS

```
from sklearn.metrics import r2_score

R2 = r2_score(y_esperado, y_previsto)

print('R2:', R2)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- Como vimos, existem alguns pontos de atenção em relação a métrica R2, desta forma, uma alternativa mais versátil é o **R2 Ajustado**, que funciona como uma calibração do R2 quando temos mais de uma feature.
- O R2 Ajustado busca representar a porcentagem da variância no modelo ajustado, porém essa métrica possui um viés reduzido, uma vez que estamos adicionando novas features (características) ao modelo.
- No R2 Ajustado penalizamos o R2 ponderando pelo número de amostras (*N*) e quantidade de features (*p):*

$$R_a^2 = 1 - (1 - R^2)\frac{(N-1)}{(N-p-1)}$$

1. Métricas para problemas de Regressão

MÉTRICAS

$$R_a^2 = 1 - (1 - R^2)\frac{(N-1)}{(N-p-1)}$$

- N representa o número de amostras (quantidade de linhas)
- p representa o número de features (características)
- Pontos de atenção:
  - Pode ser utilizado para avaliar modelos com mais precisão e segurança
  - É aplicável na avaliação de modelos com mais de uma feature
  - Não apresenta um viés dependente dos dados de entrada.
- Tanto R2, como R2 Ajustado costumam serem mais utilizadas para avaliar relações e modelos mais simples e, em grande maioria, lineares.

1. Métricas para problemas de Regressão

MÉTRICAS

```python
from sklearn.metrics import r2_score

def adjusted_r2(y_esperado, y_previsto, X_treino):

    R2 = r2_score(y_esperado, y_previsto)
    N = len(y_esperado)
    p = X_treino.shape[1]

    r2_ajustado = (1 - ((1 - r2) * (N - 1)) / (N - p - 1))

    return r2_ajustado

r2_ajustado = adjusted_r2(y_esperado, y_previsto, X_treino)
print('R2 Ajustado:', r2_ajustado)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- **MSE** ou *Mean Square Error* (Erro Quadrático Médio) talvez seja uma das métricas mais utilizadas.
- A ideia é calcular a diferença entre o valor previsto e o valor esperado. O MSE pode ser calculado por:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

1. Métricas para problemas de Regressão

- **Pontos de atenção:**
  - Como essa métrica eleva ao quadrado, as predições muito distantes do real aumentam o valor da métrica, ou seja, mse é muito afetado por outliers.
  - É uma métrica interessante para os problemas onde grandes erros não são tolerados, como é o caso de exames médicos e projeções de preços.
  - Por fim, a interpretabilidade direta é um problema, uma vez que os valores previstos estão na unidade u e a medida MSE está em u^2.

1. Métricas para problemas de Regressão

MÉTRICAS

```
from sklearn.metrics import mean_squared_error

mse = mean_squared_error(y_esperado, y_previsto)

print('MSE:', mse)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- **RMSE** (*Root Mean Square Error*) ou Erro quadrático médio resolve o problema da diferença entre as unidades.
- Desta forma essa métrica melhora a interpretabilidade, acertando a unidade.
- Porém, assim como a **MSE** essa métrica medida penaliza as predições com valores muito distantes do esperado (muito afetado por outliers).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

1. Métricas para problemas de Regressão

MÉTRICAS

```
import numpy
from sklearn.metrics import mean_squared_error

mse = mean_squared_error(y_esperado, y_previsto)

rmse = np.sqrt(mse)

print('RMSE:', rmse)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- **RMSLE** (Root Mean Squared Logarithmic Error) ou Erro Médio Quadrático e Logarítmico
- Essa métrica, apesar de apresentar uma fórmula um pouco mais extensa, realiza um cálculo similar ao do RMSE.
- O uso da função logarítmica tem por objetivo evitar a penalização onde os valores da diferenças ente o previsto e esperado são muito grandes.
- É mais robusta a outliers

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - (\log \hat{y}_i + 1))}$$

1. Métricas para problemas de Regressão

MÉTRICAS

- **Pontos de atenção:**
  - Se o valor predito e o real forem valores pequenos → RMSLE == RMSE (aproximadamente)
  - Se apenas um dos dois é grande → RMSE > RMSLE
  - Se ambos os valores são grandes → RMSE > RMSLE

1. Métricas para problemas de Regressão

MÉTRICAS

```
from sklearn.metrics import mean_squared_log_error

rmsle = np.sqrt(mean_squared_log_error(y_esperado, y_previsto))

print('RMSLE:', rmsle)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- **MAE** (*Mean Absolute Error*) ou Erro Absoluto Média consiste na média das distâncias entre valores esperados e previstos.
- Diferente do MSE e do RMSE, ela não penaliza tão severamente os outliers do modelo.
- O valor mínimo é 0 e não tem valor máximo. É calculado por:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

1. Métricas para problemas de Regressão

MÉTRICAS

- **Pontos de atenção:**
  - É um métrica sólida para modelos que devem prever muitos dados ou dados sazonais (por exemplo, número de casos de doenças).
  - Utilizar a mesma unidade de medida u, diferente do MSE.
  - Usa o valor absoluto em contraste com o MSE, por isso sofre menos os efeitos dos *outliers*.

1. Métricas para problemas de Regressão

MÉTRICAS

```
from sklearn.metrics import mean_absolute_error

mae = mean_absolute_error(y_esperado, y_previsto)

print('MAE:', mae)
```

1. Métricas para problemas de Regressão

MÉTRICAS

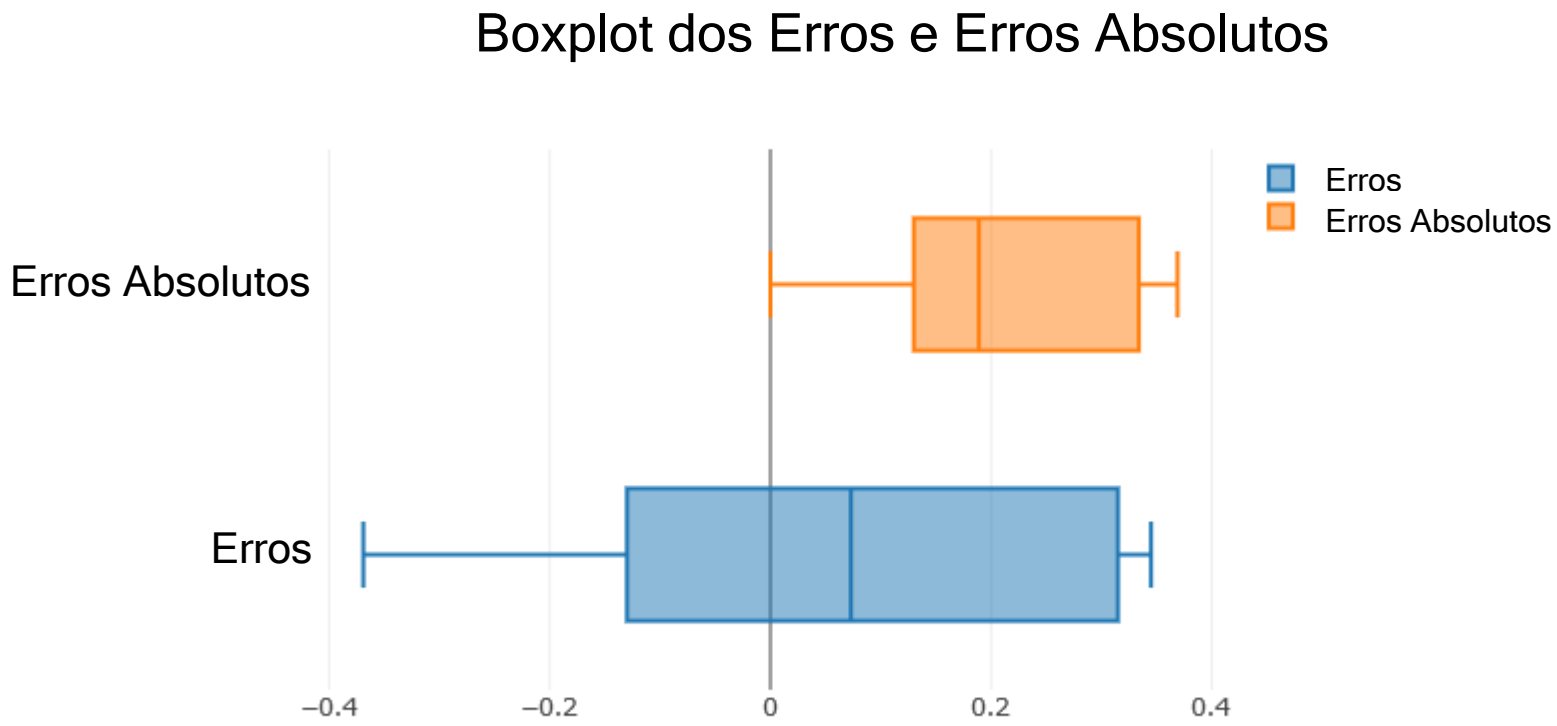
- MedAE (Median Absolute Error) ou Erro Mediano Absoluto
- É uma métrica interessante pois é robusto a *outliers*.
- A função de custo é calculado computando a mediana de todas as diferenças absolutas entre o esperado e o previsto.

$$MedAE = median(|y_i - \hat{y}_i|)$$

- Essa métrica é menos utilizada, ela dá alguma indicação na distribuição de erros absolutos, especialmente se apresentado em um Boxplot.

1. Métricas para problemas de Regressão

MÉTRICAS



Fonte: https://bit.ly/32Qd9mG

1. Métricas para problemas de Regressão

MÉTRICAS

```
from sklearn.metrics import median_absolute_error

medae = median_absolute_error(y_esperado, y_previsto)

print('MEDAE:', medae)
```

1. Métricas para problemas de Regressão

MÉTRICAS

- MAPE (Mean Absolute Porcentage Error) ou Erro Percentual Absoluto médio calcula a porcentagem obtida através da divisão da diferença entre predito (ŷ) e real pelo valor real (y).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

1. Métricas para problemas de Regressão

MÉTRICAS

- **Pontos de atenção:**
  - Por se tratar de uma porcentagem, essa métrica torna-se **extremamente intuitiva**. Por exemplo, ter um MAPE=12% significa que, em média, nosso modelo faz previsões que erram por 12% do valor real.
  - Devido a sua formulação, essa métrica não lida tão bem se tratando de problemas com uma variação muito grande no alvo, como por exemplo, uma regressão que prevê uma variável que vai e 1 a 2.000.
  - Em linhas gerais, podemos considerar:
    - Menor que 10% é uma predição alta (muito boa)
    - Entre 10% e 20% é uma predição boa
    - Entre 20% e 50% é uma predição razoável
    - Mais que 50% é imprecisa

1. Métricas para problemas de Regressão

MÉTRICAS

```
import numpy

def mape(y_esperado, y_previsto):
    return np.mean(np.abs((y_esperado - y_previsto) / y_esperado))

print('MAPE:', mape(y_esperado, y_previsto))
```

2. Algoritmos para Problemas de Regressão

2. Algoritmos para problemas de Regressão

MODELOS BASEADOS EM ÁRVORES

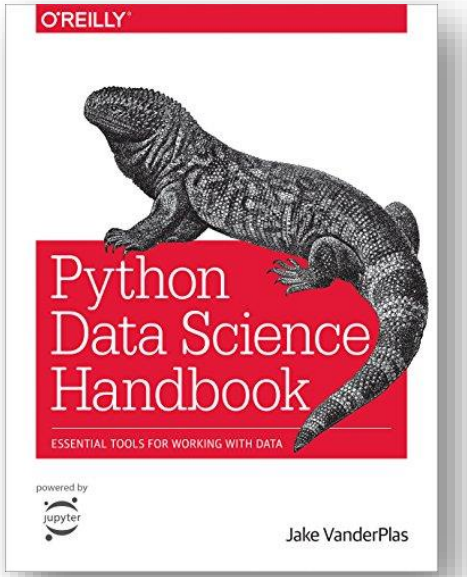
- Vamos testar algoritmos baseados em árvore para problemas de regressão.
- Todos os algoritmos baseados em árvores que vimos, além da versão para classificação, também possuem versões para regressão.
- A principal diferença entre árvores de decisão e regressão é que nas folhas da árvore será previsto um valor contínuo ao invés de categórico.
- Vamos testar as seguintes algoritmos de regressão baseados em árvore:
  - `DecisionTreeRegressor`
  - `RandomForestRegressor`
  - `LGBMRegressor`
  - `XGBRegressor`
  - `CatboostRegressor`

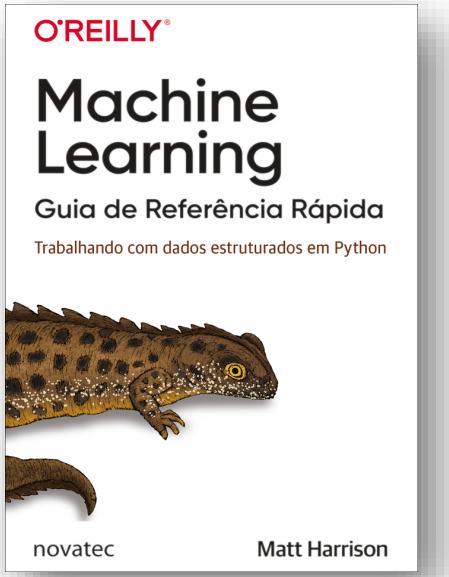
</> Abra o arquivo **"aula14-parte2-modelagem-regressao.ipynb"**

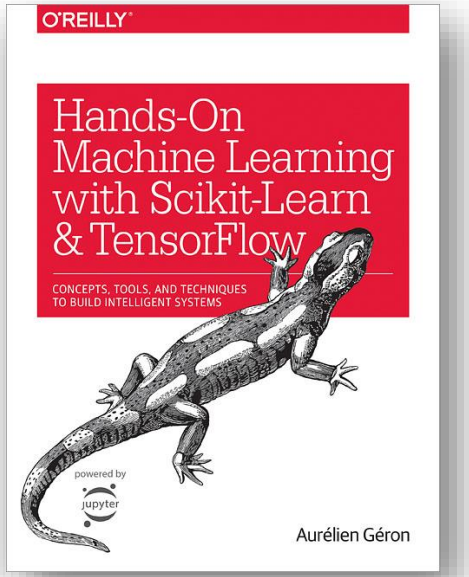
Referências Bibliográficas

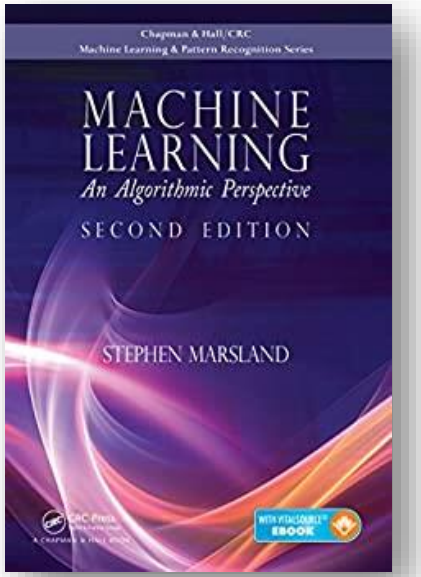
Referências Bibliográficas

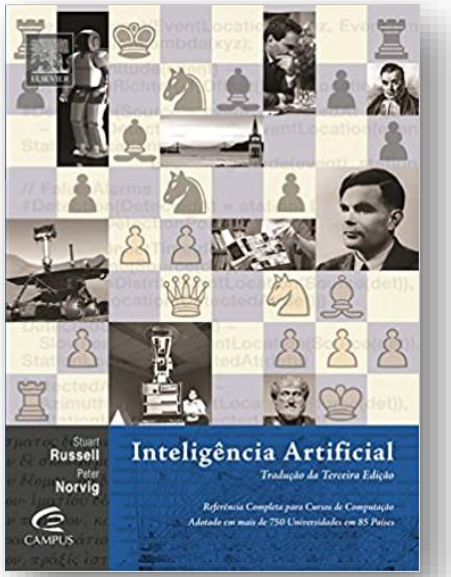
LIVROS





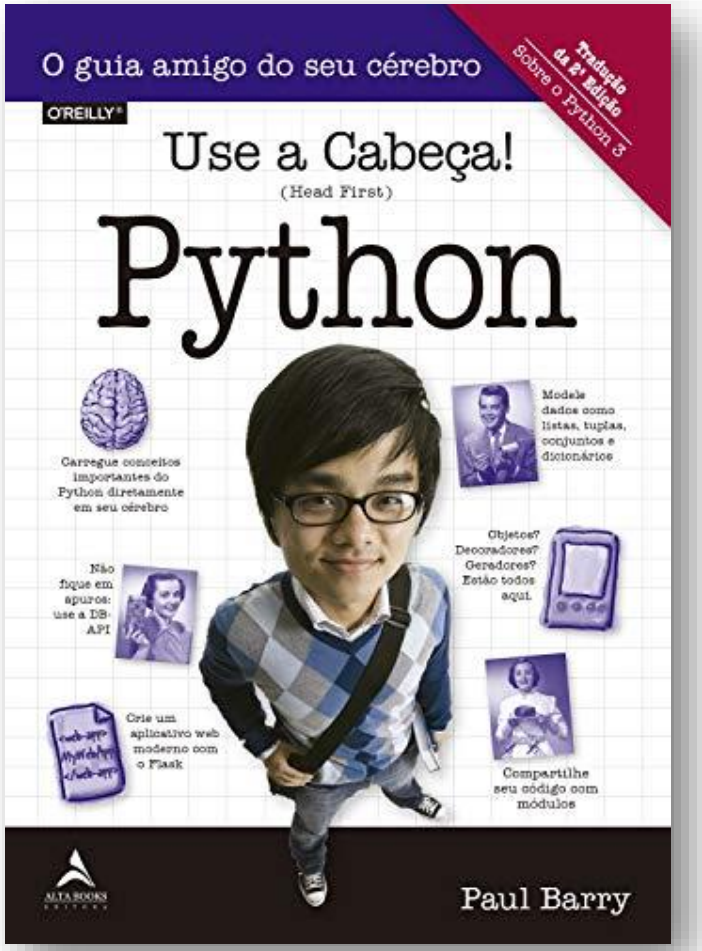


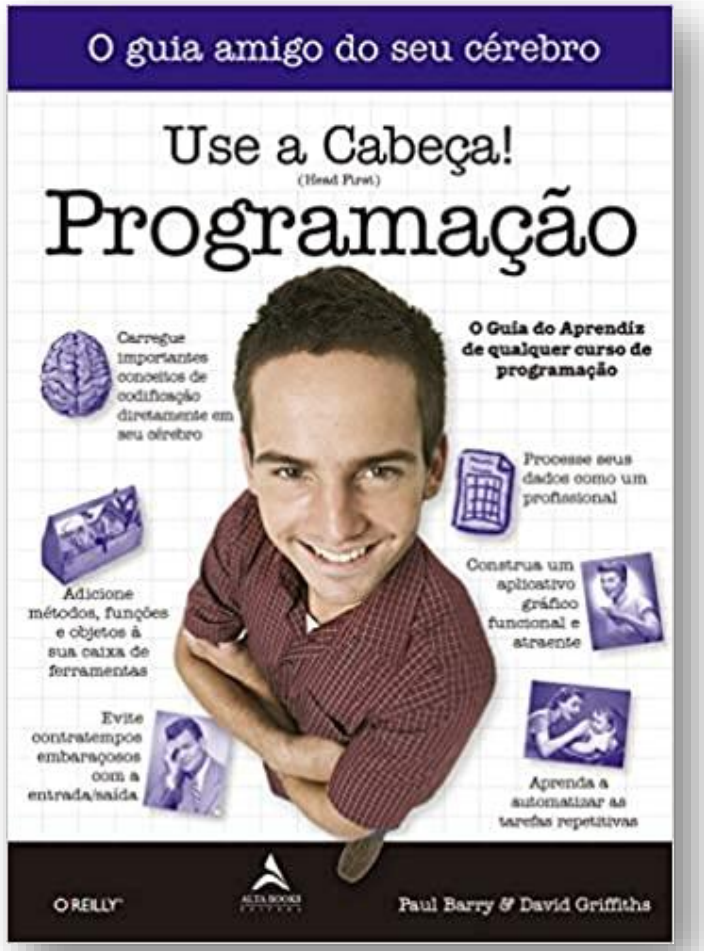


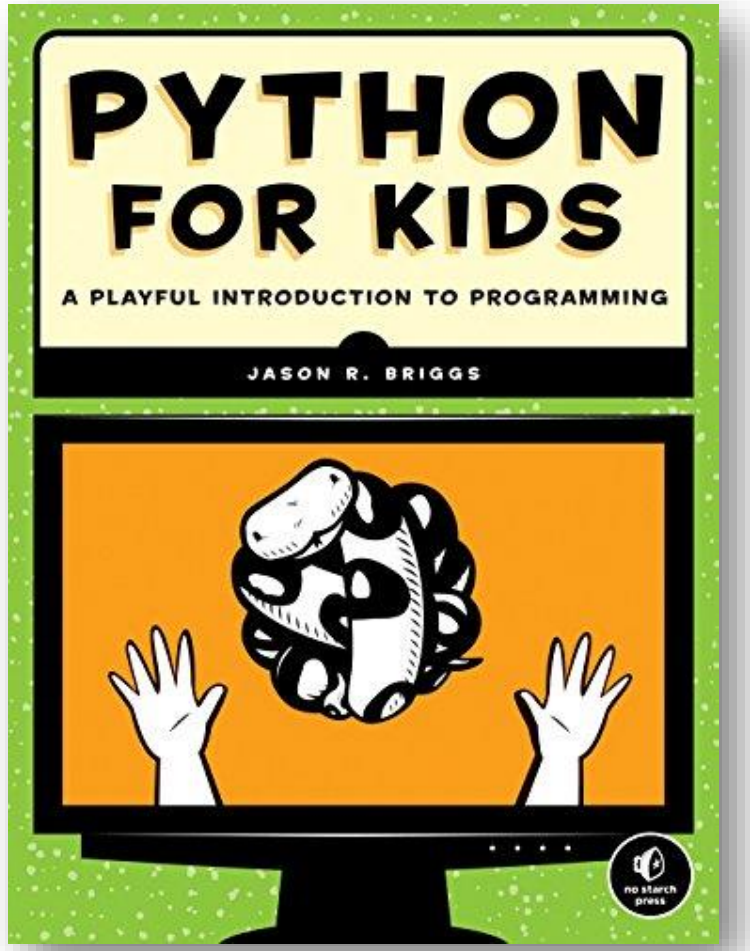


Referências Bibliográficas

LIVROS

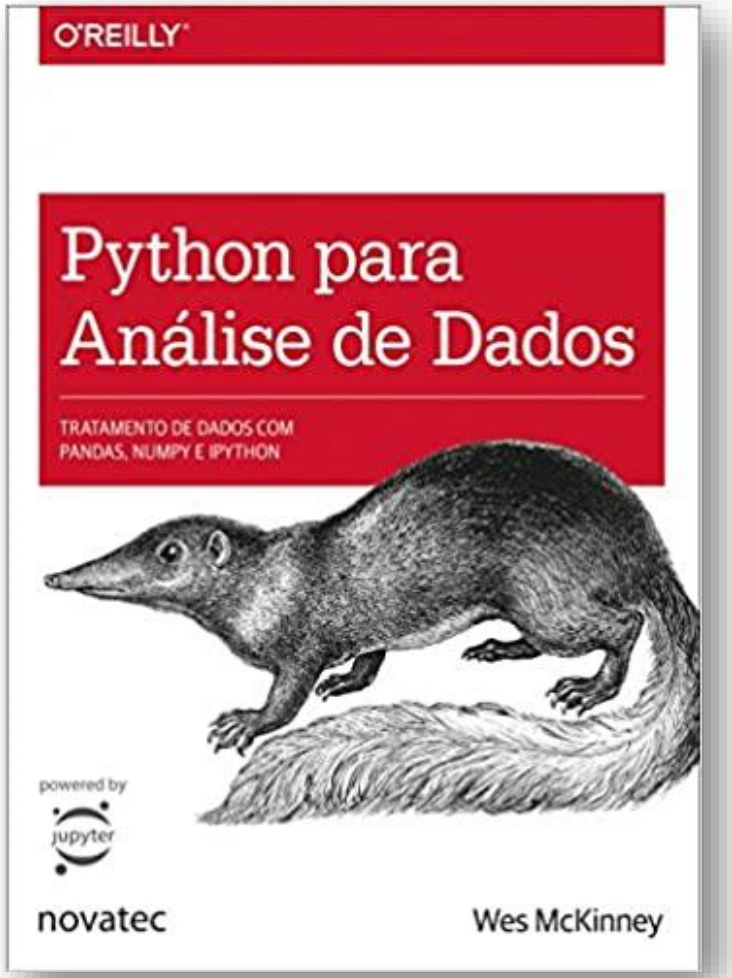


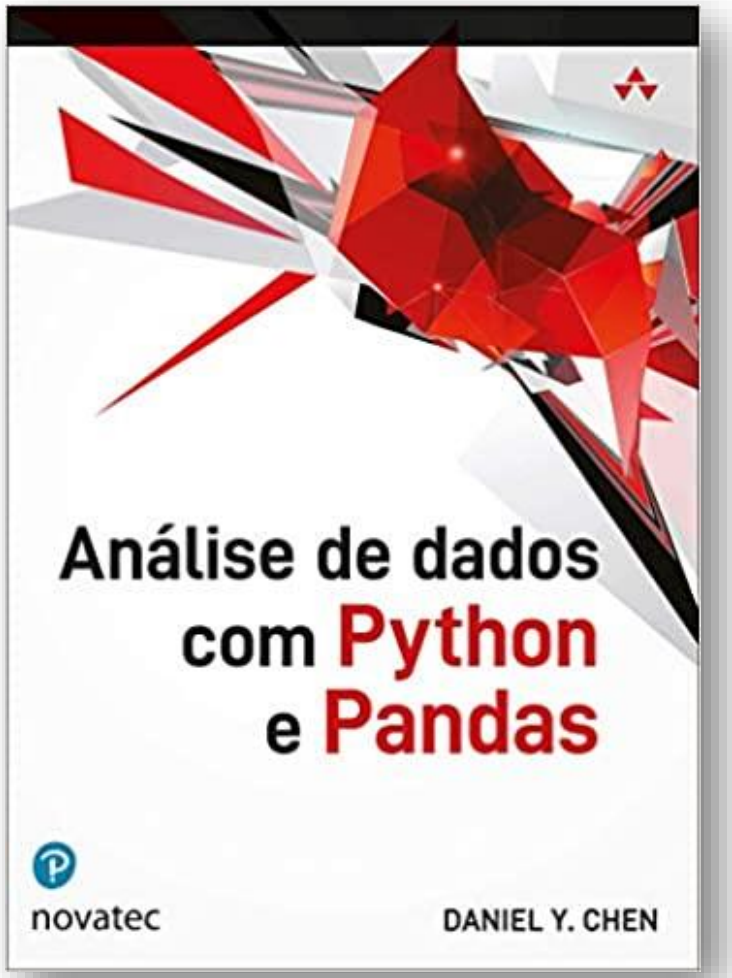


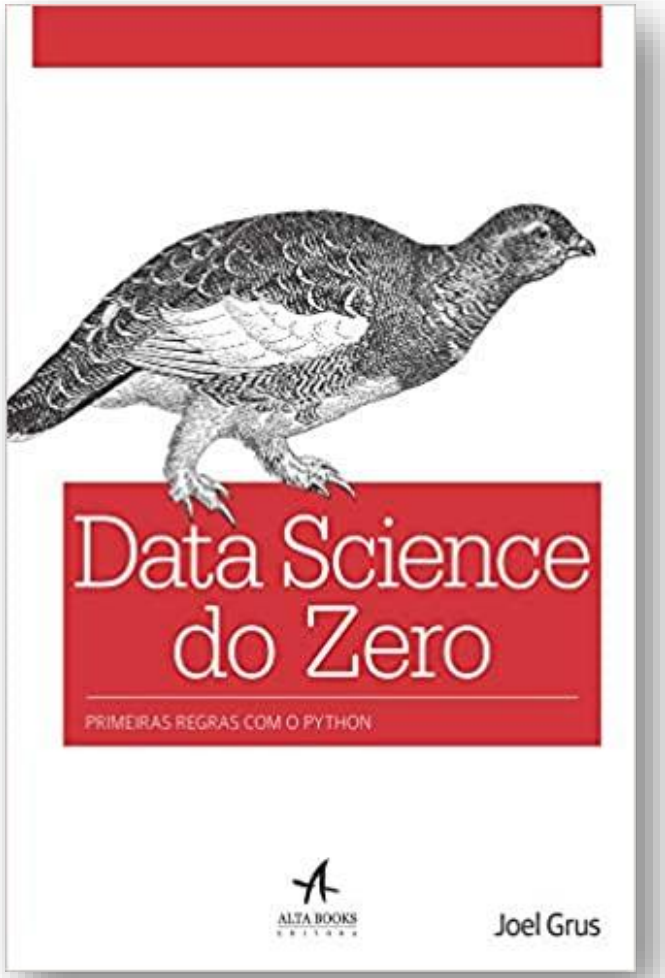


Referências Bibliográficas

LIVROS

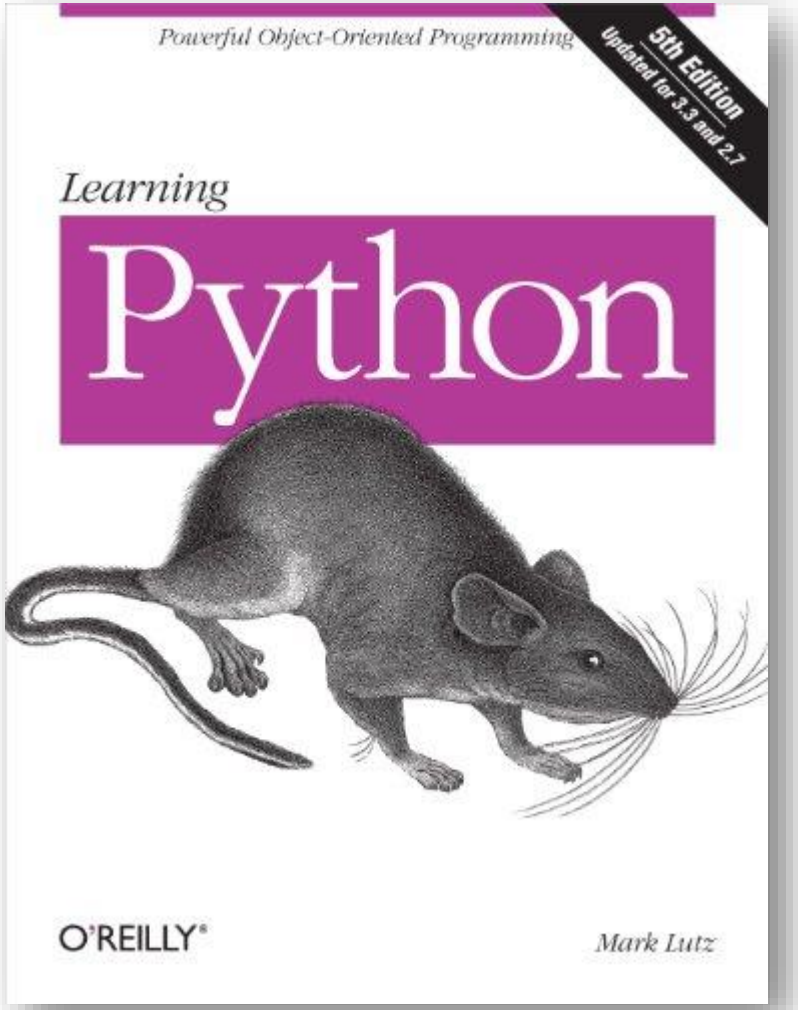


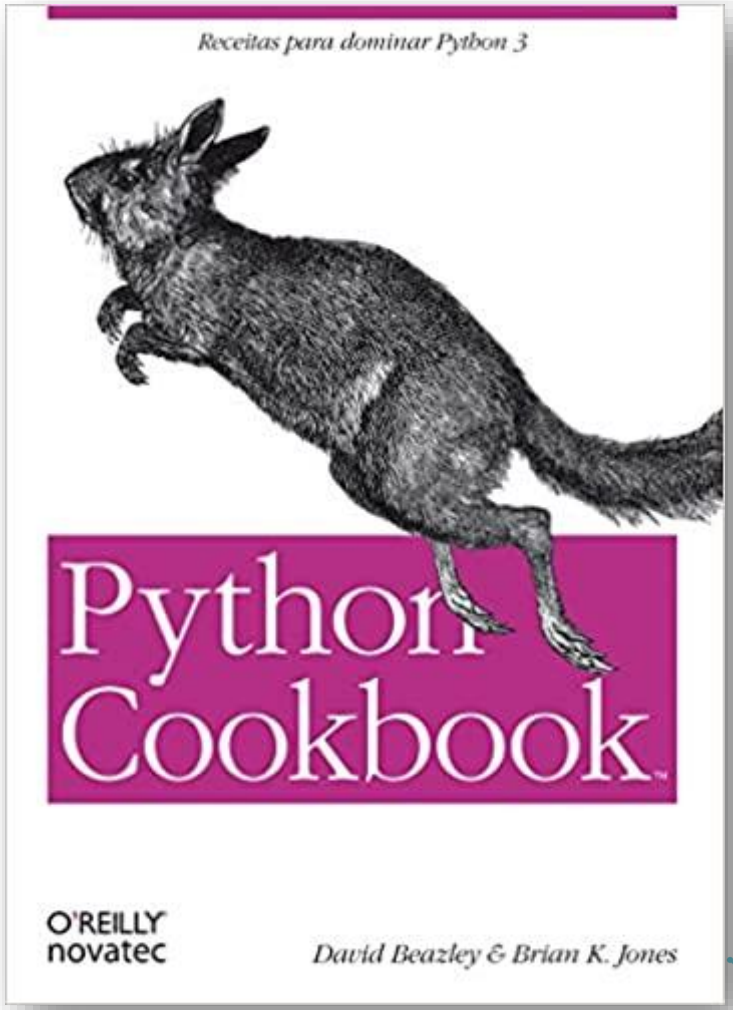




Referências Bibliográficas

LIVROS





Referências Bibliográficas

LINKS, ÍCONES, IMAGENS

- As referências de links utilizados podem ser visualizados em http://urls.dinomagri.com/refs
- Tutoriais disponíveis no site oficial do Pandas - http://pandas.pydata.org/pandas-docs/stable/
- Livro de receitas disponíveis no site oficial do Pandas - http://pandas.pydata.org/pandas-docs/stable/cookbook.html

- As imagens foram Icon made by Srip, Pixel perfect, Eucalyp, Nhor-Phai e Prettycons from www.flaticon.com