Analytics e Inteligência Artificial

Aula 16

Aprendizagem Não-Supervisionada

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. Aprendizagem Não-Supervisionada
   i. O que é?
   ii. Segmentação (Clustering)
2. Algoritmos de Clusterização
   i. KMeans
   ii. Cluster Hierárquico
   iii. DBSCAN
3. Métricas de avaliação da qualidade do ajuste
   i. Método do Cotovelo
   ii. Método da Silhueta
   iii. Mapa de Calor
   iv. Surrogate Tree
4. 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. Aprendizagem Não-Supervisionada

1. Aprendizagem Não-Supervisionada

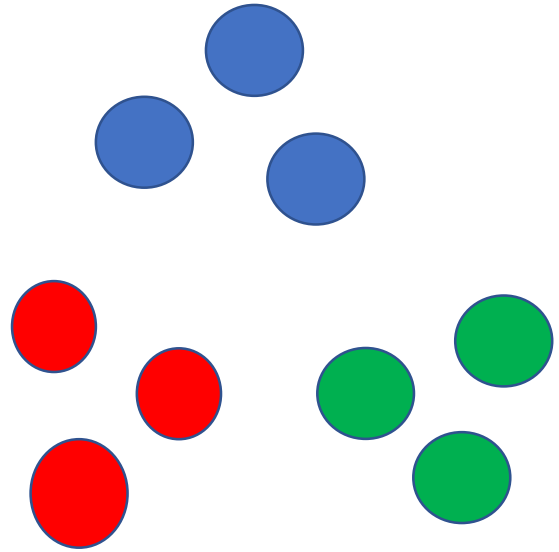
O QUE É?

- É uma forma de aprendizagem de máquina em que não temos os alvos (target ou labels) ou marcações para cada amostra dos nossos dados.
- Podemos dizer que deixamos os dados falarem por si só. Temos dois tipos comuns de aprendizagem não-supervisionada:
  - Segmentação
  - Redução de Dimensionalidade

1. Aprendizagem Não-Supervisionada

SEGMENTAÇÃO

- Queremos dividir os nossos objetos/dados baseados em features (características).
- Podemos fazer isso na mão, utilizando várias regras e if-elses? Podemos!
- Mas o objetivo do aprendizado não-supervisionado é deixar que a máquina escolha a melhor forma de se fazer essa divisão dos dados!
- Exemplos:
  - Segmentação de Clientes
  - Compressão de Imagem
  - Detecção de anomalias



2. Algoritmos de Clusterização

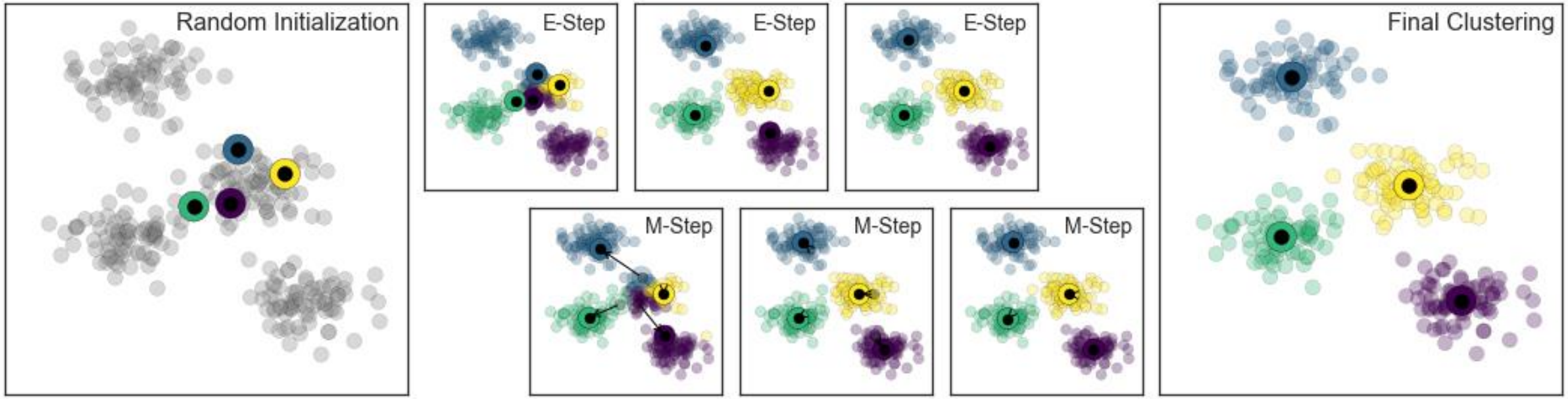
2. Algoritmos de Clusterização

KMEANS

- **K-Means** encontra clusters nos dados usando um algoritmo chamado expectation-maximization (E-M).
- Esse algoritmo consiste dos seguintes passos, descritos abaixo em escrito e na figura ao lado.
    1. Chuta de forma aleatória alguns clusters
    2. Repete até convergir (quando a soma dos quadrados internas a cada cluster não muda mais):
        1. E-Step: atribui os data points ao cluster com centro mais próximo
        2. M-Step: calcula o novo centro do cluster

2. Algoritmos de Clusterização

KMEANS

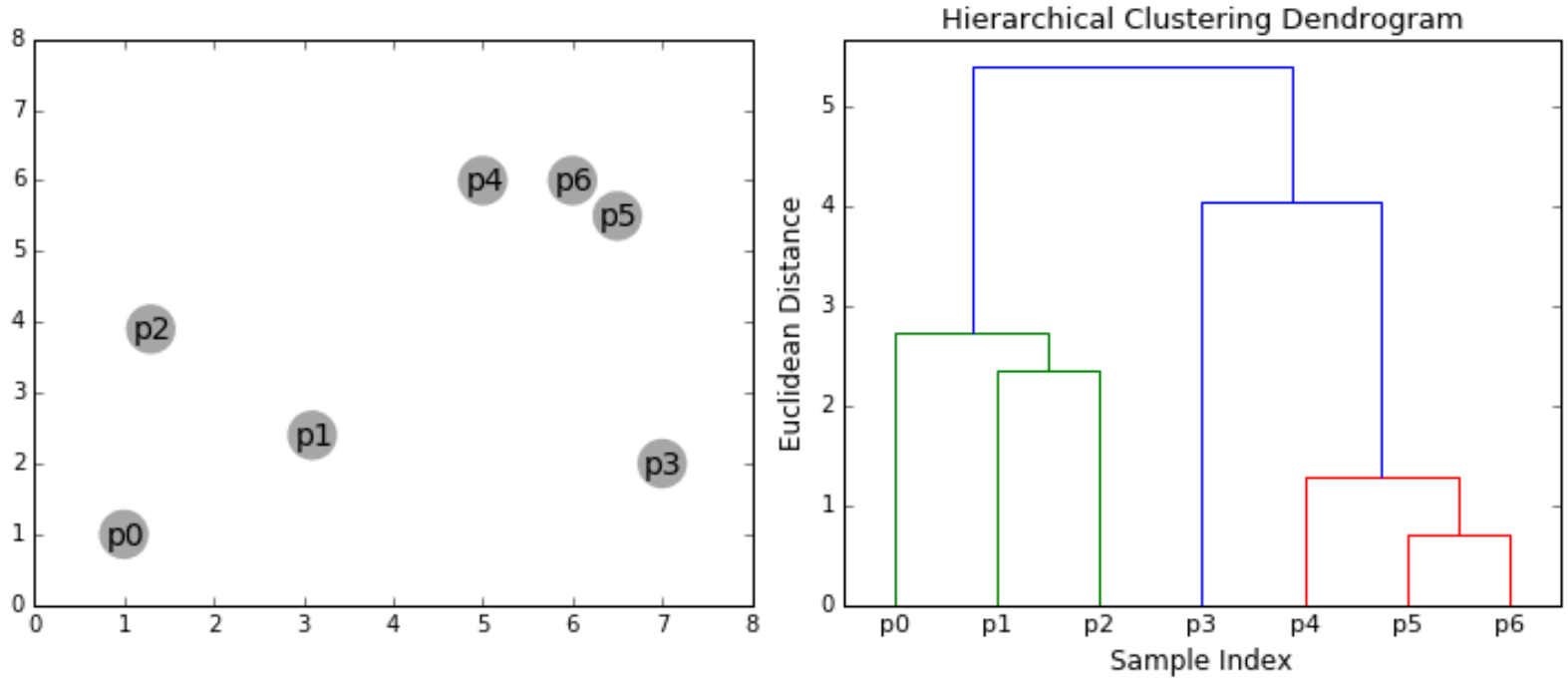


Fonte: [Python For Data Science Handbook]

2. Algoritmos de Clusterização

- CLUSTER HIERÁRQUICO

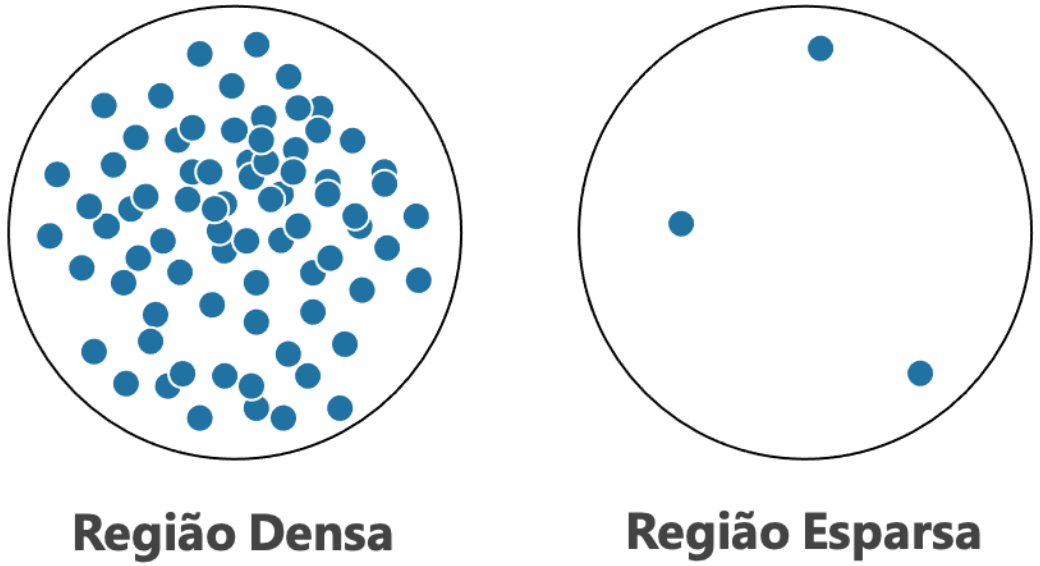
- O **Cluster Hierárquico** trata cada amostra na base de dados como um único cluster e agrupa as amostras baseado em uma medida de similaridade que com muita frequência é a distância euclidiana.



2. Algoritmos de Clusterização

DBSCAN

- Antes de entender como o algoritmo **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*) funciona, precisamos entender de maneira geral como funcionam os algoritmos de clusterização baseados em **densidade**.

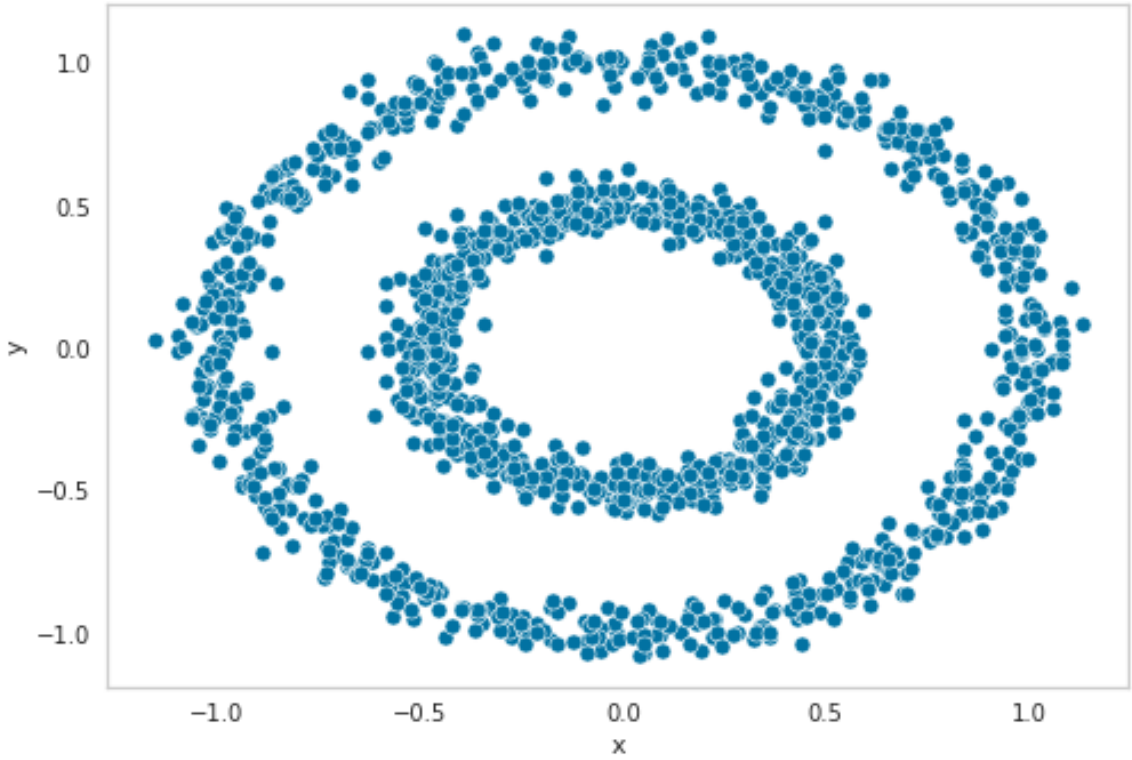


2. Algoritmos de Clusterização

DBSCAN

- Os algoritmos de clusterização baseada em densidade identificam os clusters distintos no conjunto de dados com a ideia de que um cluster no espaço de dados é uma região **adjacente de alta densidade de pontos**, separada de outros clusters por regiões **adjacentes de baixa densidade de pontos**.

- Como podemos visualizar existem diversos pontos adjacentes que contém alta densidade de pontos (círculo interno e externo), sendo separado por regiões adjacentes de baixa densidade de pontos.



2. Algoritmos de Clusterização

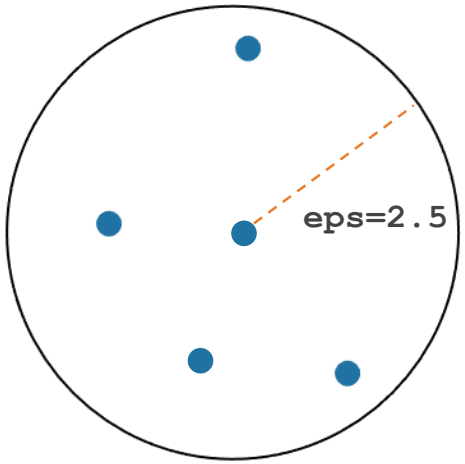
DBSCAN

- Desta forma, o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) poderá descobrir clusters com diferentes formas e tamanhos em uma grande quantidade de dados, inclusive em conjuntos de dados que contêm ruídos.
- O algoritmo DBSCAN utiliza dois parâmetros:
  - **`minPts`** - É o número mínimo de pontos necessários para formar um cluster denso.
  - **`eps (ε)`** - Épsilon é a distância máxima (euclidiana) entre um par de pontos. Os dois pontos são considerados vizinhos se e apenas se eles são separados por uma distância menor ou igual ao valor do épsilon da distância que será utilizada para localizar os pontos vizinhos de cada ponto.
  - **`metric`** - É a métrica utilizada para calcular a distância entre os pontos (por exemplo, distância euclidiana)

2. Algoritmos de Clusterização

DBSCAN

- Vamos entender melhor esses parâmetros:
  - `minPts=4`
  - `eps=2.5`



- Como a Densidade = 5 e ela é maior que o minPts, a região é considerada densa.

- Assim, cada ponto de dados poderá ser classificado em:
  - `Core Point`
  - `Border Point`
  - `Noise Point`

Fonte: https://bit.ly/3lw5K2d

2. Algoritmos de Clusterização

DBSCAN

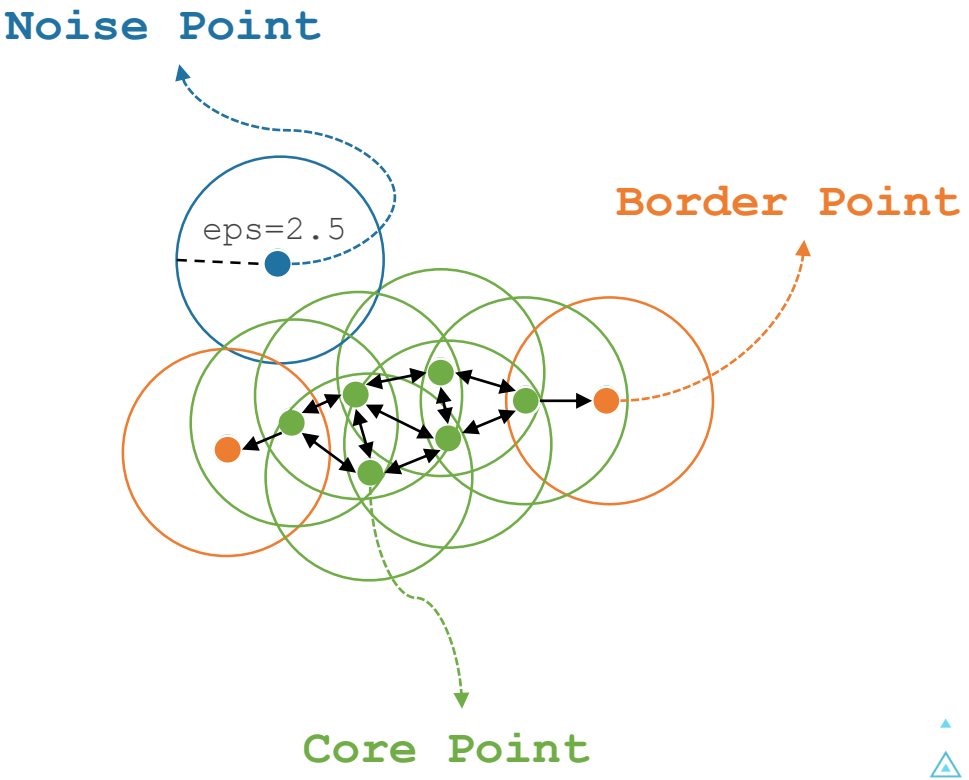
- `Core Point`
  - Um ponto é considerado **Core Point** se tiver **pelo menos** `minPts` dentro de um raio `eps` de distância a partir dele mesmo.
  - Sempre pertence a uma região densa.
- `Border Point`
  - Um ponto é considerado **Border Point** se tiver pelo menos um `Core Point` dentro de um raio `eps` de distância a partir dele mesmo.
- `Noise Point`
  - **Noise Point** é qualquer ponto que não seja um `Core Point` ou `Border Point`. E tenha uma quantidade de pontos a um raio `eps` de distância menor que `minPts` a partir dele mesmo.

Fonte: https://bit.ly/3lw5K2d

2. Algoritmos de Clusterização

DBSCAN

- Considere os dados abaixo e os valores para os parâmetros:
  - `minPts=4`
  - `eps=2.5`
- **`Core Point`** – faz parte do raio do épsilon e passa no critério do `minPts`.
- **`Border Point`** – ainda faz parte do cluster pois está dentro do raio do épsilon de um Core Point, porém não passa pelo critério do `minPts`.
- **`Noise Point`** – não é atribuído a um cluster.

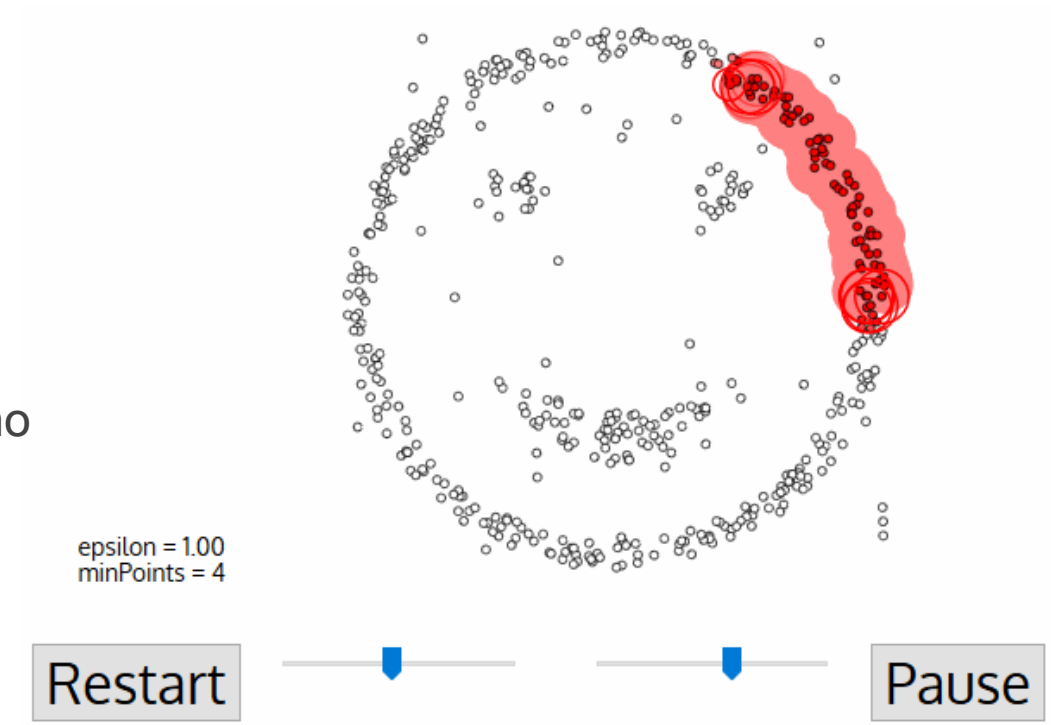


Fonte: https://bit.ly/3lw5K2d

2. Algoritmos de Clusterização

DBSCAN

- Os passos para o algoritmo **DBSCAN:**
  - O algoritmo seleciona arbitrariamente um ponto no conjunto de dados (até que todos os pontos sejam visitados).
  - Se existe pelo menos `minPts` dentro do raio épsilon até o ponto, então considere todos esses pontos como parte do mesmo cluster.
  - Os cluster são então expandidos repetindo recursivamente o cálculo da vizinhança para cada ponto vizinho.



Fonte: https://bit.ly/3lw5K2d, https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

2. Algoritmos de Clusterização

DBSCAN

- Sugestões de valores para os parâmetros `minPts` e `eps`:
  - `minPts` - Uma regra que podemos utilizar para definir a quantidade mínima de pontos é utilizar a quantidade de dimensões (D) que existe no conjunto de dados mais um, ou seja, `minPts >= D + 1`
    - Se `minPts = 1` não faz sentido, uma vez que cada ponto já será um cluster.
    - Se `minPts = 2` o resultado será o mesmo do agrupamento hierárquico com a métrica single-link e o corte no dendrograma na altura do valor do épsilon.
    - Portanto, `minPts` tem que ser igual ou maior que 3.

Fonte: https://bit.ly/3lw5K2d

2. Algoritmos de Clusterização

DBSCAN

- Sugestões de valores para os parâmetros `minPts` e `eps`:
  - `eps` poderá ser escolhido usando um gráfico de k-distâncias, onde iremos calcular a distância para o vizinho mais próximo (`k = minPts - 1`) ordenado do maior para o menor valor.
    - Bons valores de épsilon estão onde o gráfico apresenta um cotovelo.
    - Se `eps` for muito pequeno, uma grande parte dos dados não será agrupada
    - Se `eps` for muito grande, a maioria dos pontos estarão em um mesmo cluster.
    - Em geral valores pequenos do `eps` são desejados.

Fonte: https://bit.ly/3lw5K2d

2. Algoritmos de Clusterização

DBSCAN

- Vantagens do DBSCAN
  - É robusto para outliers.
  - É ótimo para separar clusters com alta densidade dos de baixa densidade
  - Não tem a necessidade de definir o número de clusters.

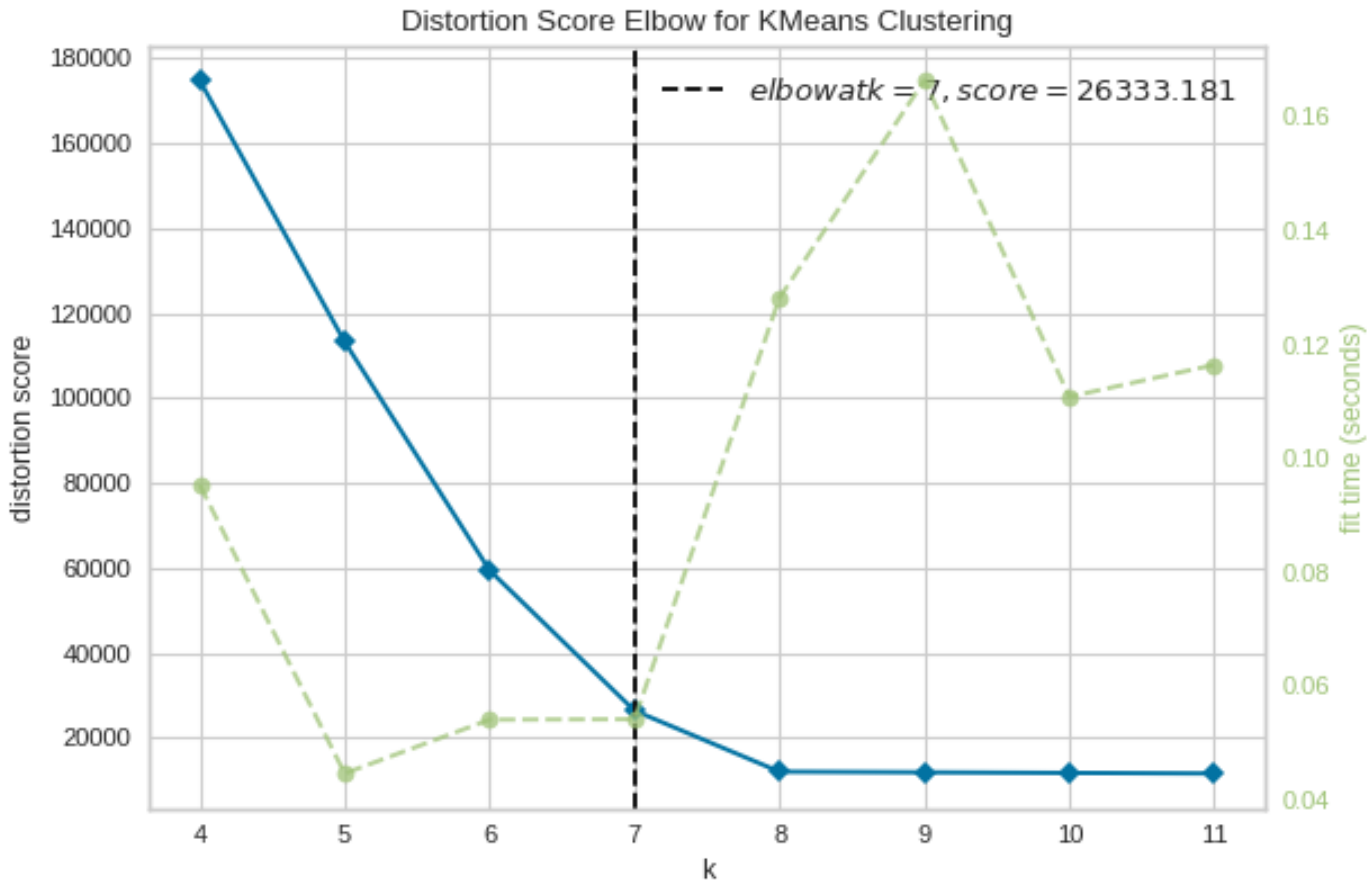
- Desvantagens do DBSCAN
  - Não funciona bem se a densidade é muito similar entre os clusters.
  - Escolher o valor para o épsilon pode ser difícil quando os dados estão em dimensões superiores.

Fonte: https://bit.ly/3lw5K2d

3. Métricas de avaliação da qualidade do ajuste

3. Métricas de avaliação da qualidade do ajuste

MÉTODO DO COTOVELO



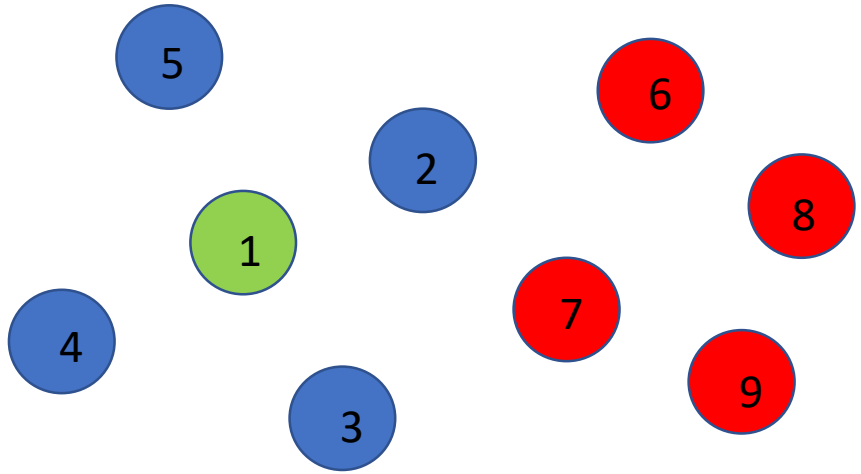
Fonte: [yellowbricks docs]

3. Métricas de avaliação da qualidade do ajuste

MÉTODO DA SILHUETA

- Calcula um score para cada ponto que compara a sua distância intra-cluster com a distância inter-cluster. Como exemplo, vamos calcular o silhueta score para o ponto verde que pertence ao cluster azul.

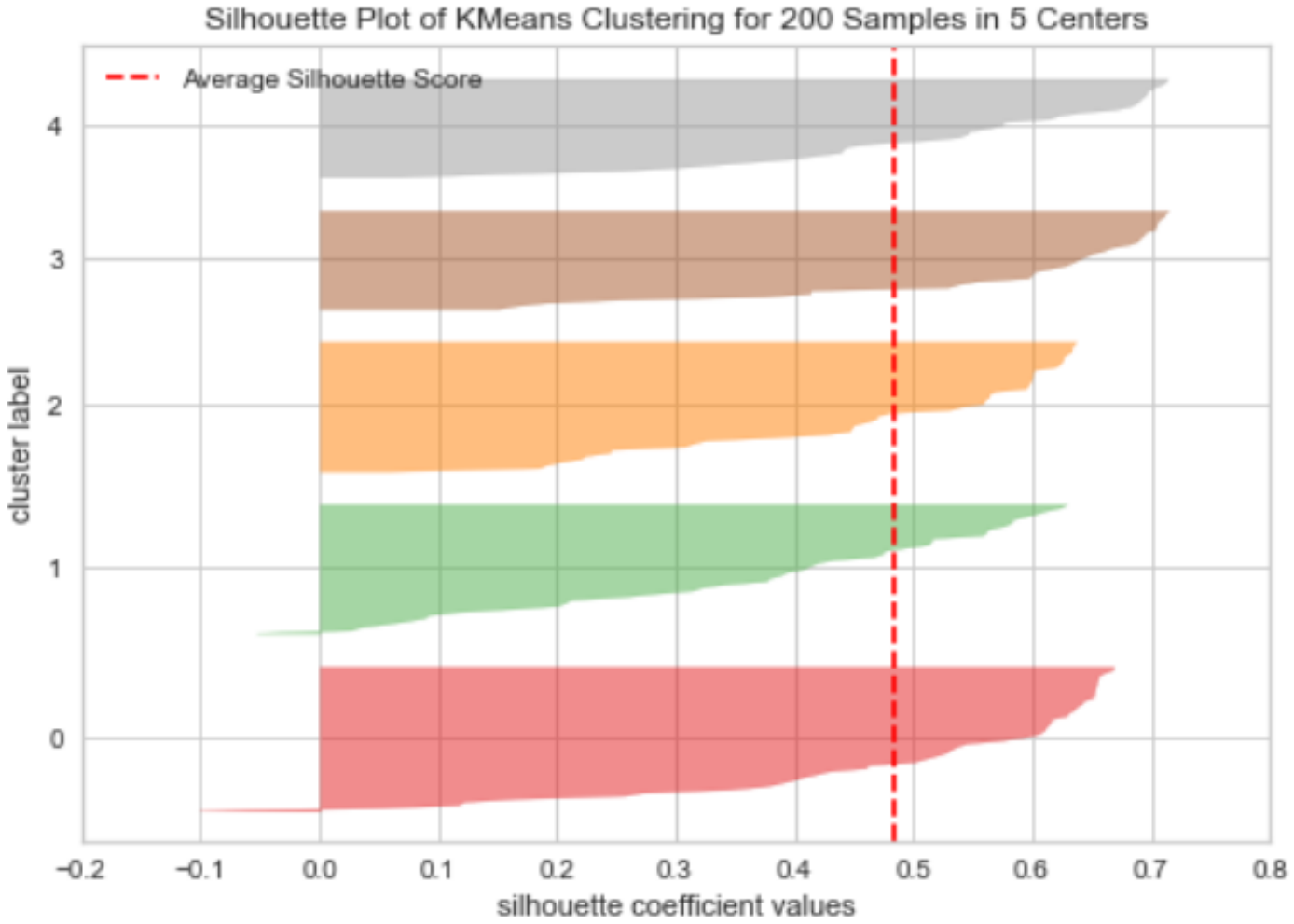
- Distância Intra-Cluster: `A = (d12 + d13 + d14 + d15) / 4`
- Distância Inter-Cluster: `B = (d16 + d17 + d18 + d19) / 4`
- Silhueta Score = `(B - A) / MAX(B,A)`



- Logo, quanto mais próximo de 1, melhor! Próximo de 0 temos que não conseguimos detectar bem os clusters e -1 temos que os clusters foram detectados de forma errada!

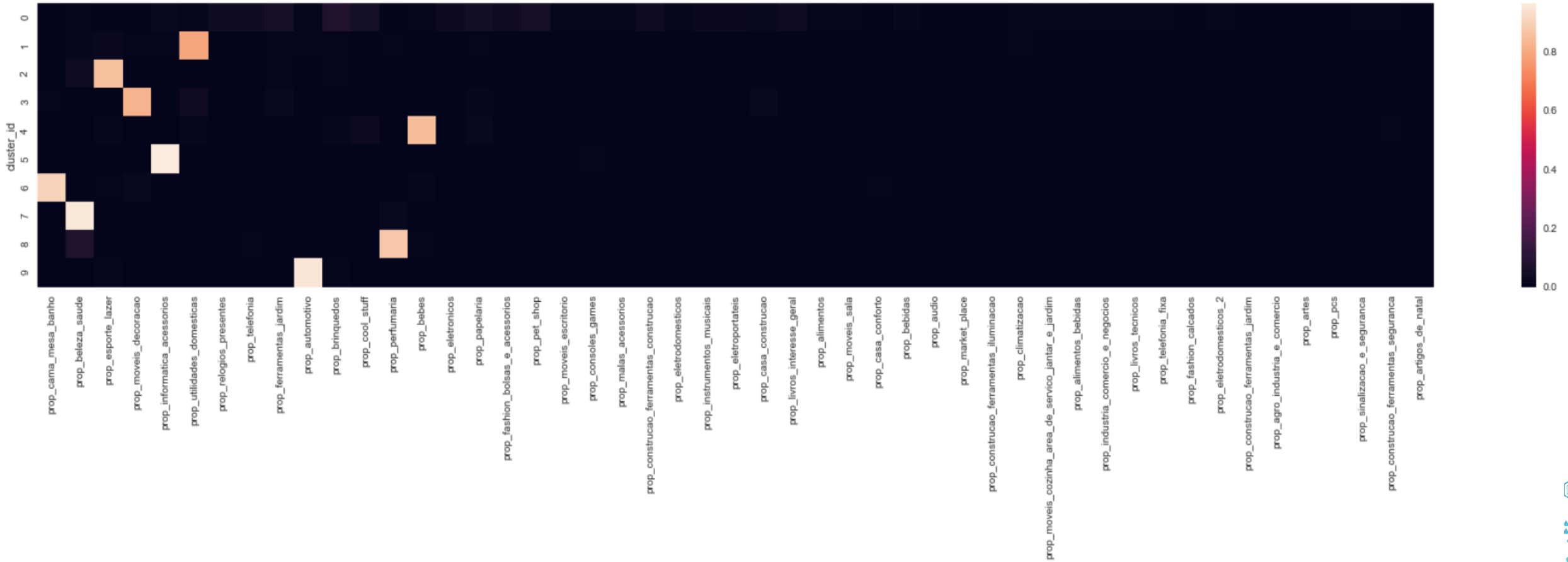
3. Métricas de avaliação da qualidade do ajuste

MÉTODO DA SILHUETA



3. Métricas de avaliação da qualidade do ajuste

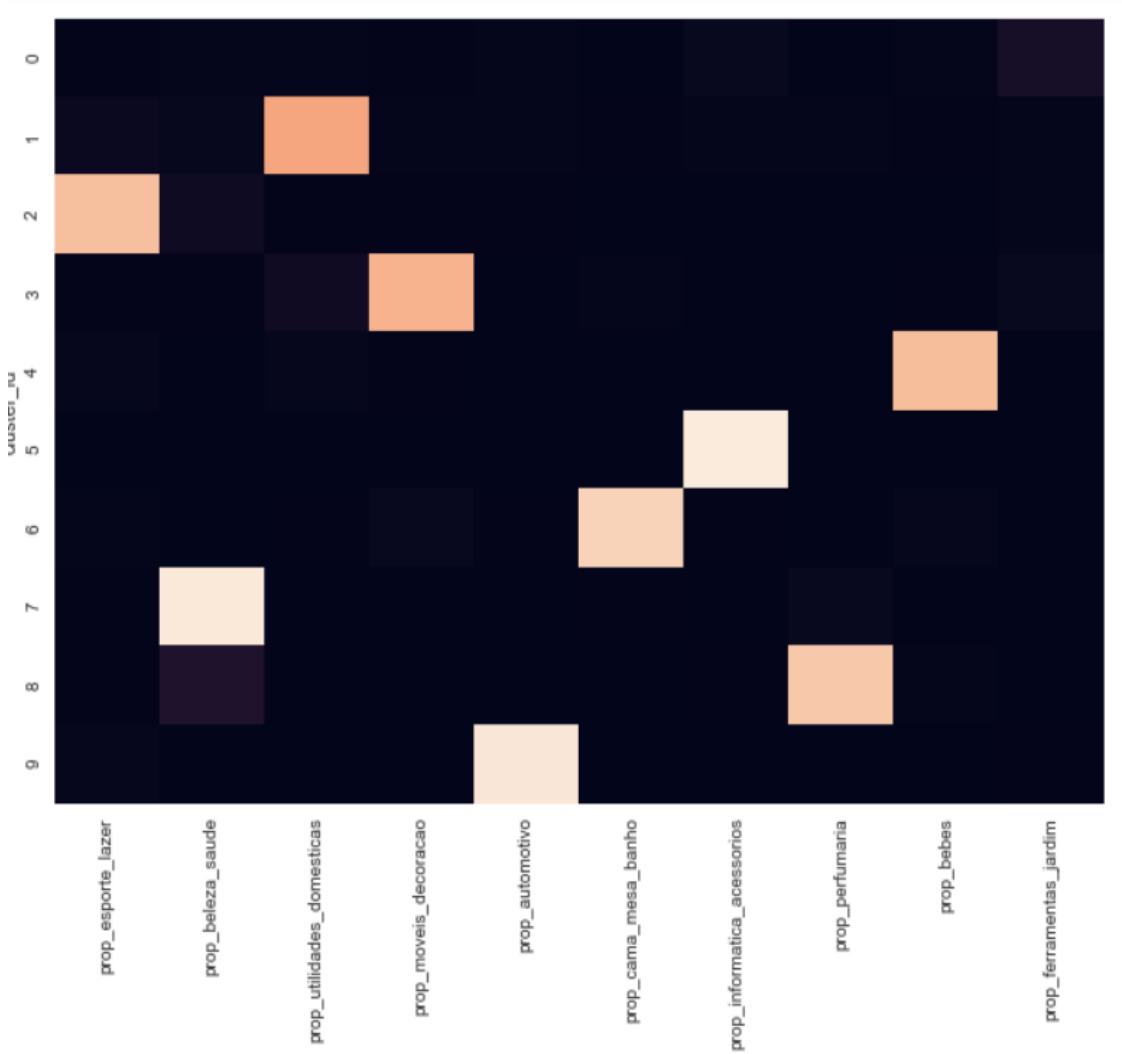
MAPA DE CALOR



3. Métricas de avaliação da qualidade do ajuste

SURROGATE TREE

- O objetivo aqui é treinarmos uma árvore de decisão com os clusters como *labels* e usar a árvore de decisão como modelo para selecionar as *features* mais importantes.



3. Métricas de avaliação da qualidade do ajuste

PRÁTICA

</> Abra o arquivo **"aula16-parte1-clusterização.ipynb"**

4. Exercícios

4. Exercícios

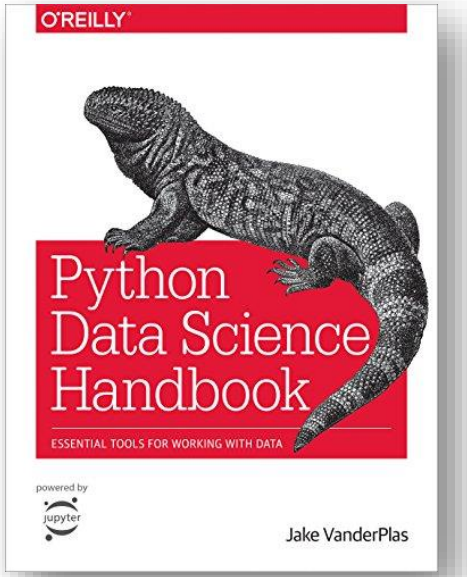
PRÁTICA

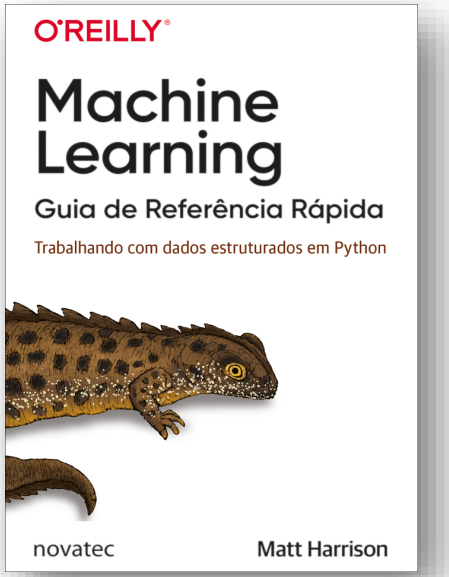
</> Abra o arquivo **"aula16-parte2-case-olist.ipynb"**

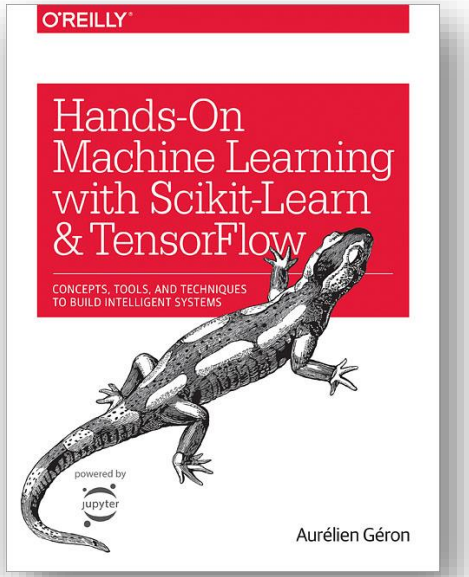
Referências Bibliográficas

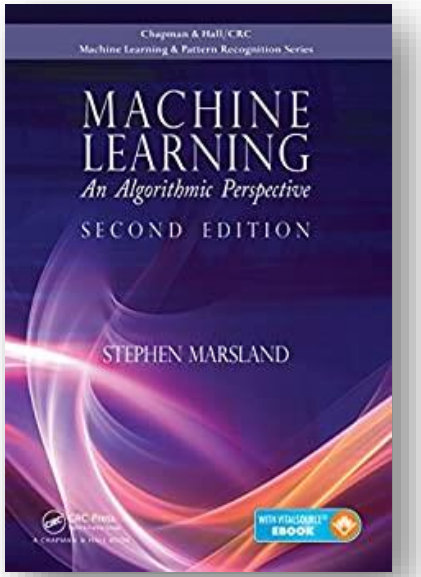
Referências Bibliográficas

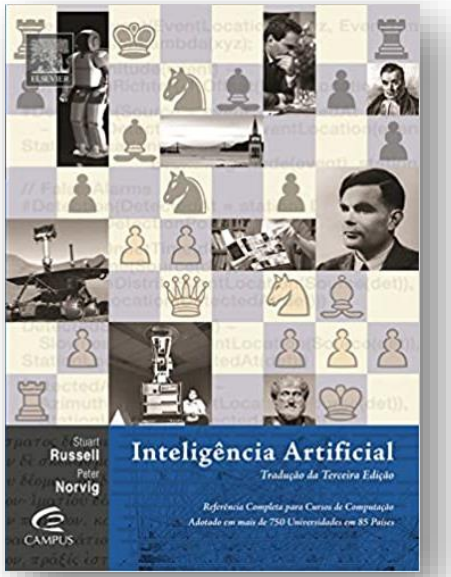
LIVROS





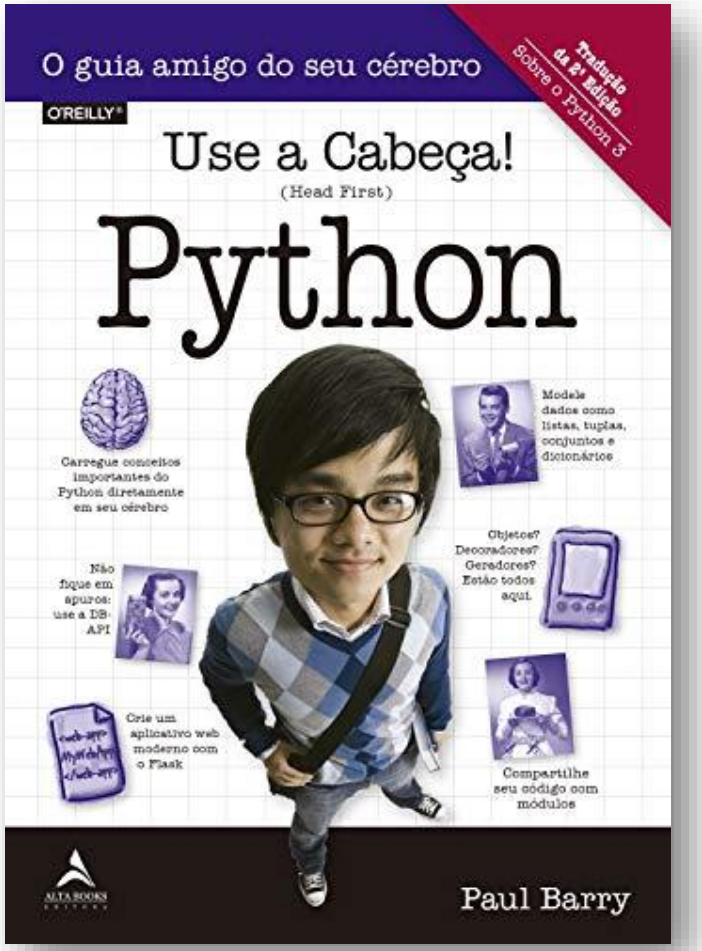


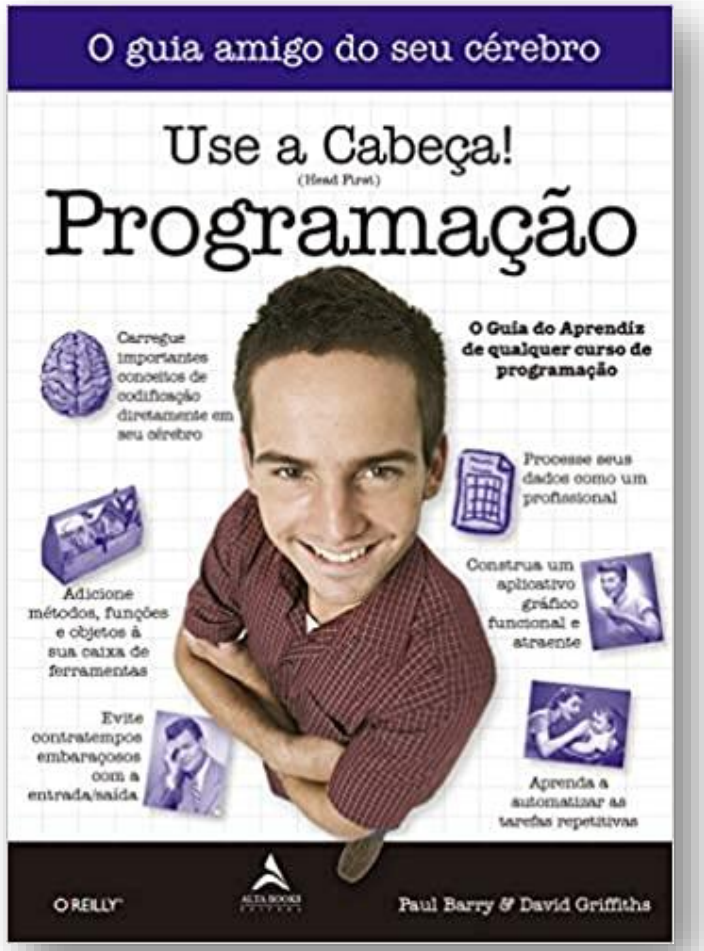


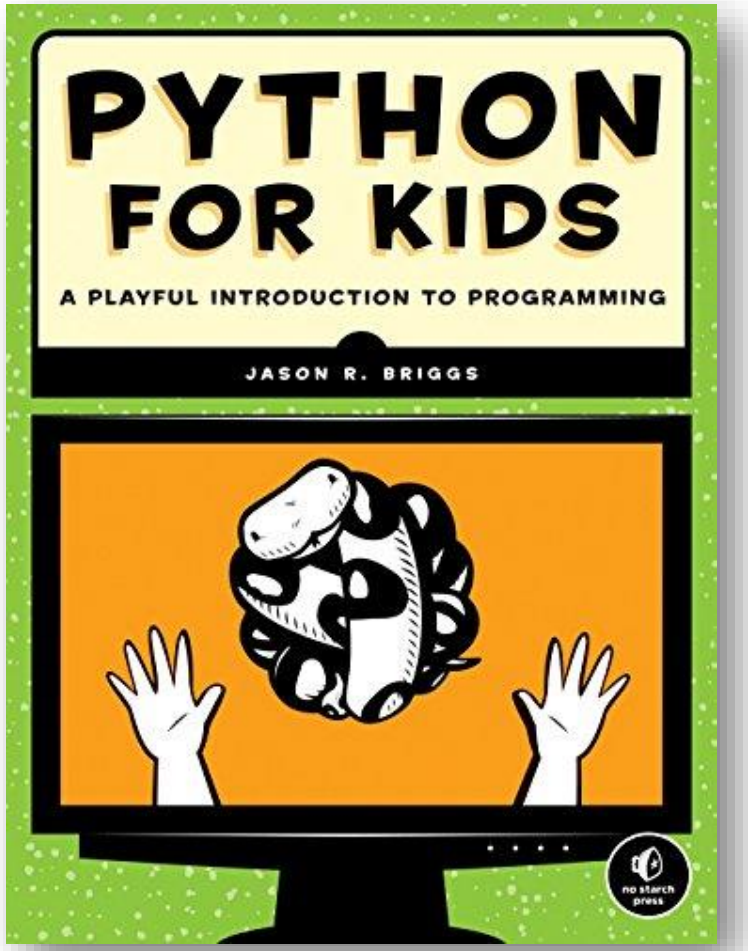


Referências Bibliográficas

LIVROS

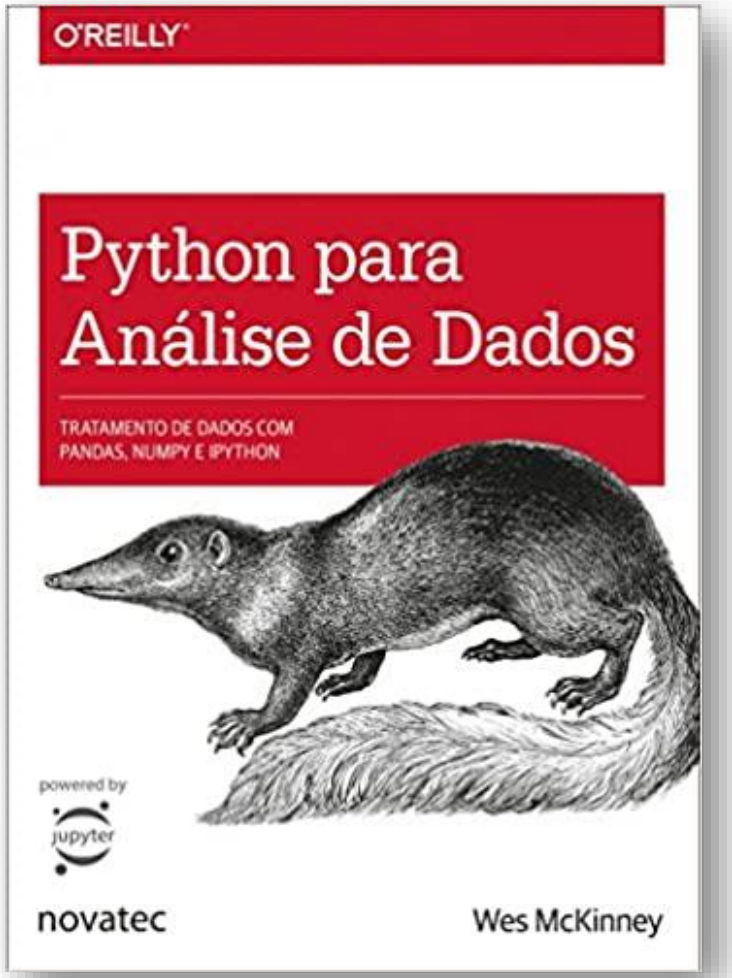


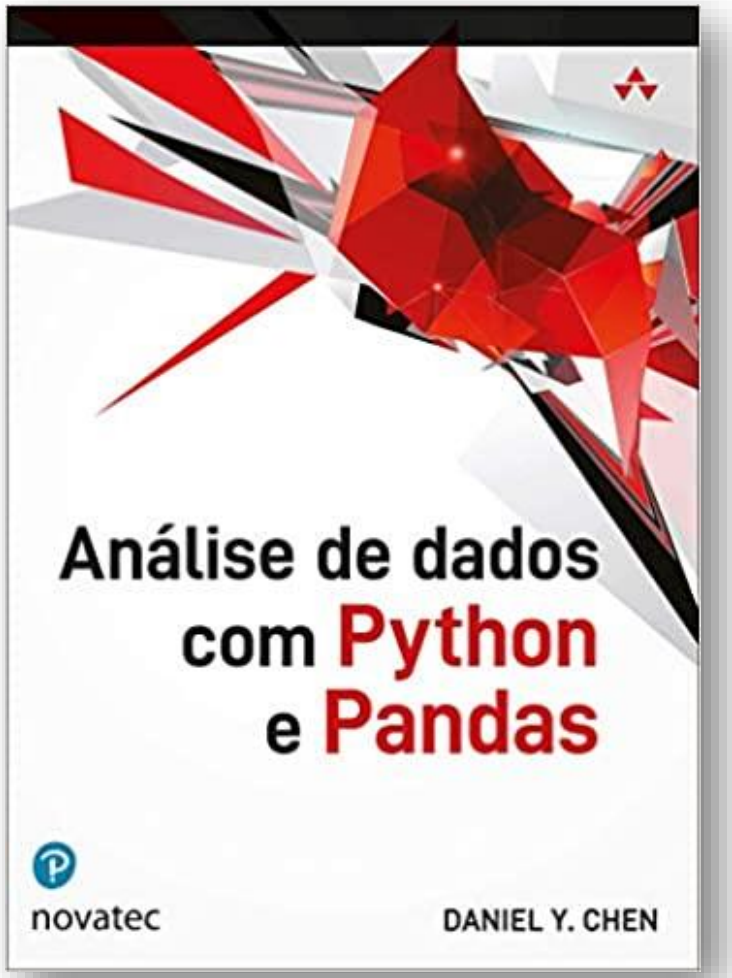


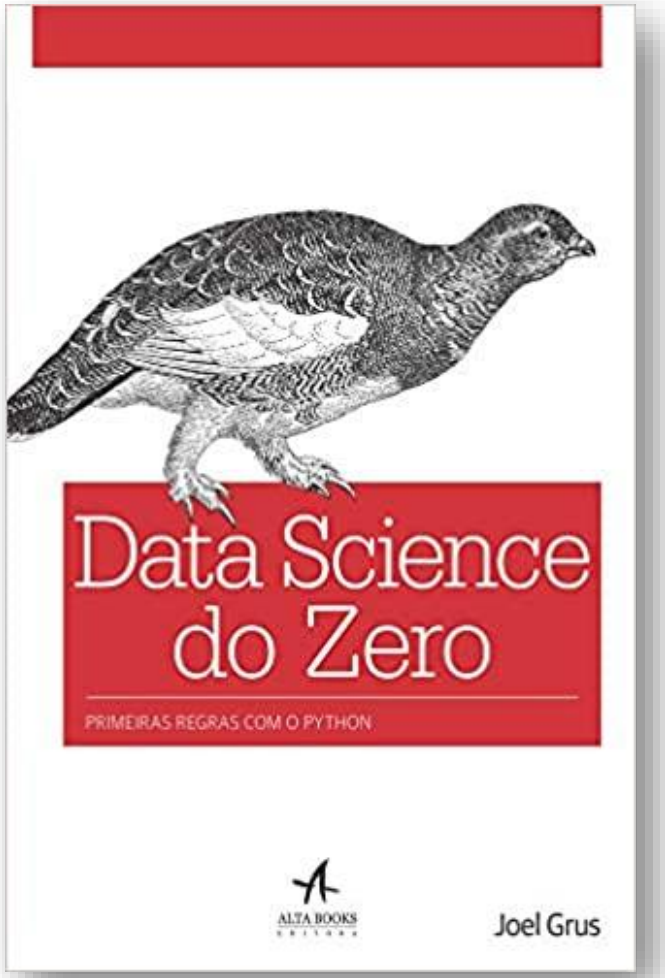


Referências Bibliográficas

LIVROS

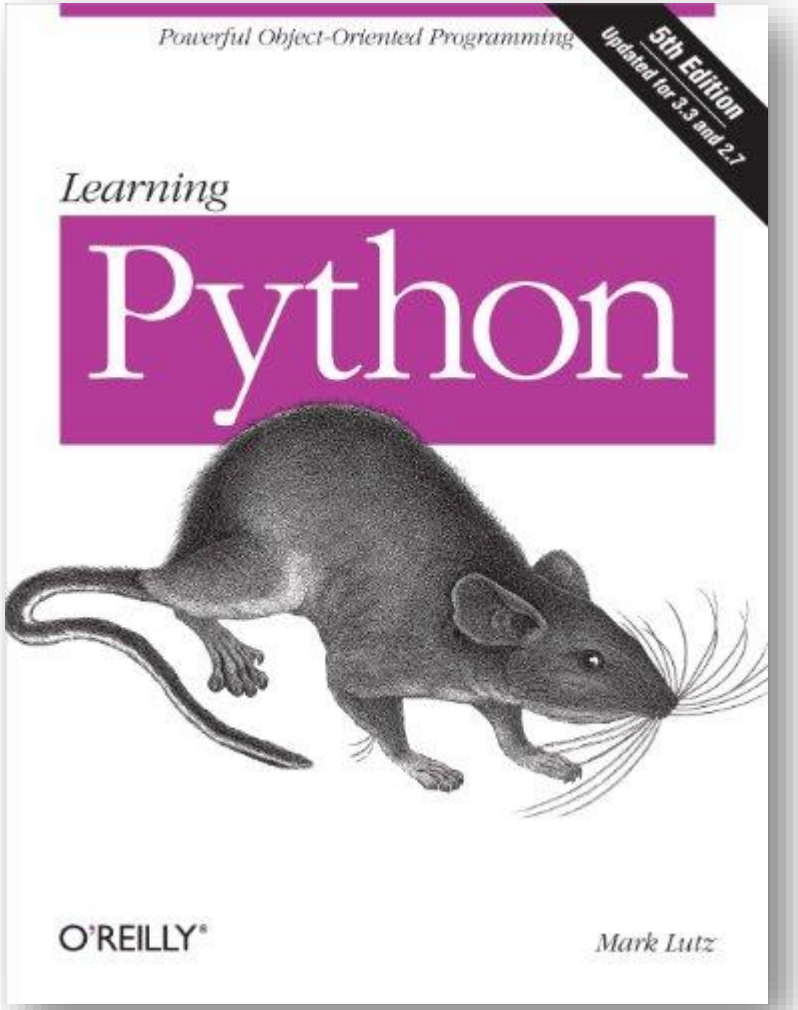


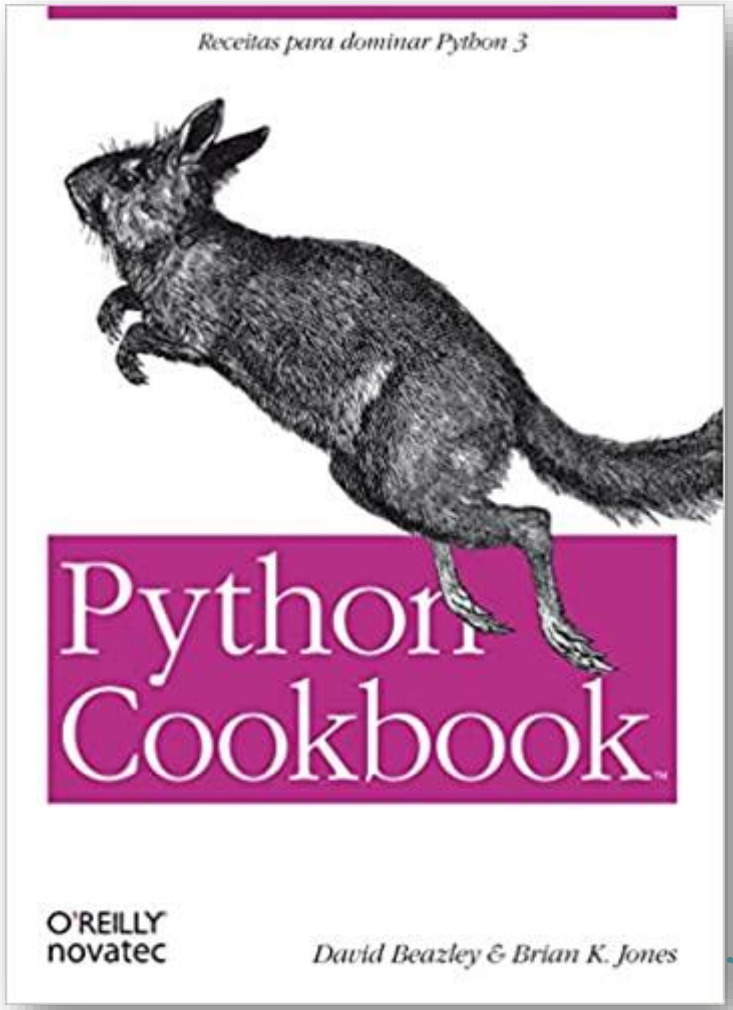




Referências Bibliográficas

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 e Prettycons from www.flaticon.com