CODING IN BRAZIL

Ciência de dados

Wellington Silva

Brasil
2021

Sumário

Lista de ilustrações

1 Introdução

Origem

A mineração de dados surgiu como área de pesquisa e aplicação independente em meados da década de 1990. Entretanto, as suas origens na matemática, estatística e computação são muito anteriores a esse período.

Objetivo

Preparação e análise das grandes massas de dados, tendo a finalidade de encontrar o conhecimento. Portanto, para cumprir tal finalidade, reuni áreas distintas, como estatística; matemática; engenharia; inteligência artificial; banco de dados; sistemas de informação; visualização; antropologia; e o especialista do domínio dos dados, que se complementam e formam a área de ciência de dados.

KDD

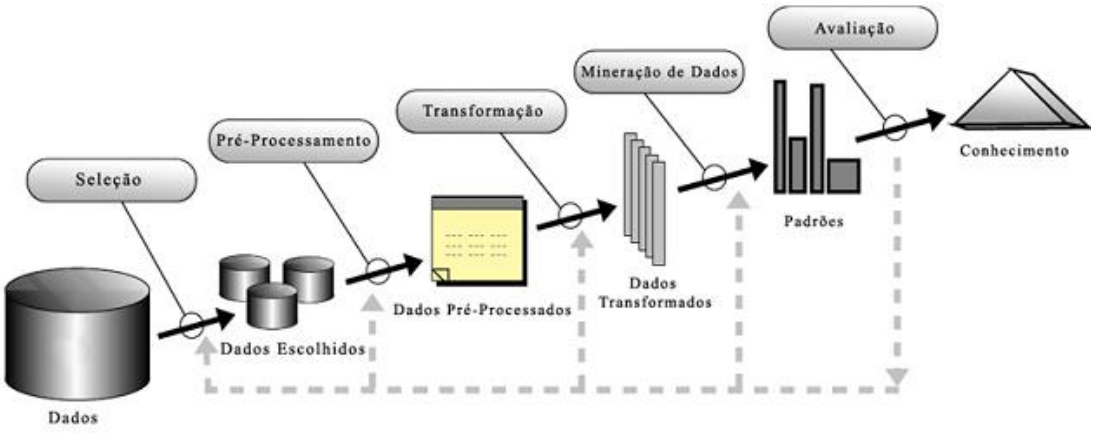

Figura 1 – Ilustração das etapas da extração de conhecimento

1. Dados: conjunto de dados organizados de forma *qualitativa* ou *quantitativa* sobre determinado tema, no qual possibilidade a extração de informação que pode resultar em conhecimento.

2. Pré-processamento dos dados: Selecionar os dados de acordo com a demanda do estudo, descartando assim dados irrelevantes, a fim de tornar a análise dos eficiente e eficaz. As etapas são distribuídas:

   - limpeza: remoção de ruídos de dados inconsistentes e ausentes;
   - integração:combinação dos dados de diferentes fontes;

- seleção: escolha de dados relevantes à análise; e
- transformação: consolidação dos dados em formato apropriado.

3. Mineração de dados: Utilização de métricas e medidas estatísticas, para representar o conjunto de dados e a sua distribuição. Tais medidas são análise descritiva, agrupamento, predição, associação e detecção de anomalias.

4. underlineAvaliação: Identificar os padrões obtidos pela representação do conhecimento são válidos, ou seja, representativo.

1.1 Ferramentas

Edição auxiliada por histórico de comando:

IPython é um interpretador interativo para várias linguagens de programação, mas especialmente focado em Python.

- JupyterLab
- Jupyter Notebook
- Colab Notebooks

Kaggle (https://www.kaggle.com/)

Tradicional:
* Vim
* Atom
* Sublime-text
* VSCode
* Spyder3
* PyDev
* PyCharm da JetBrains
* Komodi IDE
* [Kite](https://www.kite.com/)
IDLE

- Weka (www.cs.waikato.ac.nz/ml/weka)
- Matlab
- R Studio (www.r-project.org)
  - Bioconductor (www.bioconductor.org)
- Wolfram Mathematica (www.wolfram.com/mathematica)

- RapidMiner (rapidminer.com)
- SAS (sas.com)
- SSPS by IBM (www-01.ibm.com/software/analytics/spss)
- Orange (orange.biolab.si)
- Mahout by Apache (mahout.apache.org)
- ELKI (elki.dbs.ifi.lmu.de): aprendizagem não supervisionado
- LIBSVM (www.csie.ntu.edu.tw/ cjlin/libsvm)

Visualização de dados

* Tableau: visualização dinâmica dos dados e dashboard personalizado
* PowerBI: integração com os serviços Microsoft
* Data Studio: integração com os serviços Google
* Qlik: licitação

SGBD

* [PostgreSQL](www.postgresql.org)
* [ORACLE](www.oracle.com)
* [SQL Server](www.microsoft.com/en-us/server-cloud/products/sql-server)
* [DB2](www-01.ibm.com/software/data/db2)

Apache

* [Airflow](https://airflow.apache.org/): plataforma de gerenciamento de fluxo de trabalho de código aberto

Integração dos dados

* [Kafka](https://kafka.apache.org/)
* [Spark](https://spark.apache.org/)
* [NiFi](https://nifi.apache.org/)

Microsserviços

* Docker
* Docker Compose
* Kubernet

1.2 Material complementar

Livros:

- Introdução a mineração de dados por Ferrari e Silva (2017)
- Data Science para Negócios por Fawcett e Provost (2018)
- Python para análise de dados por McKinney (2019)
- Introdução à Ciência de Dados Fundamentos e Aplicações [1]

Cursos:

- ML4all - UFPR [2]

Blog:

- DIKW by Towards Data Science [3]
- Curso R [4]
- Tests as linear by Lindeloev [5]
- JTemporal [6]

Base de dados:

- UCI Machine Learning Repository [7]
- KDnuggets [8]
- Governo Brasileiro [9]
  - Brasil IO [10]
  - Gasto de parlamentar [11]
- Governo Americano [12]

[1] https://www.ime.usp.br/ jmsinger/MAE5755/cdados2019ago06.pdf
[2] http://cursos.leg.ufpr.br/ML4all/1parte/
[3] https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb
[4] https://blog.curso-r.com/
[5] https://lindeloev.github.io/tests-as-linear/
[6] https://jtemporal.com/
[7] http://archive.ics.uci.edu/ml/index.php
[8] https://www.kdnuggets.com/datasets/index.html
[9] https://dados.gov.br/
[10] https://brasil.io/
[11] https://serenata.ai/
[12] https://www.data.gov/

- Governo do Inglês [13]
- PyData Book [14]

[13] https://data.gov.uk/
[14] https://github.com/wesm/pydata-book

2 Python

2.1 Gerencimaneto de pacote

Python Package Index - PIP [15] Sistema de gerenciamento de pacotes padrão de facto usado para instalar e gerenciar pacotes de software escritos em Python. Muitos pacotes podem ser encontrados na fonte padrão para pacotes e suas dependências.

Anaconda [16] Distribuição gratuita e de código aberto das linguagens de programação Python e R para computação científica, que visa simplificar o gerenciamento e a implantação de pacote.

Conda [17] Gerenciador de pacotes e sistema de gerenciamento de ambiente de código aberto, plataforma cruzada e independente de linguagem.

2.2 Essencial

NumPy é um pacote para a linguagem Python que suporta arrays e matrizes multidimensionais, possuindo uma larga coleção de funções matemáticas para trabalhar com estas estruturas.

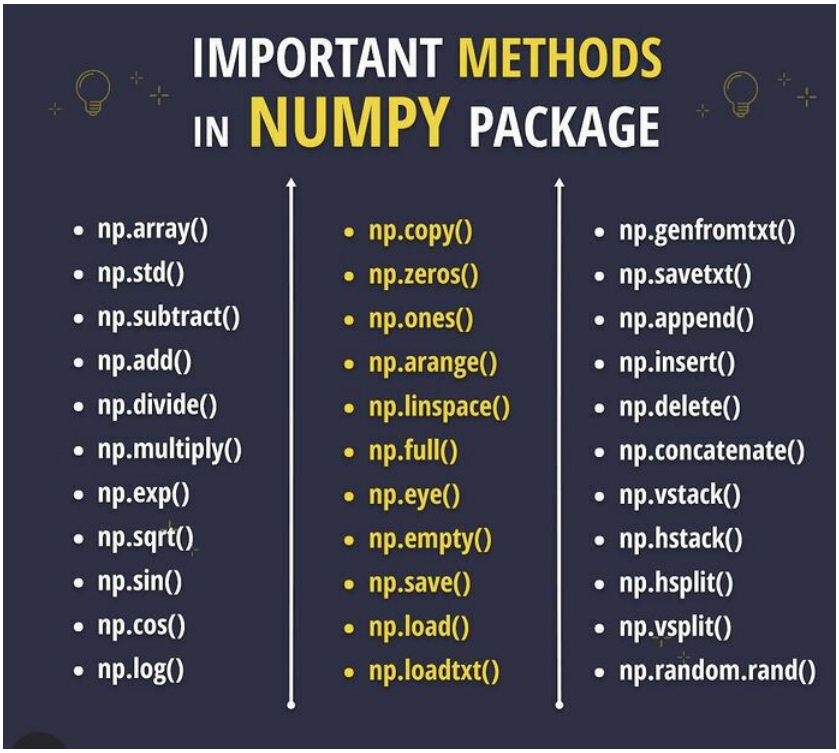

Figura 2 – NumPy

Pandas é uma biblioteca de software criada para a linguagem Python para manipulação e análise de dados. Em particular, oferece estruturas e operações para manipular tabelas numéricas e séries temporais. O nome é derivado de painel data.

[15] https://pypi.org/project/pip
[16] https://www.anaconda.com
[17] https://docs.conda.io/en/latest

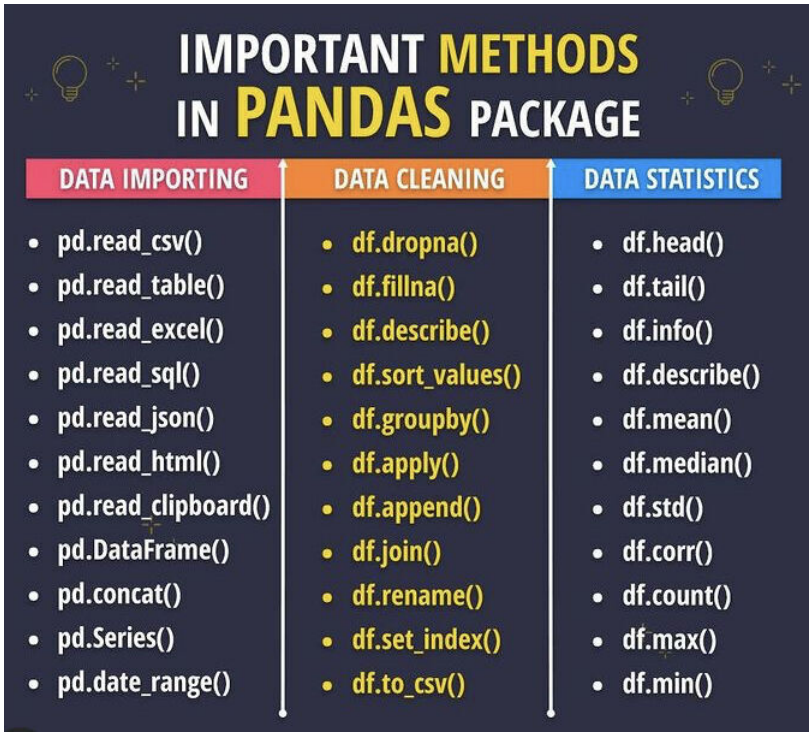

Figura 3 – Pandas

2.3 Matemática

SymPy é uma biblioteca Python para computação simbólica. Ela fornece ferramentas de álgebra computacional tanto como uma aplicação independente como, também, uma biblioteca para outras aplicações.

SciPy é uma biblioteca Open Source em linguagem Python que foi feita para matemáticos, cientistas e engenheiros. Também tem o nome de uma popular conferência de programação científica com Python.

Statsmodels é um pacote Python que permite aos usuários explorar dados, estimar modelos estatísticos e executar testes estatísticos

2.4 Visualização de dados

Matplotlib [18] é uma biblioteca para geração de gráficos e visualizações de dados em geral, feita para e da linguagem de programação Python e sua extensão de matemática NumPy.

Seaborn [19]

2.5 Web Scraping

LXML [20]

[18] https://matplotlib.org/
[19] https://seaborn.pydata.org/
[20] https://lxml.de

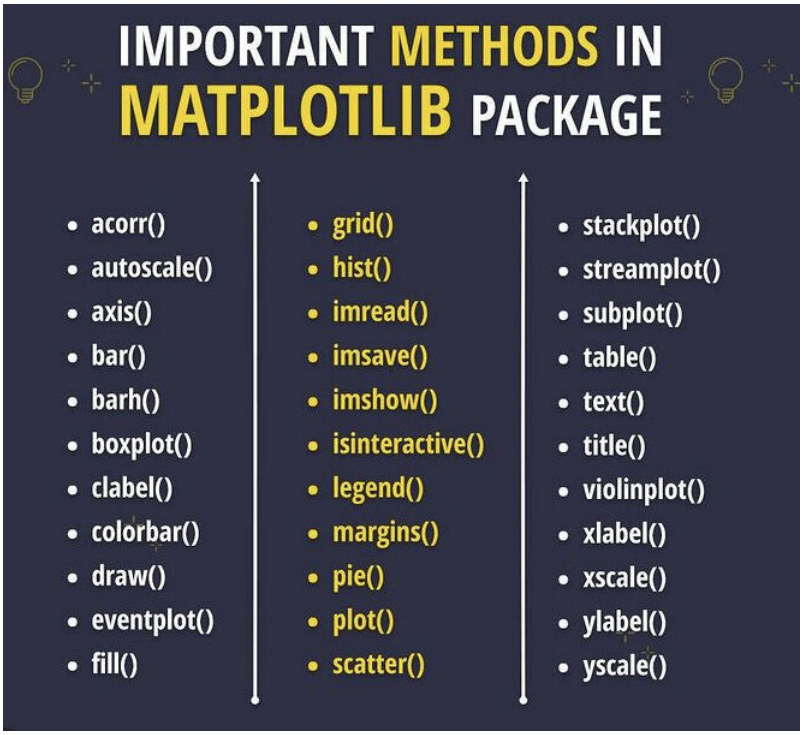

Figura 4 – Matplotlib

Figura 5 – Seaborn

HTLM5LIB [21]

Beautiful Soup [22]

2.6 Aprendizagem de máquina

Scikit-learn [23] Biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python

TensorFlow [24] Biblioteca de código aberto para aprendizado de máquina aplicável a uma ampla variedade de tarefas. É um sistema para criação e treinamento de redes neurais para detectar e decifrar padrões e correlações

Keras [25] Biblioteca de rede neural de código aberto escrita em Python. Ele é capaz de rodar em cima de TensorFlow, Microsoft Cognitive Toolkit, R, Theano, ou PlaidML. Projetado para permitir experimentação rápida com redes neurais profundas, ele se concentra em ser fácil de usar, modular e extensível

2.7 Processamento de imagens

Python Imaging Library Biblioteca da linguagem de programação Python que adiciona suporte à abertura e gravação de muitos formatos de imagem diferentes.

OpenCV Biblioteca multiplataforma, totalmente livre ao uso acadêmico e comercial, para o desenvolvimento de aplicativos na área de Visão computacional

Scikit-image Biblioteca de processamento de imagens de código aberto para a linguagem de programação Python. Inclui algoritmos para segmentação, transformações geométricas, manipulação do espaço de cores, análise, filtragem, morfologia, detecção de recursos e muito mais.

PyTorch Biblioteca de aprendizado de máquina de código aberto baseada na biblioteca Torch, usada para aplicativos como visão computacional e processamento de linguagem natural

[21] https://html5lib.readthedocs.io/en/latest
[22] https://www.crummy.com/software/BeautifulSoup/bs4/doc
[23] https://scikit-learn.org
[24] https://www.tensorflow.org
[25] https://keras.io

Referências

FAWCETT, T.; PROVOST, F. *Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. [S.l.]: Alta Books Editora, 2018.

FERRARI, D. G.; SILVA, L. N. D. C. *Introdução a mineração de dados*. [S.l.]: Saraiva Educação SA, 2017.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. [S.l.]: Novatec Editora, 2019.