## BIG DATA Analytics: Mineração e Análise de Dados

*Rogério de Oliveira*

# Apresentação da Componente Curricular

Nesta componente você vai entender os principais conceitos de Aprendizagem de Máquina. Você entenderá alguns os príncipios do aprendizado de máquina, como sobreajuste, generalização e o papel da otimização nesse aprendizado. Irá conhecer e explorar modelos Supervisionados e não Supervisionados de Aprendizado e aplicá-los nas tarefas mais comuns de Regressão, Classificação e Clusterização de Dados. Verá também princípios de técnicas de Mineração de Regras de Associação, Sistemas de Recomendação e de Modelos Neurais. Para todos esses conceitos você irá implementar soluções práticas empregando os principais pacotes Python para o aprendizado de máquina, como o scikit-learn.

## Como empregar este material

Ao longo de textos apresentamos e exploramos os conceitos com várias implementações em Python e seu pacotes. Mas **este não é um curso de programação**. Nosso foco é nos conceitos do Aprendizado de Máquina e o uso de implementações aqui é útil para explorarmos e aprofundarmos os conceitos aprendidos. Assim, você pode apreciar e compreender vários dos conceitos que serão apresentados com uma simples *leitura diagonal* dos códigos e os exercícios e atividades propostos irão requerer apenas o reuso de códidos e soluções prontas, com pequenas adaptações a cada problema, que propriamente a codificação e implementação de soluções completas.

De qualquer modo, como outros texto que incluem programação, o código é parte integrante do texto e é importante que você acompanhe a leitura dos códigos fornecidos com execuções, experimentações e se tiver interesse, consultando a documentação das api's e pacotes empregados. O texto discute o que é essencial nos códigos mas não discute cada comando e detalhe da codificação. Isso pode ser lido diretamente no código dos programas, e opções menos importantes podem ser exploradas pelo leitor na documentação dos pacotes.

Em cada trilha você ainda vai encontrar sugestões de consultas *Para Saber Mais* e referências, muitas delas eletrônicas para seu fácil acesso. Em todas as trilhas (exceto a trilha 1) você vai encontrar também casos práticos de implementação. Por último, ao longo do texto alguns códigos são empregados unicamente para produzir figuras ou reproduzir resultados que ajudam a explicar o conceito ou a ideia que é apresentada. Esses códigos não têm um interesse particular e você pode deixar de explorá-los se quiser (eles aparecem com o comentário *you can skip this code!*).

## Programa

Os conteúdos da disciplina estão divididos do seguinte modo, e consideramos útil adicionar também os casos e principais algoritmos empregados em cada uma das 8 trilhas de aprendizagem:

1. [**Apresentação**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T0_Apresentacao.ipynb)
2. [**Introdução: Mineração, Ciência de Dados e o Aprendizado de Máquina**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T1_Introducao_a_Ciencia_de_Dados_e_ML.ipynb)
3. [**Regressão e Classificação: Regressão Linear e Logística**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T2_Regressao_e_Classificacao.ipynb)
4. [**Classificação e Métricas**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T3_Classificacao_e_Metricas.ipynb)
5. [**K-Vizinhos Mais Próximos, Validação Cruzada e GridSearch**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T4_Knn_CV_GridSearch.ipynb)
6. [**Árvores de Decisão, Seleção de Atributos e outros Classificadores**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T5_DecisionTrees_MutualInfo_others.ipynb)
7. [**Aprendizado não Supervisionado: Clustering**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T6_Clustering.ipynb)
8. [**Aprendizado não Supervisionado: Regras de Associação e Filtros de Conteúdo**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T7_Regras_de_Associacao_e_Filtros.ipynb)
9. [**Redes Neurais e Deep Learning**](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_T8_Modelos_Neurais.ipynb)

|  |  |  |
| --- | --- | --- |
| **Trilha** | **Casos Práticos** | **Principais Algoritmos** |
| 1. | - | - |
| 2. | CASO: Estimando o Preço de Veículos | Regressão Linear |
|  | CASO: Estimando o tipo de Transmissão dos Veículos | Regressão Logística |
| 3. | CASO: 10 year risk of coronary heart disease CHD | Regressão Logística |
| 4. | CASO: Breast Cancer biopsy | K-Vizinhos mais Próximos |
| 5. | CASO: Classifying Political Parties Based on Congressional Votes | Decision Trees & Random Forest |
| 6. | CASO: Segmentando Estados para Políticas e Campanhas de não Violência | K-Médias |
|  | CASO: Wholesale Customer Data | Clustering Hierárquico |
| 7. | CASO: UK Online Retail Data Set | Apriori |
|  | CASO: Spotify Music Recommendation | Knn não Supervisionado |
| 8. | CASO: Kyphosis Diagnostic | MLP e Deep Learning |

# Atividades de Aprofundamento

Ao final dessas atividades você terá implementado e discutido ao menos um modelo de regressão, um de classificação e um de clusterização de dados.

* [1. Regressão com statsmodels](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A1_LinearRegressionStatsmodels.ipynb)
* [2. Discussão: Multicolinearidade, Hotencode e R2-ajustado](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A2_LinearRegressionDiscussao.ipynb)
* [3. Classificação com scikit-learn](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A3_Classificacao.ipynb)
* [4. Discussão: Seleção de Modelos](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A4_ClassificacaoDiscussao.ipynb)
* [5. Clusterização com scikit-learn](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A5_Clustering.ipynb)
* [6. Discussão: Clusters, métricas e Classificação](https://colab.research.google.com/github/Rogerio-mack/BIG_DATA_Analytics_Mineracao_e_Analise_de_Dados/blob/main/BIG_A6_ClusteringDiscussao.ipynb)

# Principais Referências

Alpaydin, E. **Machine Learning** (The MIT Press Essential Knowledge). The MIT Press. 2019.

Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. **Dive into Deep Learning** (2020). Disponível em: <https://d2l.ai/index.html> Acesso em: 21 de abril de 2021.

Fridman, L. **MIT Deep Learning Basics: Introduction and Overview with TensorFlow**. Disponível em: <https://blog.tensorflow.org/2019/02/mit-deep-learning-basics-introduction-tensorflow.html> Acesso em: 21 de abril de 2021.

Géron, A. **Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems**, 2nd ed. (2019) O'Reilly

Ian Goodfellow and Yoshua Bengio and Aaron Courville, **Deep Learning**. MIT Press (2016). Also available online: <http://www.deeplearningbook.org>.

Jake VanderPlas. **Python Data Science Handbook** O'Reilly Media, Inc. (2016). ISBN: 9781491912058. Disponível em: <https://jakevdp.github.io/PythonDataScienceHandbook/>. Acesso: 06 de Novembro de 2021.

Kelleher, J. D.; Tierney, Brendan. **Data Science** (The MIT Press Essential Knowledge). The MIT Press. 2018.

Kelleher, J. D. **Deep Learning** (2018) The MIT Press.

Koren, Y., Bell, R., & Volinsky, C. (2009). **Matrix Factorization Techniques for Recommender Systems**. Computer, 42(8), 30–37. <doi:10.1109/mc.2009.263>

Kotu, Vijay; Deshpande, Balachandre **Data Science: concepts and practice**. 2nd ed. Cambridge, [England]: Morgan Kaufmann, c2019. E-book (570 p.) ISBN 9780128147627 (electronic bk.). Disponível em: <http://pergamum.mackenzie.br:8080/pergamumweb/vinculos/00003c/00003cef.jpg>.

Larose, Chantal D.; Larose, Daniel T. **Data Science Using Python and R** Hoboken: Wiley, c2019. E-book (259 p.) (Wiley Series on Methods and Applications in Data Mining Ser.). ISBN 9781119526834 (electronic bk.). Disponível em: <https://www3.mackenzie.br/biblioteca_virtual/index.php?tipoBiblio=ebookcentral&flashObg=n>

Manapat, Michael. **Introduction to Machine Learning with Python**. Em An Introduction to Machine Learning. EMag Edição 50 (Abr 2017)

Molnar Christoph, **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable** Disponível em: <https://christophm.github.io/interpretable-ml-book/> Acesso em: 06 de Novembro de 2021.

Oliveira, R. de. **Probabilidade e Estatística Aplicada com R** Disponível em: <https://github.com/Rogerio-mack/Probabilidade-Estatistica-Aplicada-R> Acesso em: 06 de Novembro de 2021.

Oliveira, R. de. **Deep Learning I** Disponível em: <https://github.com/Rogerio-mack/Deep-Learning-I> Acesso em: 16 de novembro de 2021.

Peikari, M., Salama, S., Nofech-Mozes, S. et al. **A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification**. Sci Rep 8, 7193 (2018). <https://doi.org/10.1038/s41598-018-24876-0>

Schrage, M. **Recommendation Engines** (The MIT Press Essential Knowledge). The MIT Press. 2020.

Skansi, S. **Introduction Deep Learning** (2018) Springer.

Tan, P.-N., Steinbach, M.,, Kumar, V. (2005). **Introduction to Data Mining.** Addison Wesley. ISBN: 0321321367 e também o Book Site, Disponível em: <https://www-users.cse.umn.edu/~kumar001/dmbook/index.php> Acesso em: 16 de novembro de 2021.

\_\_\_. **An introduction to machine learning with scikit-learn** Disponível em: <https://scikit-learn.org/stable/tutorial/basic/tutorial.html> Acesso em: 06 de Novembro de 2021.

\_\_\_. **scikit-learn: machine learning in Python** Disponível em: <http://scipy-lectures.org/packages/scikit-learn/index.html> Acesso em: 06 de Novembro de 2021.