X-Ray da metodologia aplicada

***Autor:*** *Lucas Lo Ami* ***Última atualização:*** *19 de jan. de 2022*

***Objetivo do documento:*** *explicar o processo de mentoria conduzido com meus alunos, bem como as principais referências utilizadas e compartilhadas com os discentes ao longo das sessões. É importante ressaltar que este documento contém materiais complementares a cursos de ciência e análise de dados, não substituindo os mesmos. O mesmo vale para a metodologia.*

| **Disclaimer**: por motivos de contrato de não competição, não posso mentorar pessoas que tenham relações profissionais com empresas concorrentes do meu atual empregado. Portanto, esse é um item que sempre preciso avaliar antes de tomar a decisão de seguir em frente com as sessões. |
| --- |

# **A jornada do cientista de dados**

O cientista de dados é um profissional cada vez mais requisitado no mercado, não apenas brasileiro mas no internacional também. A depender do contexto de negócios no qual está inserido, este profissional pode ter um viés mais próximo de um analista de dados ou de um especialista técnico. O Airbnb [descreve os diferentes tipos de cientistas de dados](https://www.linkedin.com/pulse/one-data-science-job-doesnt-fit-all-elena-grewal/) que existem por lá. Colson [faz alguns comentários](https://www.quora.com/What-do-you-think-of-the-3-types-of-data-science-described-by-Airbnb-Analytics-Inference-and-Algorithms-Is-this-a-useful-way-to-differentiate-and-organize-the-types-of-data-scientists-at-your-company) a respeito desse texto que complementam a ideia do mesmo.

As mentorias ofertadas por mim reforçam o perfil generalista e de [profissional "T"](https://en.wikipedia.org/wiki/T-shaped_skills), ou seja, elas visam dar uma visão ampla sobre os diversos tipos de problema enfrentados por cientistas de dados no mercado e, se possível, o aluno conseguirá aprofundar em um tema específico. Com isso, é esperado que o aluno fortaleça as habilidades descritas na tabela 01:

| **Habilidade** | **Descrição** |
| --- | --- |
| Resolução de problemas | capacidade de formulá-los e definir escopo dos mesmos de forma clara e objetiva, bem como desenhar um plano de ação para resolvê-los. Além disso, é importante que o aluno faça uma análise crítica do contexto |
| Visão de negócios | capacidade de entender as diferentes fontes de receita, custos e processos em diferentes mercados, assim como diferenciais competitivos de cada empresa |
| Comunicação | transmitir de maneira clara e objetiva as principais descobertas obtidas em suas análises bem como os resultados dos projetos realizados |
| Visualização de dados | saber aplicar as melhores práticas na construção de gráficos para maximizar a compreensão dos stakeholders |
| Estatística e machine learning | Saber aplicar métodos de estatística descritiva e inferencial adequadamente bem como conhecer os principais tipos de problema resolvidos por aprendizado de máquina |

Tabela 01: habilidades do cientista de dados

O desenvolvimento de tais habilidades e amadurecimento do perfil generalista ou "T" é obtido por meio de projetos alinhados com os desafios do mundo real e sessões de mentoria com esclarecimento de dúvidas e direcionamentos sobre pontos de melhoria nos mesmos. Espera-se que a partir disso o aluno se torne apto a desenvolver projetos de muito impacto nas empresas que trabalhar, [vide framework da Sequoia Capital](https://medium.com/sequoia-capital/progression-of-a-data-scientist-e1bebf8c8420) de progresso de carreira de um cientista de dados.

## Passo-a-passo das sessões

Antes de mais nada, é importante frisar que cabe ao aluno aderir **ou não** a metodologia descrita abaixo. Ela foi desenhada para ajudar os alunos a se tornarem profissionais mais completos e servir de guia para os mesmos durante as sessões de mentoria, maximizando a qualidade das sessões. Pode ser descrita com os passos abaixo:

1. Alinhamento de expectativas: busco compreender o nível de maturidade do aluno em estatística, programação e negócios, além da carga horária de estudos que ele tem planejado. Explico como funciona a metodologia que aplico nas mentorias e definimos próximos passos
2. Definição da rotina de estudos: baseado nos feedbacks do primeiro item, definimos uma rotina inicial de estudos para o aluno, que consiste de projetos, exercícios e/ou livros para serem consumidos.
3. Exercícios de aquecimento: caso o aluno não tenha boa desenvoltura em programação e estatística, o primeiro item de estudos consiste em resolver alguns desafios do HackerRank nessas duas temáticas.
4. Definição de projetos a serem implementados: usa-se como base a lista de projetos sugeridos abaixo, bem como outros que possam ser de interesse do aluno
5. Criação do github e padrão de portfólio
6. Ciclo de execução de projeto (feito até o final das mentorias)
   1. Entender o projeto a fundo: qual o problema ele visa resolver? Quais técnicas geralmente são usadas para resolver o desafio? Quais os melhores livros e artigos que trazem esses conceitos?
   2. Disponibilização de conteúdo de referência para criação do projeto
   3. Aluno implementa o projeto e compartilha com o mentor
   4. Revisão do projeto realizado pelo aluno

O número de projetos realizados depende da disponibilidade e velocidade de aprendizado de cada aluno, mas recomendo ter pelos a seguinte quantidade de projetos ao final da jornada:

* 1 projeto de análise de dados
* 2 projetos de regressão
* 2 projetos de classificação
* 1 projeto de clusterização
* 1 projeto de otimização linear ou sistema de recomendação

A sessão de [portfólio](#_a629z0rmzg5y) contém sugestões de todos os projetos que devem compor o arcabouço do aluno. Caso este tenha uma agenda lotada e baixa carga horária de estudos, recomendo uma abordagem alternativa de mentorias, na qual ao invés de fazer diversos projetos, o aluno se compromete a ler alguns livros para diversificar seus conhecimentos. Nessa opção, o discente deverá ler ao menos;

* 1 livro sobre negócios
* 1 livro sobre tomada de decisão com uso de dados
* 1 livro sobre projetos gerais de data science no mercado

# **Literatura recomendada**

A tabela -2 contém um compilado de livros e vídeos que ajudam o aluno a desenvolver as habilidades citadas no começo deste documento. Recomendo que, inicialmente, as leituras se iniciem pela categoria "negócios", para que assim o aluno consiga avaliar criticamente as diferentes empresas nas quais irá trabalhar e, ao mesmo tempo, conectar essa análise com dados e os projetos que farão.

Os livros de ciência de dados da subcategoria geral são importantes para dar uma visão mais ampla do que já é feito atualmente no mercado.

O aluno não precisa ler todos esses conteúdos, mas recomendo que tenham a maioria dos livros disponíveis para consulta quando necessário no dia-a-dia de trabalho. Caso haja dificuldade para adquirir os livros, basta conversar comigo.

| **Categoria** | **Subcategoria** | **Tipo** | **Conteúdo** |
| --- | --- | --- | --- |
| Negócios | Geral | Livro | 1. Lean Analytics  2. Zero To One  3. Talking to Humans  4. A estratégia do oceano azul |
| Negócios | Problem Solving | Livro | 1. Behind Every Good Decision  2. |
| Negócios | Problem Solving | Vídeo | 1. [How do consulting firms work (hypothesis-based problem solving explained)](https://www.youtube.com/watch?v=TBvJzXxRuxs&t=2s)  2. [Comprehensive McKinsey Hypotheses Based Case Interview Approach](https://www.youtube.com/watch?v=dwsQ2f2wYpU)  3. [Bulletproof Problem Solving: Sydney Airport Case Study](https://www.youtube.com/watch?v=XmXaV7oB_NI) |
| Data Science | Geral | Livro | 1. Data Smart  2. Data Science para Negócios  3. Como Mentir com estatística |
| Data Science | Machine Learning | Livro | 1. Machine Learning: A Probabilistic Perspective  2. The Elements of Statistical Learning |
| Data Science | Bayes | Livro | 1. Statistical Rethinking |
| Data Science | Visualização de dados | Livro | 1. Storytelling with data  2. How Charts Lie |
| Data Science | Séries Temporais | Livro | [Forecasting: Principles and Practice](https://otexts.com/fpp2/moving-averages.html) |

Tabela 02: principais livros e vídeos para os alunos consumirem

# **Portfólio recomendado**

Os conteúdos dos projetos versam não apenas sobre temas abordados em sala de aula nos cursos de ciência de dados, mas também exploram outros tópicos que acredito serem importantes no desenvolvimento profissional de um cientista de dados. A tabela 03 contém um resumo destes projetos.

| **Categoria** | **Objetivos/Descrição** | **Datasets** |
| --- | --- | --- |
| Análise de dados  (1 projeto) | Exercitar a exploração de dados com viés de negócios, desde a elaboração das perguntas-chave a serem respondidas, até desenvolvimento de visualizações de dados usando as melhores práticas. Implementar com as bibliotecas do Python ajuda a aumentar as habilidades com essa tecnologia | [Brazilian E-Commerce Public Dataset by Olist](https://www.kaggle.com/olistbr/brazilian-ecommerce) |
| Análise de dados  (1 projeto) | Desenvolver o olhar crítico sobre visualizações de dados mal elaboradas e propor melhorias práticas para estas | [Gráfico de Furacão - slide da aula](https://docs.google.com/presentation/d/1zWY671ZKOU0EuQqWO6cmZvTED78qFKTp_Ld6tDuZsY4/edit?usp=sharing) |
| Análise de dados  (1 projeto) | Exercitar o conhecimento de testes de hipótese (estatística) através da análise de teste A/B | [Udacity A/B Testing](https://www.kaggle.com/zhangluyuan/ab-testing?select=ab_data.csv)  [A/B testing web analytics data](https://scholarworks.montana.edu/xmlui/handle/1/3507) |
| Aprendizado supervisionado - regressão  ( 2 ou 3 projetos) | Criar diferentes projetos a fim de explorar desde os conceitos mais simples de cada algoritmo até diferentes estratégias de manipulação de *features* e os porquês disso. O aluno deverá, ao final dessa etapa, entender:   * Funcionamento dos algoritmos mais simples como regressão linear, polinomial, árvore de regressão, *support vector machines* * Regularização, quais os tipos que existem, quando aplicar, quando não aplicar e efeitos gerados por elas * *Feature scaling* (normalização), quais os tipos que existem, quando aplicar, quando não aplicar e efeitos gerados pelas técnicas existentes * Estratégias para lidar com valores faltantes no dataset * Quais as principais métricas de avaliação de modelos de regressão, seus prós e contras | [Brazilian E-Commerce Public Dataset by Olist](https://www.kaggle.com/olistbr/brazilian-ecommerce)  [Solar Flare](https://archive-beta.ics.uci.edu/ml/datasets/solar+flare)  [Online Retail II](https://archive-beta.ics.uci.edu/ml/datasets/online+retail+ii) |
| Aprendizado supervisionado - classificação  ( 2 ou 3 projetos) | Criar diferentes projetos a fim de explorar desde os conceitos mais simples de cada algoritmo até diferentes estratégias para lidar com dataset desbalanceados e outros problemas clássicos. O aluno deverá, ao final dessa etapa, entender:   * Funcionamento dos algoritmos mais simples como regressão logística, árvore de decisão, *naive* bayes * Como tratar datasets desbalanceados * Diferentes estratégias para lidar com features categóricas dentro da solução proposta * Diferentes técnicas para fazer seleção de *features* * Quais as principais métricas de avaliação de modelos de classificação, seus prós e contras | [SMS Spam Collection](https://archive-beta.ics.uci.edu/ml/datasets/sms+spam+collection)  [South German Credit](https://archive-beta.ics.uci.edu/ml/datasets/south+german+credit)  [Default of credit card clients](https://archive-beta.ics.uci.edu/ml/datasets/default+of+credit+card+clients) |
| Aprendizado supervisionado - séries temporais  (1 projeto) | Criar diferentes projetos a fim de explorar desde os conceitos mais simples de cada algoritmo, suas limitações além dos principais componentes de uma série temporal. O aluno deverá, ao final dessa etapa, entender:   * Os principais componentes de uma série temporal * O funcionamento dos algoritmos mais clássicos como ARIMA-based models * Como endereçar sazonalidade dentro do dataset * Técnicas estatísticas para determinar se uma série é estacionária ou não | [Minimum Daily Temperatures Dataset](https://www.kaggle.com/paulbrabban/daily-minimum-temperatures-in-melbourne)  [Monthly Mean Total Sunspot Number](https://www.kaggle.com/robervalt/sunspots)  [Ozone Level Detection Dataset](https://www.kaggle.com/prashant111/ozone-level-detection/version/1)  [Brazilian E-Commerce Public Dataset by Olist](https://www.kaggle.com/olistbr/brazilian-ecommerce)  [Online Retail II](https://archive-beta.ics.uci.edu/ml/datasets/online+retail+ii)  [Daily and Sports Activities](https://archive-beta.ics.uci.edu/ml/datasets/daily+and+sports+activities) |
| Aprendizado supervisionado - NLP  ( 2 ou 3 projetos) | Entender as principais técnicas utilizadas em NLP, desde as mais clássicas até as mais modernas (com uso de transformers) e alguns casos de uso da área. O aluno deverá, ao final dessa etapa, entender:   * Quais são as features clássicas: contagem de termos, frequência de termos, tf-idf e como utilizá-las * O que são n-gramas e como são utilizados * Como aplicar diferentes técnicas de limpeza de texto * O que são Lemmatizers, quais os seus objetivos e como utilizá-los * Quais são as features mais informativas de NLP e como utilizá-las: LSA, word2vec, doc2vec, word embeddings | [Reuters-21578 Text Categorization Collection](https://archive-beta.ics.uci.edu/ml/datasets/reuters+21578+text+categorization+collection)  [Drug Review Dataset (Drugs.com)](https://archive-beta.ics.uci.edu/ml/datasets/drug+review+dataset+drugs+com) |
| Aprendizado supervisionado - ensembles  (1 projeto) | Compreender as diferentes técnicas de *ensembles* existentes, quando aplicar cada uma delas bem como suas vantagens e desvantagens. O aluno deverá, ao final dessa etapa, entender:   * A diferença entre métodos de *averaging* e de *boosting* * Como aplicar os principais algoritmos existentes: XGBoost e RandomForest | Os mesmos utilizados nos problemas de regressão e classificação |
| Aprendizado não supervisionado - clustering  (1 projeto) | Compreender as diferentes técnicas de clusterização existentes, definir o número ótimo de clusters e as limitações de cada uma das técnicas. O aluno deverá, ao final dessa etapa, entender:   * A diferença entre modelos de conectividade, modelos de centróide e modelos de densidade * Como aplicar os principais algoritmos existentes: k-means, k-medians, cluster hierárquico, DBSCAN * Como determinar o número ótimo de clusters através da técnica do cotovelo e silhueta | [Hierarchical Sales data](https://archive-beta.ics.uci.edu/ml/datasets/hierarchical+sales+data)  [Human Activity Recognition Using Smartphones](https://archive-beta.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones) |
| Aprendizado não supervisionado - análise de cesta  (1 projeto) | Compreender os principais casos de uso de regras de associação e seus algoritmos. O aluno deverá, ao final dessa etapa, entender:   * Como aplicar as técnicas de maneira adequada e analisar os resultados * Funcionamento dos algoritmos Apriori, SETM e AIS | [Association Rule Learning-Apriori](https://www.kaggle.com/sivaram1987/association-rule-learningapriori) |
| Sistemas de recomendação  (1 projeto) | Compreender as principais técnicas existentes, seus trade-offs e bibliotecas para implementar soluções. O aluno deverá, ao final dessa etapa, entender:   * Funcionamento dos principais algoritmos clássicos: *collaborative filtering*, *content-based filtering* e *context-based filtering* | [MovieLens Dataset](https://grouplens.org/datasets/movielens/25m/)  [Million Song Dataset](http://millionsongdataset.com/) |
| Inferência bayesiana  (1 projeto) | Compreender como funciona a estatística bayesiana, em quais cenários ela é melhor que a frequentista, como fazer simulações em Python. O aluno deverá, ao final dessa etapa, entender:   * Princípios da estatística bayesiana * Como funciona o teste A/B bayesiano * Como fazer um modelo de regressão bayesiano * Como fazer modelos hierárquicos | [Hierarchical Sales data](https://archive-beta.ics.uci.edu/ml/datasets/hierarchical+sales+data) |
| Inferência causal  (1 projeto) | Compreender os principais conceitos que permeiam inferência causal e as principais técnicas utilizadas. | [Awesome Causality Data](https://github.com/rguo12/awesome-causality-data) |

Tabela 03: projetos que devem compor o portfólio do aluno

## Materiais de apoio

A tabela 04 contém uma lista de artigos e tutoriais que mais compartilho com os alunos e que os ajudam ao longo da jornada de mentoria, principalmente quando estiverem realizando seus projetos. A depender dos tópicos e temas explorados nas sessões, posso enviar mais links de referência.

| **Categoria** | **Subcategoria** | **Tipo** | **Conteúdo** |
| --- | --- | --- | --- |
| Machine Learning | Datasets desbalanceados | Artigo | [10 Techniques to deal with Imbalanced Classes in Machine Learning](https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/)  [How to Handle Imbalanced Classes in Machine Learning](https://elitedatascience.com/imbalanced-classes) |
| Seleção de features | Artigo | [How to choose a feature selection method for machine learning](https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/) |
| Feature scaling | Artigo | [All about feature scaling](https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35) |
| Vieses de métricas de seleção | Artigo | [Forcast KPI: RMSE, MAE, MAPE & Bias](https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d) |
| Learning curves | Artigo | [Learning curves explained with Python SkLearn Example](https://vitalflux.com/learning-curves-explained-python-sklearn-example/) |
| Tutorial - lib | [SkLearn Learning Curves](https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py) |
| Overfitting e underfitting | Artigo | [How to Identify Overfitting Machine Learning Models in Scikit-Learn](https://machinelearningmastery.com/overfitting-machine-learning-models/) |
| Shapley values | Livro | [Shapley values chapter](https://christophm.github.io/interpretable-ml-book/shapley.html) |
| Vídeo | [Understanding The Shapley Value](https://www.youtube.com/watch?v=9OFMRiAVH-w) |
| Séries Temporais | Geral | Curso | [Timeseries forecasting](https://www.udacity.com/course/time-series-forecasting--ud980) |
| Teste de hipótese | Artigo | <https://towardsdatascience.com/test-for-existence-of-a-trend-in-a-time-series-3a44f242c329> |
| Data Science | Retenção de clientes | Artigo | [Data Science para retenção de clientes e o case do mercado livre](https://ilumeo.com.br/todos-posts/2021/12/10/data-science-para-retencao-de-clientes-e-o-case-do-mercado-livre) |
| Artigo | [Predict Customer Churn in Python](https://towardsdatascience.com/predict-customer-churn-in-python-e8cd6d3aaa7) |
| Vídeo | [Predict Customer Churn with Machine Learning](https://www.youtube.com/watch?v=90GR8q3eYG0%5D) |
| Technical interview | Artigo | [130 Data Science Interview Questions and Answers [Latest]](https://data-flair.training/blogs/data-science-interview-questions/) |
| Artigo | [71 Data Science Interview Questions and Answers – Crack Technical Interview Now!](https://data-flair.training/blogs/data-science-interview-questions-and-answers/) |
| Artigo | [R Data Science Interview Questions based on top projects](https://data-flair.training/blogs/r-data-science-interview-questions/) |
| Estatística | Teste A/B | Artigo | [A/B Testing for Data Science using Python – A Must-Read Guide for Data Scientists](https://www.analyticsvidhya.com/blog/2020/10/ab-testing-data-science/) |
| Artigo | [A Simple Guide to A/B Testing for Data Science](https://towardsdatascience.com/a-simple-guide-to-a-b-testing-for-data-science-73d08bdd0076) |
| Artigo | [Frequentist vs. Bayesian approach in A/B testing](https://www.dynamicyield.com/lesson/bayesian-testing/) |
| Artigo | [Talking Bayes to Business: A/B Testing Use Case | Shopify](https://www.youtube.com/watch?v=J6kqvWnUE2Q) |
| Artigo | [How to Double A/B Testing Speed with CUPED](https://towardsdatascience.com/how-to-double-a-b-testing-speed-with-cuped-f80460825a90) |
| Artigo | [Bayesian A/B Testing with Python: the easy guide](https://towardsdatascience.com/bayesian-a-b-testing-with-python-the-easy-guide-d638f89e0b8a) |
| Artigo | [MASP — Nosso framework de marketing digital](https://medium.com/loftbr/masp-nosso-framework-de-marketing-digital-3ec46bfc2f96) |
| Causal Inference | Geral | Vídeo | [Science Before Statistics: Causal Inference](https://www.youtube.com/watch?v=KNPYUVmY3NM) |

Tabela 04: tutoriais de apoio para resolução dos desafios