

GUIA PARA DESCOBRIR

Como se tornar um Cientista de Dados?

POR LET'S DATA



Copyright © 2021 Let's Data
Bernardo Lago, Felipe Schiavon de Oliveira e Leon Sólon Silva

PUBLICADO POR LET'S DATA

[HTTPS://LETSDATA.AI](https://letsdata.ai) ↗

Este guia está protegido por leis de direitos autorais. Todos os direitos sobre o guia são reservados. Você não tem permissão para vender este guia nem para copiar/reproduzir o conteúdo do guia em sites, blogs, jornais ou quaisquer outros veículos de distribuição e mídia sem a devida citação da fonte. Qualquer tipo de violação dos direitos autorais estará sujeita a ações legais.

Primeira versão, 18 de agosto de 2021

Conteúdo

I

Apresentação

1	Introdução	8
1.1	Introdução	8
2	Quem é o Let's Data?	10

II

O que é Data Science?

3	Introdução	14
---	-------------------------	----

4	Histórico, contexto e definição	16
4.1	Histórico e contexto	16
4.2	Definições	22
4.3	Síntese	30
5	Aplicações em Ciência de Dados	33
5.1	Família e Vida pessoal	34
5.2	Marketing, Propaganda e Internet	35
5.3	Risco financeiro e Seguros	35
5.4	Saúde	36
5.5	Aplicação da lei	36
5.6	Detecção de fraude, Segurança e Eficiência logística	37
5.7	Governo, Política e Educação	38
5.8	Linguagem, Pensamento e Psicologia	39
5.9	Recursos Humanos	40
6	Ciência de Dados x Inteligência Artificial	41

III

O que aprender para se tornar um Cientista de Dados?

7	Introdução	45
8	Conhecimentos básicos	48
8.1	Noções Básicas de Bancos de Dados	48
8.2	Formatos de dados	50
8.3	ETL	52
8.4	Git / Github	53
8.5	Matrizes e Fundamentos de Álgebra Linear	54

8.6	CRISP-DM	55
9	Programação	58
9.1	Python	59
9.2	Pacotes Python	61
9.2.1	Pandas	61
9.2.2	NumPy	62
9.3	Jupyter Notebook	62
9.4	Ambientes virtuais	63
10	Análise Exploratória de Dados	65
11	Estatística	70
12	Visualização de Dados	74
13	Machine Learning	79
13.1	Introdução	79
13.2	Aprendizado Supervisionado	82
13.2.1	Regressão	82
13.2.2	Métricas de desempenho para regressão	83
13.2.3	Classificação	84
13.2.4	Métricas de desempenho para classificação	85
13.3	Aprendizado Não Supervisionado	88
13.3.1	Clusterização	90
13.3.2	Detecção de anomalias	91
13.3.3	Redução de dimensionalidade	91
13.4	Conceitos importantes	92
13.4.1	Overfitting e Underfitting	92
13.4.2	Dados de treino, validação e teste	94
13.4.3	Função de custo e Gradiente descendente	96

IV Mercado de trabalho para o Cientista de Dados

14	Vagas e salário	99
14.1	Brasil e Mundo	99
15	Outros tipos de trabalho	105
15.1	Freelance	106
15.1.1	Upwork	108
15.1.2	Codementor	110
15.1.3	Toptal	113
15.2	Consultoria	117

V Conclusão

16	Conclusão	119
16.1	Nossos canais	120
	Referências	121
	Livros	121
	Artigos	121
	Sites	122
	Índice remissivo	123



Apresentação

1	Introdução	8
1.1	Introdução	
2	Quem é o Let's Data?	10

1. Introdução

1.1 Introdução

Citação 1.1.1 *In our world of Big Data, businesses are relying on data scientists to glean insight from their large, ever-expanding, diverse set of data ... while many people think of data science as a profession, it's better to think of data science as a way of thinking, a way to extract insights using the scientific method.*

Em nosso mundo de Big Data, as empresas estão contando com cientistas de dados para obter insights de seu grande, cada vez maior e diversificado conjunto de dados ... embora muitas pessoas pensem na ciência de dados como

uma profissão, é melhor pensar na ciência de dados como uma forma de pensar, uma forma de extrair insights usando o método científico.

- Bob E. Hayes

Escrevemos este livro para responder três perguntas relevantes:

1. O que é Ciência de Dados?
2. O que aprender para se tornar um Cientista de Dados?
3. Como é o mercado de trabalho para o Cientista de Dados?

Cada uma dessas perguntas é uma parte deste livro. Escrevemos as respostas de uma maneira simples, para que você possa ler e aprender sem se sentir sobrecarregado. Você verá, no tópico seguinte, que essa foi a principal motivação não só deste livro existir, mas da criação do Let's Data.

2. Quem é o Let's Data?

O Let's Data nasceu da união de três pessoas - Leon Sólon, Felipe Schiavon e Bernardo Lago - com o propósito de tornar o aprendizado de Ciência de Dados simples e prazeroso. Mas, antes desse propósito, havia uma grande dor: “o que devo fazer para me tornar um cientista de dados?”.

Essa é uma dor muito comum na área devido a muitos fatores: 1) os conhecimentos exigidos de um cientista de dados são muito diversificados; 2) a área é relativamente nova e passa por um processo de amadurecimento constante; 3) a ciência de dados exige habilidades práticas com o foco de resolver problemas.

Bem, diante dessa selva com diversos obstáculos, o que se deve fazer para se tornar um cientista de dados? Na nossa

trajetória, encontramos algumas respostas. Mas, mesmo assim, não ficamos satisfeitos porque elas pareciam incompletas. Prosseguimos... e ao continuarmos nossa jornada, tivemos vários insights e fizemos descobertas de como as coisas poderiam ser mais simples. Bastava que a explicação de um determinado conceito (que aprendemos em cursos, livros, etc) fosse mais clara. Bastava entender a relação entre dois conteúdos que havíamos aprendido separadamente (isso acontece muito com Estatística x Programação, por exemplo). Enfim, o que “bastava” eram várias coisas, mas que eram totalmente passíveis de serem resolvidas. Decidimos mapear essas coisas todas e ajudarmos outras pessoas a fazerem essa jornada e se tornarem um Cientista de Dados!

Um pouquinho sobre nós:

- **León Sólon Silva** é bacharel em Ciência da Computação e mestre em Ciência de Dados pela Universidade de Brasília - UnB. Possui mais de 20 anos de experiência na área de TI e mais de 5 anos atuando como cientista de dados. Atualmente é Auditor-Fiscal da Receita Federal (trabalhando com ciência de dados) e cientista de dados da startup Bludworks.
- **Felipe Schiavon de Oliveira** é bacharel em Administração pela Universidade de Brasília - UnB, especialista em Gestão Estratégia de Pessoas pela Universidade Gama Filho, especialista em Inteligência Artificial e Machine Le-

arning pela PUC Minas e mestrando em Ciência de Dados na UnB. Atualmente trabalha no Tribunal de Justiça do Distrito Federal e Territórios, atuando como cientista de dados.

- **Bernardo Lago** é engenheiro eletricista pelo Instituto de Educação Superior de Brasília, pós-graduado em Data Science for Marketing pela Nova Information Management School - Nova IMS da Universidade Nova de Lisboa - UNL e mestrando em Estatística e Gestão da Informação, com ênfase em Marketing Research e CRM, também pela Nova IMS - UNL.



O que é Data Science?

3	Introdução	14
4	Histórico, contexto e definição	16
4.1	Histórico e contexto	
4.2	Definições	
4.3	Síntese	
5	Aplicações em Ciência de Dados	33
5.1	Família e Vida pessoal	
5.2	Marketing, Propaganda e Internet	
5.3	Risco financeiro e Seguros	
5.4	Saúde	
5.5	Aplicação da lei	
5.6	Detecção de fraude, Segurança e Eficiência logística	
5.7	Governo, Política e Educação	
5.8	Linguagem, Pensamento e Psicologia	
5.9	Recursos Humanos	
6	Ciência de Dados x Inteligência Artificial	41

3. Introdução

Citação 3.0.1 *Data Science is a process, not an event. It is the process of using data to understand different things, to understand the world.*

Ciência de dados é um processo, não um evento. É o processo de usar dados para entender diferentes coisas, para entender o mundo.

- Shingai Manjengwa

Neste capítulo, vamos falar do termo Ciência de Dados e vamos responder algumas perguntas:

- O que significa Ciência de Dados?
- Como esse termo surgiu?
- O que faz um Cientista de Dados?
- Quais são as aplicações de Data Science?
- Ciência de Dados é a mesma coisa que Inteligência Artificial?

Vamos começar! Segue o fio...

4. Histórico, contexto e definição

4.1 Histórico e contexto

Em 2012, uma matéria mudaria a história da Ciência de Dados! Ela foi publicada na Harvard Business Review e o título da matéria era: Data Scientist: The Sexiest Job of the 21st Century. Os autores dessa matéria são Thomas H. Davenport e D. J. Patil.

O subtítulo da matéria também é muito interessante: *Meet the people who can coax treasure out of messy, unstructured data* (Conheça as pessoas que podem extrair tesouros de dados bagunçados e não estruturados).

Bem, já tem muitas coisas interessantes que podemos analisar aqui:

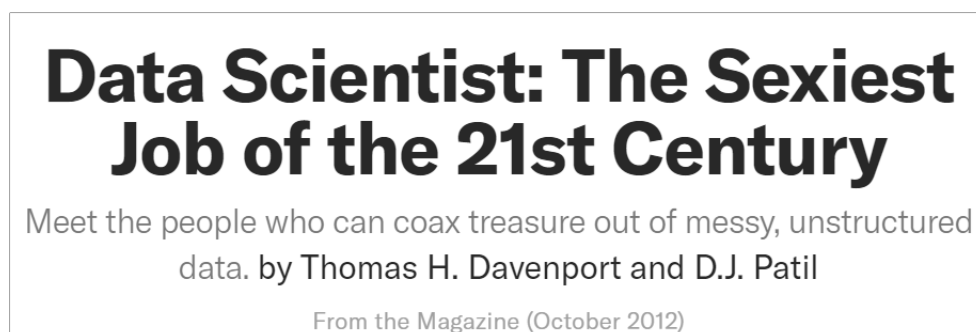



Figura 4.1: Artigo “Data Scientist: The Sexiest Job of the 21st Century”

- O artigo foi publicado numa revista de negócios, não numa revista de tecnologia ou de alguma área STEM. Guarda isso porque tem um ponto muito importante a esse respeito que vamos falar daqui a pouco.
- O título do artigo, que apesar de ter um apelo extremamente atrativo, tem o objetivo de mostrar o valor dessa carreira. É por isso que ela seria, então, considerada “The Sexiest Job do Século 21”.
- O subtítulo já nos mostra um objetivo da ciência de dados: extrair valor a partir de dados. E os autores da matéria já qualificam o tipo de dado... não são dados limpos, organizados, estruturados. São dados desorganizados, bagunçados. Então guarda isso também porque a gente vai falar do famoso Big Data daqui a pouco.

Vocabulário 4.1 STEM é uma sigla em inglês que significa Ciências Naturais, Tecnologia, Engenharia e Matemática (*Science, Technology, Engineering and Mathematics*).

E o que esses autores escreverem nesse artigo? Basicamente, tentaram definir o que um cientista de dados faz. Além disso, eles notaram que a demanda por cientistas de dados estava crescendo para além da disponibilidade de profissionais. Ou seja, além de a demanda já ser maior do que a oferta, essa demanda ainda estava crescendo a uma velocidade superior.

Lembre-se: estamos em 2012, data da publicação do artigo. Foi justamente naquele ano que o termo Data Science ganhou uma página na Wikipedia , inclusive!

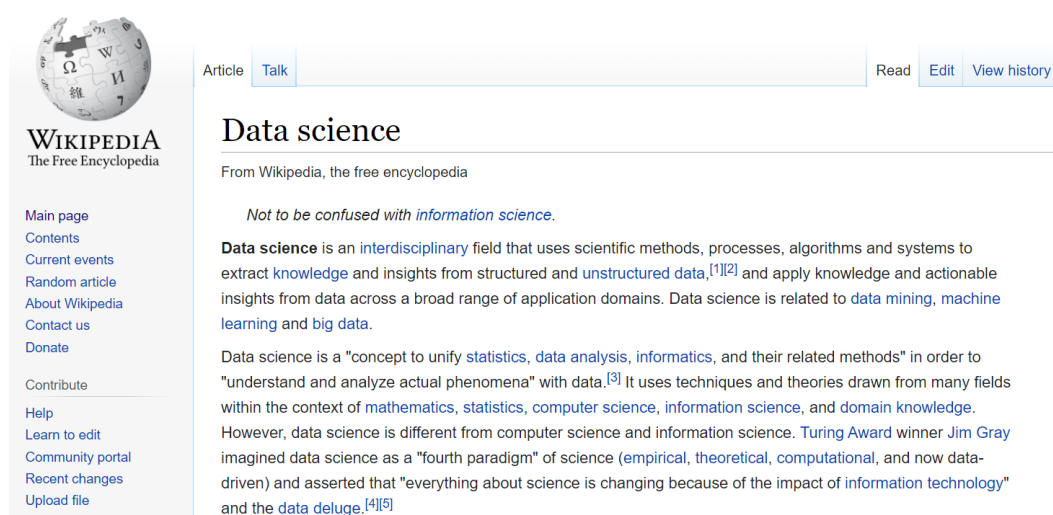


Figura 4.2: Wikipedia - Data Science

É a partir dessa época que podemos verificar o crescimento do interesse pelo termo Data Science. Vamos olhar o gráfico de busca desse termo nos EUA usando o Google Trends. Vejam que a partir de 2012 temos o início do crescimento das buscas.



Figura 4.3: Google Trends - “Data Science” nos EUA

Vamos olhar como é o mesmo gráfico no Brasil, comparando o gráfico nos EUA. Percebam que há uma diferença de quase 4 anos entre o início do crescimento lá e aqui no nosso país.

Quando usamos o termo em português - Ciência de Dados, essa diferença fica ainda maior. Diferentemente do que alguns podem ter pensado, apesar de o termo original ser em inglês, ele é ainda mais utilizado do que o termo em português, até hoje. Percebam que o interesse por "ciência de dados" só começou a crescer em dezembro de 2017.

Mas vamos voltar um pouco mais para entendermos a origem do termo **Cientista de Dados**. Ele foi criado em 2008, por D. J. Patil - ele foi um dos autores que escreveu o artigo que citamos anteriormente e trabalhava no LinkedIn na época - , e Jeff Hammerbacher - que trabalhava no Facebook.

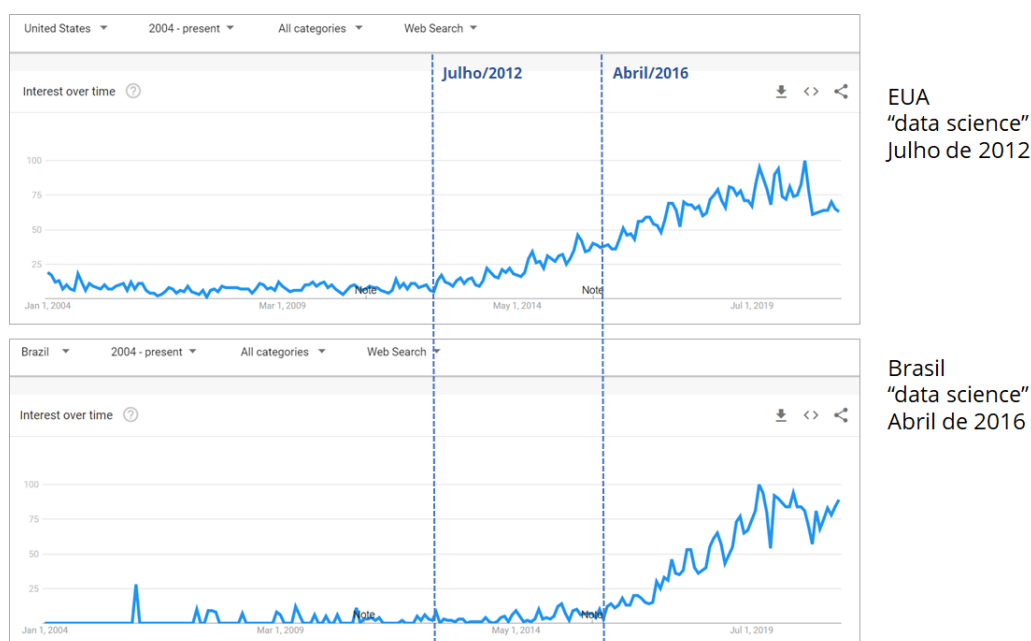


Figura 4.4: Google Trends - "Data Science" nos EUA e no Brasil

A partir daí é que o termo Data Scientist ou Cientista de Dados surgiu como cargo.

Ok. Então temos em 2008 esse termo novo, que virou uma nova profissão ou cargo - Cientista de Dados - e em 2012 esse artigo que disseminou e atraiu muita gente pra área por divulgar que era a profissão mais sexy do século XXI e que a demanda por cientistas de dados era bem maior que a oferta.

Também já sabemos que Ciência de Dados tem a ver com **extrair valor a partir de dados desorganizados e desestruturados**. Mas talvez você esteja pensando: mas a ciência, em sua maior parte, não trabalha com análise de dados? A expressão **Ciência de Dados** não seria uma redundância?

Bem, a verdade é que sim: grande parte do trabalho cien-

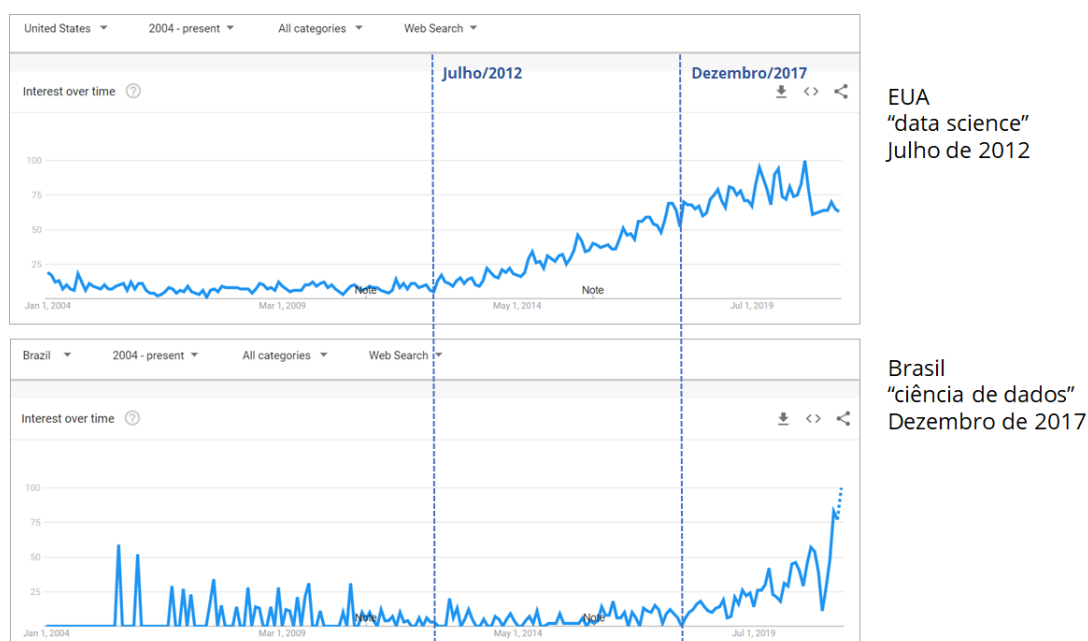


Figura 4.5: Google Trends - “Data Science” nos EUA e “Ciência de dados” no Brasil

tífico, do que chamamos de **ciência**, envolve coletar, tratar ou ao menos analisar dados. E a combinação dessas duas palavras - **Ciência** e **Dados** - não nos indica claramente o que a ciência de dados significa.

A verdade é que não existe uma definição única e unânime. Não há um consenso do que Data Science ou Ciência de Dados é. Talvez por ser um paradigma novo, talvez por ser uma área que abrange muitos domínios do conhecimento...

Na realidade, a Ciência de Dados, como termo, surgiu na academia há muito tempo, ainda nos anos 60. Mas embora o termo Ciência de Dados tenha sido cunhado originalmente na academia, a proliferação do seu uso foi impulsionada principalmente pela indústria de tecnologia e, com isso, seu

significado ganhou matizes diferentes. Não vamos trazer essa discussão detalhadamente no livro para não desviarmos do nosso objetivo. Quem quiser saber mais sobre esse ponto, sugerimos a leitura do artigo *Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards* [↗](#), publicado na HDSR - Harvard Data Science Review, que aborda a origem do termo na academia em detalhes.

Vamos, então, ver algumas definições para identificarmos os principais elementos do que podemos considerar como Ciência de Dados.

4.2 Definições

Vamos ver uma primeira definição de Ciência de Dados:

Definição 4.2.1 — Ciência de dados. *The study of extracting value from data.*

O estudo de extrair valor de dados.

- Jeannette Wing

Mas seria “só” isso? Lembra do subtítulo do artigo famoso de 2012? “Meet the people who can coax treasure out

of messy, unstructured data”. Até esse subtítulo do artigo diz mais, pois caracteriza que tipo de dado estamos falando. Mas, ainda sim, temos elementos parecidos: Data Science envolve extrair valor a partir de dados, como já dissemos antes. Mas esbarramos em uma questão ao conceituar Data Science dessa forma, pois um bom conceito deve não só definir algo, mas essencialmente distinguir esse algo de forma a torná-lo único, distinto. Acontece que essa definição, por ser muito ampla, pode ser confundida com o conceito de estatística. Vamos ver a definição de estatística de acordo com a American Statistical Association:

Definição 4.2.2 — Statistics. *The science of learning from data and of measuring, controlling, and communicating uncertainty.*

A ciência de aprender com dados e de medir, controlar e comunicar a incerteza.

- American Statistical Association

E agora? Ciência de Dados e Estatística são a mesma coisa? Com certeza não, apesar de que a Ciência de Dados precisa dos conhecimentos da Estatística. Está entendendo onde começa a confusão?

Bem, vamos relembrar que o termo **Cientista de Dados** surge a partir da experiência real, do dia a dia, tendo em vista a necessidade de lidar com grandes quantidades de dados e conseguir trabalhar em projetos em que essas tarefas com dados eram preponderante para o sucesso dos projetos. Foi a partir dessa necessidade que cunharam o termo **Cientista de Dados**. E que conhecimentos esses profissionais tinham que ter?

! Habilidades de programação e experiência em organizar e analisar conjuntos de dados complexos e desorganizados eram fundamentais!

Bem, nem todos os estatísticos ou cientistas da computação eram capazes de desempenhar esses papéis de forma completa. Um estatístico podia, por exemplo, ser capaz de formular uma tese, explicar modelos, mas nunca ter trabalhado com bases de dados reais. Um cientista da computação poderia saber escrever código, desenvolver sistemas, mas não trabalhar com análise de dados. Mas a gente já sabe que conhecimentos advindos dessas duas áreas de conhecimento são importantes: estatística e computação, mais especificamente, programação.

Bem, uma definição interessante e que traz essa dimensão computacional é a seguinte:

Definição 4.2.3 — Data Science. *Data science is an approach to data analysis with a foundation in code and algorithms.*

Ciência de dados é uma abordagem para análise de dados com base em código e algoritmos.

- Matthew Brett

Para analisar dados, precisamos dos conhecimentos de estatística. Por exemplo, qual a diferença de uma variável qualitativa ordinal para uma variável qualitativa nominal? Os dados de determinada base seguem que tipo de distribuição? O tamanho da minha amostra é suficiente para inferir algo de uma determinada população? Essas são perguntas básicas relacionadas à estatística e que são importantes na análise de dados.

Bem, mas temos nessa definição, também, o aspecto computacional. O conceito traz a informação “**com base em código e algoritmos**” . Um algoritmo nada mais é do que um conjunto de instruções a serem executadas por computador. É um conjunto finito de instruções onde temos entradas (*inputs*) e saídas (*outputs*). É só isso.

Um exemplo de algoritmo: o “algoritmo” do Instagram. O que é isso? É um conjunto de regras definidas pelo Instagram para determinar, com base em entradas (os *inputs*) como serão as saídas (os *outputs*). O que seriam as entradas? As informações do seu perfil e suas postagens.

Exemplos de entradas:

- Quantidade de seguidores do seu perfil.
- Nível de engajamentos dos seus seguidores (likes, comentários, compartilhamentos...).
- Média de postagens por dia.
- Formato de postagens: Post normal, Stories, IGTV, Reels?

Alguns exemplos de saídas:

- Pessoas para as quais o Instagram vai mostrar o seu conteúdo no feed dos seus seguidores.
- Pessoas que não te seguem e que o Instagram vai mostrar o seu conteúdo.
- Conteúdos que serão mostrados.

E por aí vai. Ok, algoritmo é isso. No caso de Data Science, o algoritmo vai usar dados de entradas que nós vamos fornecer para processar essas informações com base no algoritmo que a gente escrever ou escolher e ele vai nos devolver saídas. Vamos exemplificar com um problema.

PROBLEMA

Um banco tem uma base de dados de clientes com diversas informações: idade, sexo, estado civil, valor em conta bancária, dívidas, se tem filhos, há quanto tempo é cliente, etc. E o

banco quer avaliar se novos pedidos de empréstimo devem ser aprovados ou não, com base no histórico de inadimplência. Como descobrir isso?

Bem, o banco tem clientes que nunca pediram empréstimo. Como ele vai saber se a pessoa vai ou não cumprir com o pagamento? Ele pode usar ciência de dados pra isso. O que ele pode fazer é passar um conjunto de regras (algoritmo) para que o computador aprenda, com base no histórico de empréstimos anteriores, se os novos pedidos de empréstimo serão ou não honrados.

Vamos supor que o cientista de dados decida utilizar um algoritmo de Regressão Logística. Não se preocupe com o algoritmo, ok? Aqui a ideia é só tentar tornar mais concreto o processo. O que um algoritmo de Regressão Logística faz é a mesma coisa que qualquer algoritmo faz: pega entradas, processa e gera saídas. Nesse caso, quais seriam as entradas? Aquelas que falamos: qual a idade do sujeito? ele tem filhos? ele é casado, solteiro, divorciado ou viúvo? qual o salário dele? etc.

E qual é a saída do algoritmo? Se o banco deve ou não autorizar o empréstimo!

Viu como tudo se encaixa? Vamos voltar para o conceito: “Ciência de dados é uma abordagem para **análise de dados** com base em **código e algoritmos**.”

Vamos ver os aspectos destacados na definição:

Análise de dados Quem vai analisar os dados? Primeiro, obviamente, o cientista de dados! Ele precisa entender o que tem na base dele, se os dados estão corretos, que tipo de dados tem, se precisa arrumar alguma coisa na base de dados (99,99% dos casos precisa), etc. A partir daí, ele vai pedir ajuda do computador para continuar o processo, certo? Então seguimos para outro elemento do conceito.

Código e algoritmos Qual algoritmo o cientista de dados escolheu? No nosso exemplo, um algoritmo de regressão logística. Esse algoritmo vai receber os dados de entrada e vai retornar uma saída: 0 ou 1: 0 se o empréstimo não for aprovado e 1 se o empréstimo for aprovado.

E aí, o que você acha? Esse resultado extrai valor a partir dos dados? Claro que sim né! Para um banco, que busca maximizar seu lucro, ter inadimplência é péssimo. Se o cientista de dados consegue criar um modelo capaz de prever quem vai pagar e quem não vai pagar ao banco, ele gera muito valor, concorda?

Vamos ver um último conceito pra fechar essa primeira parte.

Definição 4.2.4 — Data Science. *Using Data to achieve specified goals by designing or applying computational methods for inference or prediction.*

Usar dados para atingir objetivos específicos, elaborando ou aplicando métodos computacionais para inferência ou predição.

- Usama Fayyad and Hamit Hamutcu

Veja que o conceito é muito parecido com o anterior. Mas o que é diferente? Ele explicita um elemento importantíssimo: **"para atingir objetivos específicos"**.

O objetivo da ciência de dados é resolver problemas de negócio. Ou seja, você tem que ter em mente que precisa atingir objetivos específicos e definidos. Dependendo do contexto, pode significar exploração, descoberta, tomada de decisão, previsão, otimização... Isso está muito relacionado ao método científico para construir conhecimento a partir de observações. Na prática, os objetivos são definidos pelos líderes da empresa ou pelos responsáveis por determinada unidade. No exemplo do banco, o objetivo poderia ser:

DIMINUIR O ÍNDICE DE INADIMPLÊNCIA DOS EMPRÉSTIMOS.

Ou, mais especificamente:

CRIAR UM MODELO PREDITIVO PARA DECISÃO DE APROVAÇÃO OU REPROVAÇÃO DE PEDIDOS DE EMPRÉSTIMO.

4.3 Síntese

Depois de passarmos por tudo isso, vamos sintetizar o que vimos até aqui:

1. Ciência de Dados é um campo de atuação multidisciplinar.
2. Ciência de Dados utiliza conhecimentos estatísticos, mas não é só estatística.
3. Ciência de Dados utiliza conhecimentos da computação, mas não é só ciência da computação.
4. Ciência de Dados trabalha com dados, especialmente com grandes quantidades de dados.
5. Ciência de Dados envolve fazer descobertas e/ou extrair valor.
6. Ciência de Dados surgiu cresceu dentro do contexto da indústria de tecnologia com o objetivo de resolver problemas de negócio.

Vamos reforçar, mais uma vez, esse aspecto dos problemas de negócio. Isso é extremamente importante! Se você tem conhecimentos de estatística e de programação, não necessariamente vai conseguir aplicá-los em projetos para auxiliar na consecução de objetivos definidos. Não necessariamente vai conseguir resolver problemas. Por isso existe um elemento muito importante para o Cientista de Dados que é o conhecimento do domínio no qual essas habilidades serão aplicadas.

Vamos ver mais um exemplo, dessa vez na área de saúde:

PROJETO: DESENVOLVER MODELO PREDITIVO CAPAZ DE IDENTIFICAR PACIENTES DIABÉTICOS COM RISCO DE AGRAVAMENTO DE QUADRO DE SAÚDE.

Esse projeto é real viu pessoal? Ele inclusive ganhou o prêmio Champions of Science - Storytelling Challenge – Latin America and the Caribbean Edition, da Johnson & Johnson [!\[\]\(9dfdaff1d86ba3c1f8353b4d1b61b8c5_img.jpg\)](#)

Bem, o conhecimento do domínio da saúde aqui é fundamental. O que é ter diabetes? Quais os tipos de diabetes? Nós temos vários dados de saúde que podem ser utilizados para criarmos esse modelo preditivo, mas o que significa cada um desses dados? Se há uma deficiência no conhecimentos de informações como essas, o trabalho de Ciência de Dados fica extremamente prejudicado, com a possibilidade de o modelo que está sendo desenvolvido ser falho.

Dessa forma, podemos pensar na ciência de dados como a união de três grandes área de conhecimento:

1. Estatística (e inserimos aqui, também, a Matemática, pois conhecimentos de Álgebra Linear e Cálculo, por exemplo, também são muito importantes).
2. Computação (especificamente a habilidade de programação).
3. Conhecimento/expertise no tema.

O tema pode ser qualquer um: Marketing, Recursos Humanos, Direito, Saúde, Esporte, etc.



Figura 4.6: Diagrama - Data Science (adaptado de Drew Conway)

5. Aplicações em Ciência de Dados

Neste capítulo trazemos, de forma resumida, exemplos reais de aplicações em Ciência de Dados. Os exemplos foram selecionados do livro *Análise Preditiva: O poder de prever quem vai clicar, comprar, mentir ou morrer*, do autor Eric Siegel. Apesar de ser um livro pequeno, ele é excelente pra quem quer ter um primeiro contato com a temática ou até mesmo aprofundar seus conhecimentos.

Os exemplos se referem a estas categorias:

1. Família e Vida pessoal
2. Marketing, Propaganda e Internet
3. Risco financeiro e Seguros
4. Saúde

- 5. Aplicação da lei
- 6. Detecção de fraude, Segurança e Eficiência logística
- 7. Governo, Política e Educação
- 8. Linguagem, Pensamento e Psicologia
- 9. Recursos Humanos

5.1 Família e Vida pessoal

1) Um exemplo clássico e que ficou muito famoso: a Target (que é uma imensa rede varejista nos EUA e que vende online) previu a gravidez das clientes a partir do comportamento de compras, identificando 30% mais pessoas para entrar em contato com ofertas relacionadas a necessidades de um recém-nascido. Isso gerou um problema gigante porque a Target muitas vezes sabia que a mulher estava grávida antes dos familiares! Você imagina o barulho que isso criou na época em que aconteceu, em 2012. Esse é um exemplo de aplicação da Target, mas eles implementaram diversos outros e conseguiram ampliar sua receita entre 15 a 30%.

2) O LinkedIn considera as sugestões de pessoas que você pode conhecer como o “o produto de dados mais importante que criaram”. Esse produto se enquadra em uma aplicação de Machine Learning conhecida como **sistemas de recomendação**.

5.2 Marketing, Propaganda e Internet

3) A Vermont Country Store, que é uma rede varejista de eletrônicos nos EUA, conseguiu direcionar o seu envio de catálogos usando Ciência de Dados e isso resultou numa receita 11 vezes maior do que o investimento utilizado nos envios.

4) O Google utiliza Data Science para classificar os e-mails como spam. Quem já usa o Gmail há algum tempo deve ter percebido como esse sistema de classificação melhorou. E quando você indica que a classificação não foi feita de forma correta, você ajuda no processo de aprendizado do algoritmo e ele se torna ainda melhor.

5.3 Risco financeiro e Seguros

5) A London Stock Exchange tem cerca de 40% das suas negociações conduzidas por algoritmos. Essa é uma grande área de aplicação da ciência de dados: investimentos em ações feitas com bases em sistemas computacionais baseados em algoritmos. É o que se chama de **trading com dados**.

6) A Chase, uma empresa do ramo financeiro e de seguros dos EUA, conseguiu ampliar sua receita identificando os donos de imóveis que iriam refinanciar suas hipotecas para captar esse interesse antecipadamente, antes que essas

pessoas procurassem outros bancos.

5.4 Saúde


7) O site Risk Prediction [↗](#) criou um modelo capaz de prever o risco de morte em um procedimento cirúrgico com base nos aspectos da pessoa, incluindo, por exemplo, idade, gravidade da doença e a existência de comorbidades. Ele foi criado para auxiliar cirurgiões a estimar o risco, para que ele possa esclarecer melhor os pacientes para que tomem a decisão pela cirurgia ou não.

8) A Universidade de Stanford criou um modelo para diagnosticar o desenvolvimento de câncer de mama considerando amostras de tecido.

5.5 Aplicação da lei

9) O Instituto de Tecnologia de Israel criou um modelo que prevê 51% dos tumultos com uma precisão de 91%. Com isso, eles conseguem tomar medidas preventivas considerando a possibilidade de agitação civil, promovendo mais segurança para as pessoas.

10) Em Oregon e na Pensilvânia juízes utilizam modelos de ciência de dados para auxiliar na decisão de manutenção da prisão de condenados, avaliando a chance de reincidência

dos presos. Veja: o modelo não decide nada... ainda é o juiz, mas ele tem a chance de se apoiar nessas métricas para formar sua decisão. E se você acha isso estranho, saiba que o ser humano é um ser cheio de vieses. Existe uma grande discussão ética sobre a questão dos algoritmos, como eles processam essas informações, e isso é muito importante! Mas trazemos uma outra reflexão aqui, que é: o ser humano também falha e utiliza critérios não racionais em suas decisões. Por exemplo, o ser humano pode julgar mais severamente um réu porque está com fome. Isso é real, ok? Claro que quem faz isso não tem consciência, mas, ainda assim, isso pode acontecer. Fora isso, o ser humano possui uma inúmeros vieses psicológicos. Se você tem interesse nesse assunto, sugerimos acessar o site [The Decision Lab](#) .

5.6 Detecção de fraude, Segurança e Eficiência logística

11) Na área de detecção de fraude, usa-se Data Science, por exemplo, em transações comerciais e sistemas de pagamento, para identificar comportamentos anômalos e evitar fraudes. Bem, todo mundo sabe que temos uma quantidade enorme de fraudes, especialmente agora no mundo virtual. O Brasil, por exemplo, é um dos países que tem recordes de transações fraudulentas. Então esse tipo de tecnologia, aqui, é ainda mais fundamental!

12) No Irã uma universidade criou um modelo capaz de prever a resistência do concreto com base em sua composição e mistura, gerando mais segurança para as construções.

13) A UPS, uma empresa de logística dos EUA, usou Ciência de Dados para otimizar as rotas de entregas de produtos. Dessa forma, conseguiram reduzir 85 milhões de milhas das entregas anuais. Isso equivale a 136 milhões de quilômetros! O que que vocês acham? Vale ou não vale a pena usar Ciência de Dados nas empresas? Olha a economia que ela pode gerar, fora outras externalidades positivas como entrega mais rápida para o comprador, menos trânsito, etc.

5.7 Governo, Política e Educação

14) Cambridge Analytica. Esse foi um escândalo ocorrido em 2019 e que gerou um filme: Privacidade Hackeada (The Great Hack). Quem não assistiu esse filme, tem que assistir! A Cambridge Analytica é uma empresa de análise de dados que trabalhou com o time responsável para campanha do Donald Trump nas eleições de 2016, nos Estados Unidos. A Cambridge Analytica teria comprado acesso a informações pessoais de usuários do Facebook e usado esses dados para criar um sistema que permitiu prever e influenciar as escolhas dos eleitores nas urnas.

15) A Universidade de Phoenix criou um modelo para

identificar os alunos com mais risco de reprovação. Com isso, consegue adotar medidas preventivas, como aconselhamento, e fazem um monitoramento mais individualizado do desempenho escolar desses alunos. Outra faculdade, a Rio Salado College, consegue medir com 70% de acurácia qual será o desempenho dos alunos após o oitavo dia de aula, baseado no seu comportamento online. Isso também é interessante para poder direcionar as atividades de ensino.

5.8 Linguagem, Pensamento e Psicologia

16) A IBM desenvolveu o Watson, um computador que derrotou dois campeões no programa de televisão Jeopardy! Esse programa é tipo um Show do Milhão, em que são feitas diversas perguntas de temas gerais e vence quem consegue acertar mais. Essa é uma área da Ciência de Dados conhecida como Processamento de Linguagem Natural (Natural Language Processing - NLP) que estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos. O objetivo é fornecer aos computadores a capacidade de entender e criar textos e falas.

17) A Universidade de Buffalo criou um modelo capaz de detectar mentiras com 82% de acurácia observando apenas os movimentos dos olhos da pessoa que está sendo observada. Olha que interessante! Mas imagina isso sendo usado na sua em casa!

18) A Online Privacy Foundation patrocinou uma competição para prever psicopatia a partir dos tweets das pessoas. Um dos projetos mais simples e interessantes de realizar no Twitter é análise de sentimentos: identificar se o conteúdo da mensagem é positiva ou negativa. Mas identificar psicopatia é um nível mais elevado de trabalho!

5.9 Recursos Humanos

19) A HP criou um modelo que gera um escore de “Risco de Evasão” para cada um dos funcionários da empresa, ou seja, ela pontua o seu risco de pedir demissão. O objetivo não é punir aqueles que possuem escore alto, mas atuar previamente, planejando ações para evitar a saída do funcionário. Com essa ação, eles estimaram uma economia de 300 milhões de dólares.

20) Por fim, mas não menos importante, o Comando de Combate Naval Especial nos Estados Unidos criou um modelo para prever se um candidato irá concluir a fase inicial do treinamento, de forma a auxiliá-los na decisão de contratação dos participantes. É um treinamento muito exigente em que menos de 25% dos participantes conseguem finalizar.

6. Ciência de Dados x Inteligência Artificial

Vamos finalizar este capítulo abordando, rapidamente, um tópico que é confuso pra muita gente. Ciência de Dados e Inteligência Artificial são a mesma coisa? Se não, quais são as diferenças?

Esses dois termos são muito utilizados de forma intercambiável. Inclusive, a maioria das empresas considera uma pessoa que trabalhe essencialmente com Inteligência Artificial como Cientista de Dados! Até nas seleções isso é feito.

A Ciência de Dados envolve uma série de etapas e procedimentos, como extração de dados, manipulação, visualização e manutenção de dados, para prever a ocorrência de eventos futuros. Porém, podemos expandir ainda mais essa dimensão e pensar em aplicações que mimetizam as capacidades

humanas, como o entendimento de uma fala (processamento de linguagem natural), o reconhecimento de pessoas em imagens (visão computacional), dentre outras. Nesse caso, temos a Inteligência Artificial. Ela envolve, normalmente, a criação de algo ou a reprodução de um comportamento humano, “como se fosse algo inteligente”. Isso normalmente é feito usando algoritmos de Machine Learning (Aprendizado de Máquina). E dentro do campo do Aprendizado de Máquina, temos outro subcampo chamado de Deep Learning (Aprendizado Profundo). O Deep Learning é um ramo de aprendizado de máquina baseado em um conjunto de algoritmos que tentam modelar abstrações de alto nível de dados usando várias camadas de processamento, compostas de várias transformações.

Resumidamente, temos a Figura 6.1 que mostra que a Ciência de Dados é uma grande área e, dentro dela, temos subáreas como Inteligência Artificial, descendo um pouco mais, Machine Learning, e, mais ainda, Deep Learning.

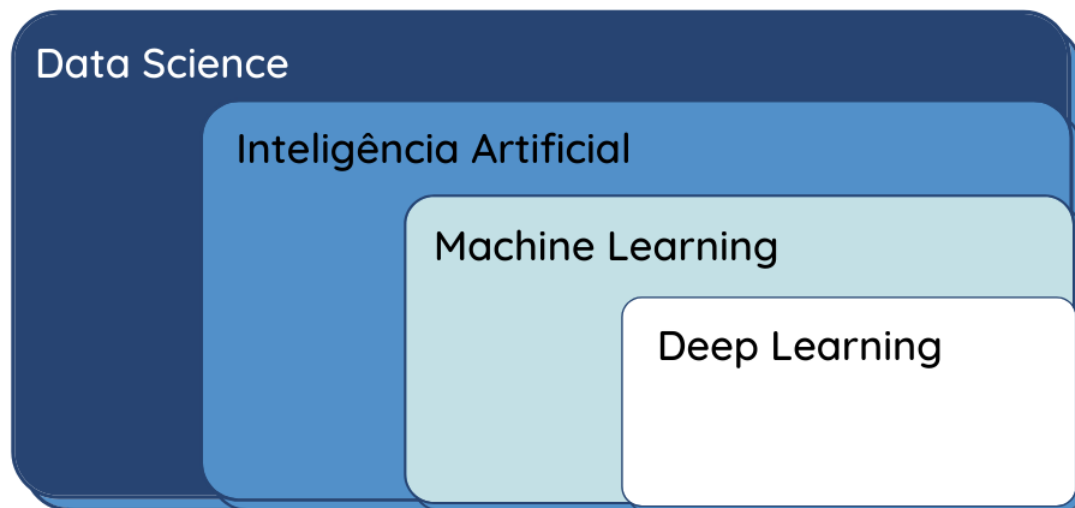


Figura 6.1: Data Science, Inteligência Artificial, Machine Learning e Deep Learning



O que aprender para se tornar um Cientista de Dados?

7	Introdução	45
8	Conhecimentos básicos .	48
8.1	Noções Básicas de Bancos de Dados	
8.2	Formatos de dados	
8.3	ETL	
8.4	Git / Github	
8.5	Matrizes e Fundamentos de Álgebra Linear	
8.6	CRISP-DM	
9	Programação	58
9.1	Python	
9.2	Pacotes Python	
9.3	Jupyter Notebook	
9.4	Ambientes virtuais	
10	Análise Exploratória de Dados	65
11	Estatística	70
12	Visualização de Dados	74
13	Machine Learning	79
13.1	Introdução	
13.2	Aprendizado Supervisionado	
13.3	Aprendizado Não Supervisionado	
13.4	Conceitos importantes	

7. Introdução

Essa é a pergunta que muita gente se faz e passa muito tempo tentando descobrir. Afinal, o que precisamos saber para nos tornarmos um Cientista de Dados?

Além de saber o que se deve aprender, é importante saber, também, em que nível se deve saber cada conhecimento. E, quando não se tem experiência no assunto, é impossível sabermos, ao certo, se o que aprendemos é suficiente. Às vezes, é até mesmo difícil saber se nós realmente sabemos determinado assunto!

Se você se sente assim, ou já se sentiu assim, bem vindo ao time! A gente já adianta pra vocês que essas perguntas são impossíveis de serem respondidas facilmente. A rigor, acreditamos que não existe uma única resposta.

Aqui neste livro, então, optamos por apresentar, em linhas gerais, os principais conhecimentos que um Cientista de Dados deve saber, organizados nestas categorias:

- Conhecimentos básicos
- Programação
- Fontes de dados
- Análise Exploratória de Dados / Manipulação de Dados
- Estatística
- Visualização de Dados
- Machine Learning

Nós apresentaremos, para cada categoria dessa, um diagrama com os tópicos que consideramos mais importantes. Vamos abrir algumas seções para falar mais sobre alguns deles, o que não significa que sejam os tópicos mais importantes. Significa, apenas, que entendemos ser importante explicar brevemente um pouco sobre o que aquele conteúdo significa, ok?

Além do mais, não se preocupe caso comece a pensar, em determinado momento: “Meu Deus, vou levar uma vida para aprender isso tudo!”. É normal! Uma coisa é certa: a carreira de Cientista de Dados exige bastante estudo, dedicação e esforço. Caso você não tenha o perfil de uma pessoa que goste de aprender, que seja motivado por desafios, que goste de resolver problemas, talvez não seja uma carreira boa para você. Existem perfis profissionais para todos os tipos de tra-

balho. Mas caso você esteja certo de que isso é pra você, não se preocupe tanto com os conhecimentos exigidos. Ninguém aprende tudo de uma vez e o aprendizado não é linear. O importante é ter constância e ritmo adequado na sua jornada!

Esperamos, então, que ao término da leitura desta parte do livro você seja capaz de listar, com mais clareza, o que te fará se tornar um Cientista de Dados! Vamos nessa.

8. Conhecimentos básicos

Neste capítulo mostraremos os conhecimentos básicos que um Cientista de Dados deve conhecer. Na Figura 8.1 você pode ver todos os conhecimentos dentro das “caixinhas”.

Vamos dar uma olhada mais de perto em alguns desses conhecimentos!

8.1 Noções Básicas de Bancos de Dados

Quando falamos de bancos de dados, uma das ferramentas mais usadas é SQL, que significa *Structured Query Language* (Linguagem de Consulta Estruturada). Ela permite acessar e manipular bancos de dados e se tornou um padrão do American National Standards Institute (ANSI) em 1986 e

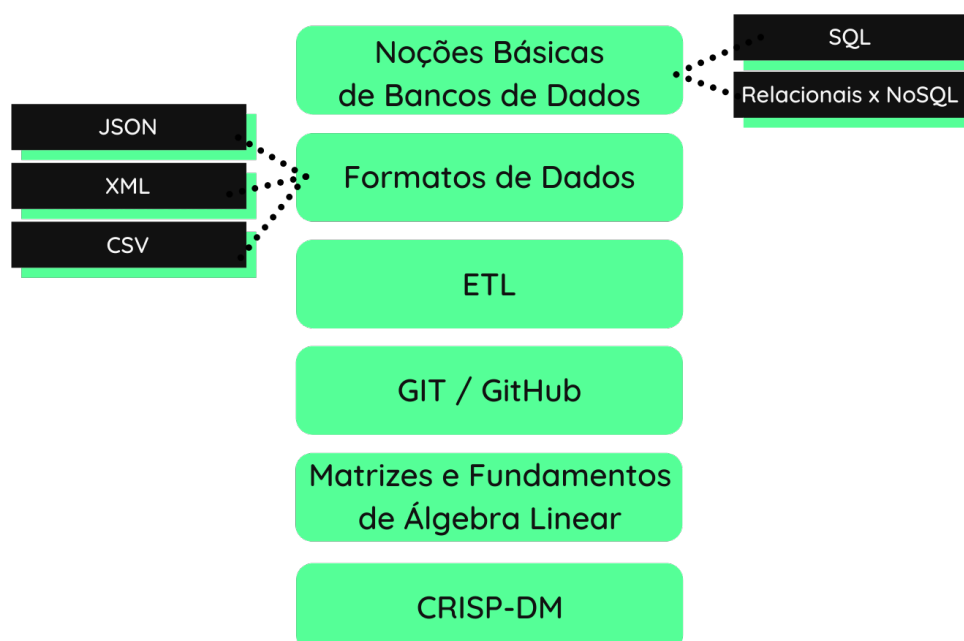


Figura 8.1: Conhecimentos básicos

da International Organization for Standardization (ISO) em 1987.

As consultas que realizamos nas bases de dados usando SQL servem para acessar as bases e obter os dados para serem utilizados nos projetos de Ciência de Dados.

Importante frisar que o Pandas (vamos falar dele no próximo capítulo) incorpora muitos conceitos de SQL, além de ter algumas funções que foram criadas a partir de estruturas de *queries* (consultas) SQL, como por exemplo a função `GROUPBY` (uma função que permite realizar o agrupamento de dados).

O SQL também é ótimo para entender como funcionam as tabelas de bases de dados, como podem ser feitas consultas

em várias tabelas ao mesmo tempo para retornar uma nova tabela com aquelas informações agrupadas. Também é muito bom para entender como funcionam Data Lakes e como ter boas práticas em criação e manutenção de bases de dados.

Importante frisar o domínio da função JOIN, pois é o pilar para consultas eficientes em SQL.

Além do SQL, que é a linguagem mais utilizada para se trabalhar com bancos de dados relacionais, também existe uma outra categoria de bancos de dados conhecidos por NoSQL.

Os bancos de dados NoSQL, também conhecidos como "Not Only SQL" (não apenas SQL) são não tabulares e armazenam dados de maneira diferente das tabelas relacionais.

Eles são amplamente usados em aplicativos da web em tempo real e Big Data, porque suas principais vantagens são alta escalabilidade e alta disponibilidade.

Os bancos de dados NoSQL também são a escolha preferida dos desenvolvedores, pois eles naturalmente permitem um paradigma de desenvolvimento ágil, adaptando-se rapidamente aos requisitos em constante mudança.

8.2 Formatos de dados

Quando vamos fazer a importação de uma base de dados, temos que saber qual é a origem desses dados. Eles podem vir em vários formatos. Os principais são: JSON, XML e CSV. Esses três tipos de arquivo são básicos, e entender suas

estruturas é importante.

JSON (JavaScript Object Notation) e XML (eXtensible Markup Language) são formatos de dados em markup. O JSON é derivado do JavaScript e o XML derivado do HTML. Os dois tem estruturas parecidas com as linguagens de programação de onde foram inspiradas e funcionam de forma similar. A seguir temos um exemplo dos dois tipos de arquivos.

JSON

```
{"TheBeatles": [
    { "firstName": "Paul", "lastName": "McCartney" },
    { "firstName": "John", "lastName": "Lennon" },
    { "firstName": "George", "lastName": "Harrison" },
    { "firstName": "Ringo", "lastName": "Starr" }
]}
```

XML

```
<TheBeatles>
  <Beatle>
    <firstName>Paul</firstName> <lastName>McCartney</lastName>
  </Beatle>
  <Beatle>
    <firstName>John</firstName> <lastName>Lennon</lastName>
  </Beatle>
  <Beatle>
    <firstName>George</firstName> <lastName>Harrison</lastName>
  </Beatle>
  <Beatle>
    <firstName>Ringo</firstName> <lastName>Starr</lastName>
  </Beatle>
```

Os formatos são parecidos e bem lógicos em termos de

estutura, mas o JSON tem um formato mais enxuto e mais simples de ler e escrever. Ele também suporta arrays, diferentemente do XML.

Além desses formatos, há ainda o CSV (Comma Separated Value), que é um formato de “tabela”, onde usualmente cada célula é separado por uma vírgula (,) e cada linha separada por um sinal de quebra de linha.

CSV

```
firstName , lastName  
Paul , McCartney  
John , Lennon  
George , Harrison  
Ringo , Starr
```

8.3 ETL

ETL significa *Extract, Transform and Load*, que corresponde às etapas de Extração, Transformação e Carregamento usadas para combinar dados de diversas fontes. Ele é comumente utilizado para construir um Data Warehouse.

Vocabulário 8.1 — Data Wharehouse. Um Data Warehouse é um repositório central de dados que podem ser analisados para tomar decisões mais adequadas. Os dados fluem de sistemas transacionais, bancos de dados relacionais e de outras fontes para o Data Warehouse. A palavra *Warehouse* significa “armazém, depósito”. Então pense no Data Warehouse como

um depósito de dados.

Nesse processo de ETL, os dados são retirados (extraídos) de um sistema-fonte, convertidos (transformados) em um formato que possa ser analisado e armazenados (carregados) em um Data Warehouse ou outro sistema, de modo a aprimorar a performance.

8.4 Git / Github

Primeiramente vamos deixar claro que Git e Github são duas ferramentas diferentes.

O Git é uma ferramenta de **controle de versão** gratuita, de código aberto, projetado para lidar com tudo, desde projetos pequenos a muito grandes, com velocidade e eficiência.

Vocabulário 8.2 — Controle de versão. O controle de versão permite gerenciar diferentes versões de um documento. Com isso, é possível gerenciar o histórico de alterações em um código que está sendo escrito, fazer o backup caso seja preciso resgatar uma versão anterior, desenvolver paralelamente (com mais de uma pessoa), o que gera segurança para o projeto.

O Github é um site baseado em Git que serve muito bem para o armazenamento dos seus projetos, pois permite que você use todas as funcionalidades do Git, mas também per-

mite que sejam hospedados os projetos dentro do site, além de renderizar Jupyter Notebooks (falaremos dele no próximo capítulo). Com a adição da nova funcionalidade *Github Pages*, você pode montar o seu portfólio todo dentro do Github com uma apresentação profissional do seu trabalho e dos seus projetos pessoais. Curiosidade: a Microsoft comprou o GitHub em 2018 por US\$ 7,5 bilhões.

8.5 Matrizes e Fundamentos de Álgebra Linear

Álgebra linear é um ramo da matemática que surgiu do estudo detalhado de sistemas de equações lineares. A álgebra linear utiliza alguns conceitos e estruturas fundamentais da matemática como vetores, espaços vetoriais, transformações lineares, sistemas de equações lineares e **matrizes**. De forma bem simplificada, a álgebra linear te ajuda a compreender espaços geométricos como planos em dimensões mais altas e executar operações matemáticas nesses planos.

Matriz é uma tabela organizada em linhas e colunas no formato $m \times n$, onde m representa o número de linhas (horizontal) e n o número de colunas (vertical).

A função das matrizes é relacionar dados numéricos. Por isso, o conceito de matriz não é só importante na Matemática, mas também em outras áreas como em Data Science, já que diversos algoritmos usam dados matriciais para suas operações.

8.6 CRISP-DM

O CRISP-DM é uma metodologia que significa *Cross Industry Standad Process for Data Mining* (Processo Padrão Inter-setorial para Mineração de Dados) e foi criada no intuito de estabelecer etapas para um projeto de Data Science (ou Data Mining), sempre com o foco nos objetivos finais do projeto. Ou seja, não se pode esquecer que aquilo que é importante para que o projeto está sendo desenvolvido e não se perder tempo de trabalho em ações que não vão agregar valor para o resultado final.

É uma metodologia cíclica, onde as etapas se complementam e podem ser revisitadas sempre que for necessário, como vemos na Figura 8.2.

Vamos ver um pouco sobre cada uma de suas etapas:

Entendimento do negócio, onde vão ser tratados os objetivos do projeto, os critérios de sucesso, o que vai ser alocado de recursos ou o que vai ser contingenciado, quais são os objetivos do processo de Data Mining e fazer o planejamento estrutural do projeto.

Entendimento dos dados, onde é primeiramente feita a coleta dos dados, logo após a descrição desses dados, bem como a exploração e verificação de qualidade da base de dados.

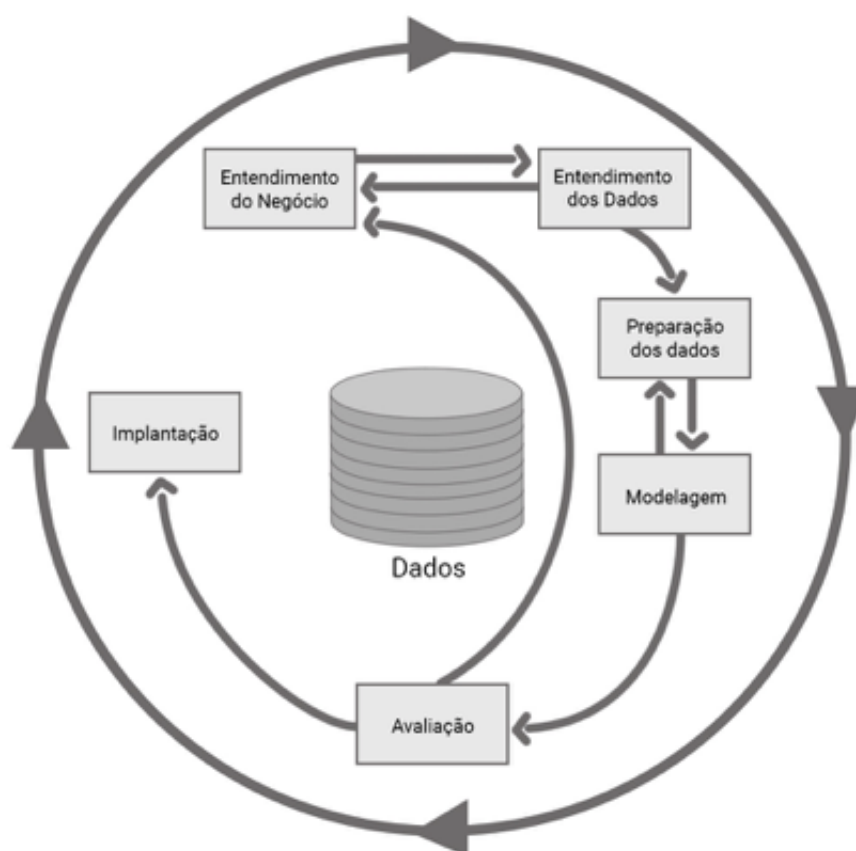


Figura 8.2: CRISP-DM (traduzido de CRISP-DM 1.0 - Step-by-step data mining guide - SPSS)

Preparação dos dados, a parte mais longa do projeto, onde demandamos tempo para tratar de seleção de dados, limpeza, tratar de outliers, transformação, construção de novas variáveis com feature engineering, integração de bases de dados e formatação de valores de acordo com as necessidade dos projetos e dos algoritmos.

Modelagem. Feita a escolha da técnica de modelagem, o

desing teste do modelo, a sua construção e avaliação, e também nesta etapa é onde vários modelos diferentes podem ser testados e avaliados a sua performance e a sua aplicabilidade.

Implantação, onde o modelo é colocado em produção, e ainda tratado de planejamento de implantação, monitoramento e manutenção, produção do relatório final e a revisão geral do projeto.

9. Programação

Neste capítulo apresentamos os conhecimentos de programação necessários para um Cientista de Dados. Vamos destacar a linguagem Python, pois ela é a mais utilizada, aceita e demandada pelo mercado. E isso ocorre não apenas na área de Ciência de Dados, mas na área de desenvolvimento em geral! Ou seja, escolher aprender Python é um tiro certo!

Dentro da temática **Programação**, elaboramos a Figura 9.1 com os conhecimentos que são importantes para sua carreira de Cientista de Dados. Vamos falar um pouco mais sobre alguns deles, o que não significa que sejam os mais importantes. Queremos apenas esclarecer alguns tópicos.

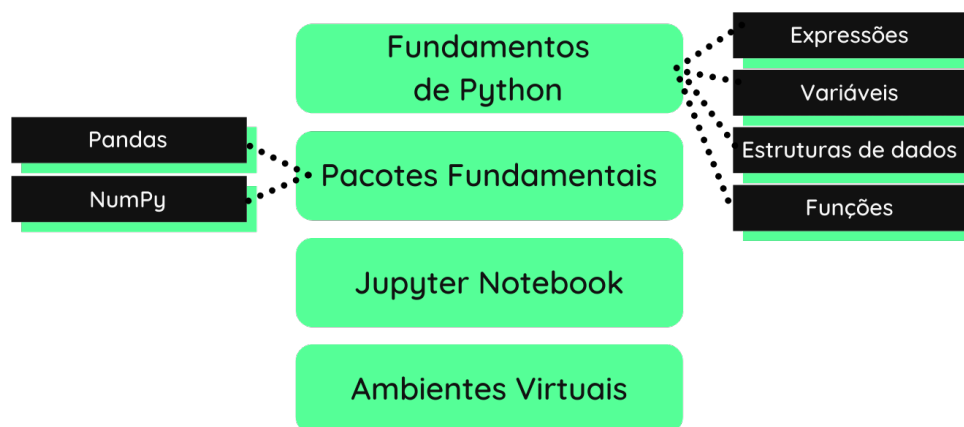


Figura 9.1: Conhecimentos de Python

9.1 Python

Segundo a IEEE Spectrum, a principal revista do IEEE (a maior organização profissional do mundo dedicada à engenharia e às ciências aplicadas), Python está no topo da lista de popularidade das linguagens de programação em 2021, conforme pode ser visto na Figura 9.2.

É ela que optamos por utilizar nossas aulas e nossos projetos. Além de Python, R também é muito difundido em Data Science, mas como Python tem bibliotecas mais completas, além de ser uma linguagem de programação generalista que não é focada em um tipo específico de uso, como o R que tem aplicação mais específica para estatística.

A maneira que a linguagem Python é distribuída facilita muito a vida dos Cientistas de Dados, porque seus pacotes são distribuídos de maneira gratuita e possuem fácil instalação através do instalador de pacotes “pip”.

Para instalar o Python, nossa ferramenta preferida é o Ana-

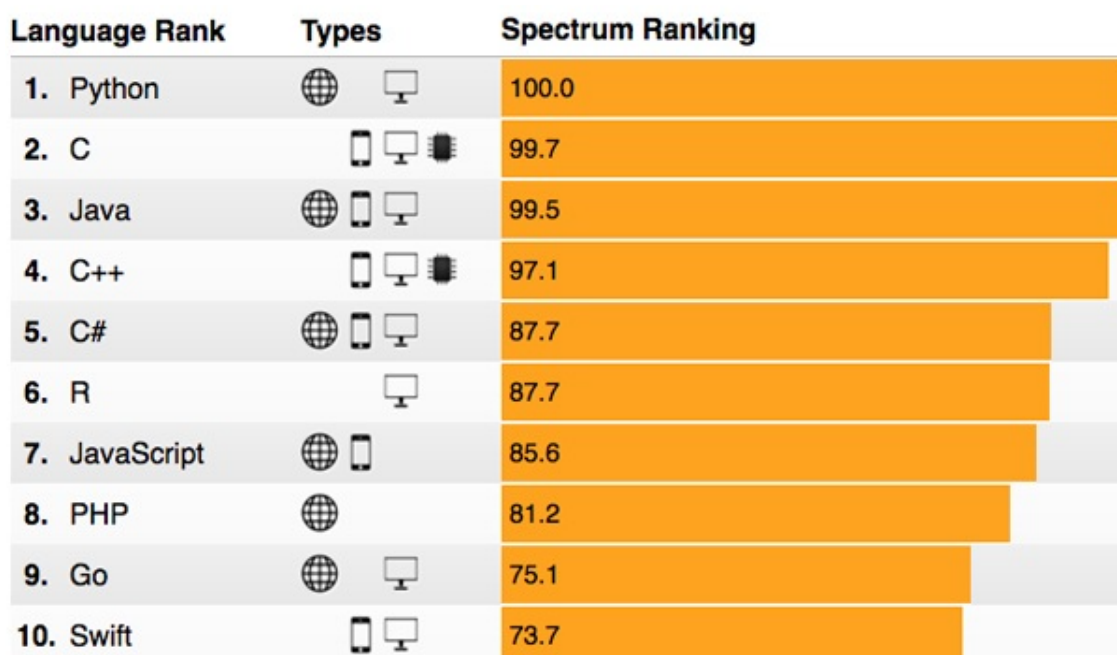


Figura 9.2: Ranking de popularidade das linguagens de programação em 2021

conda [↗](#). Nela é possível instalar pacotes, controlar versões dos próprios pacotes, instalar o Python, Jupyter Notebooks, entre outras ferramentas, e ainda criar **ambientes virtuais** para os projetos.

Vocabulário 9.1 — Ambientes virtuais. São pequenos espaços virtuais dentro da máquina onde você pode escolher quais pacotes e versão instalar, para ter sempre um ambiente limpo e organizado para os seus projetos.

9.2 Pacotes Python

Os pacotes python são funções ou conjunto de funções criadas pelos usuários que são compartilhados no repositório python, como falamos anteriormente, através do gerenciador de pacotes “pip”. Há vários pacotes focados em visualização de gráficos, alguns que são bem completos para Data Science, e outros específicos para tratar de base de dados, vamos tratar um pouco sobre alguns e principais deles aqui.

No nosso canal do YouTube tem um vídeo chamado **Preparando o Ambiente para usar Python para Data Science | Conda, Anaconda, pip, Spyder...** [🔗](#) onde ensinamos a instalar o Python puro e pelo Anaconda e explicamos a diferença e o funcionamento do Conda, Anaconda e pip, além de falar sobre os principais ambientes de desenvolvimento – IDEs para Data Science.

9.2.1 Pandas

O pacote Pandas é provavelmente o primeiro que você vai importar em cada projeto seu. Ele é quem determina funções para trabalhar com base de dados. Para importar a base para o projeto, fazer uma análise exploratória, conhecer a base de dados, feature engineering, etc. Ou seja, é um pacote muito robusto e o mais utilizado em Data Science. Além disso, é muito leve e rápido de se trabalhar.

Você pode saber mais sobre o Pandas assistindo ao vídeo **O que é o Pandas? Por que e como usar o Pandas no**

Python?, disponível no nosso canal do YouTube [↗](#).



Figura 9.3: Logomarca do pacote Pandas

9.2.2 NumPy

O NumPy é um pacote de Python que suporta operações com vetores e matrizes e é essencial para a computação científica. Quando precisamos trabalhar com arrays, matrizes e precisamos de álgebra linear, esta é a biblioteca chave, também em análise exploratória. Muito simples de usar, também leve como Pandas, é uma das bibliotecas mais usadas em Data Science.



Figura 9.4: Logomarca do pacote NumPy

9.3 Jupyter Notebook

Jupyter Notebook é um ambiente focado em “contar uma história” e muito útil quando é preciso compartilhar uma

análise de dados, uma modelagem, ou o que quer que seja no seu projeto. Como ele tem uma estrutura de células, e essas células podem ser executadas de forma independente, isso facilita muito o trabalho de Análise Exploratória de Dados (vamos falar sobre ela no próximo capítulo), por exemplo, além de ser perfeito para organizar o seu projeto.

Além disso, após cada execução de célula com o código, os resultados referentes a essa célula ficam salvos de maneira visual. Por isso, é uma boa ferramenta para compartilhar com outras pessoas o seu projeto. Além disso, o GitHub, por exemplo, tem um suporte a Jupyter Notebooks, então se for feito um upload de um notebook para o GitHub, sua visualização vai ser renderizada como estiver quando for feito o upload.

Curiosidade: O nome Jupyter é um acrônimo criado a partir das linguagens de programação que inicialmente foram aceitas pelo Projeto Jupyter: Julia, Python e R.

Caso queira saber mais sobre o Jupyter Notebook, tem um vídeo no nosso canal do YouTube chamado **Dominando o Jupyter Notebook | Funcionalidades e 20 Atalhos Mata-dores** [↗](#) que explica a história, as funcionalidades e os atalhos que otimizam o seu trabalho como Cientista de Dados.

9.4 Ambientes virtuais

Quando trabalhamos com Python, o mais comum é que precisemos usar pacotes e módulos que não fazem parte da

biblioteca padrão. E, muitas vezes, precisamos de versões específicas de um pacote para que um determinado algoritmo, sejam pelo fato de ele trabalhar com outros pacotes que só funcionam a versão de um outro pacote, seja porque há algum bug em diferentes versões que comprometem a execução do código.

O que isso significa? Que pode não ser possível que uma instalação do Python atenda aos requisitos de cada projeto de Data Science. A solução para esse tipo de problema é criar um **Ambiente Virtual**, um espaço independente que contém uma instalação específica do Python, além de vários pacotes com versões específicas.

Dessa forma, é possível definir ambientes virtuais para diferentes projetos, seja um ambiente virtual para cada projeto ou para vários projetos.

Temos duas ferramentas que são comumente utilizadas para a criação e gestão de ambientes virtuais. O módulo **venv**, que fornece suporte para a criação de Ambientes Virtuais leves com seus próprios diretórios de site, opcionalmente isolados dos diretórios de site do sistema, e o **Conda**, que, além de ser uma ferramenta de gerenciamento de ambientes, também é uma ferramenta de gerenciamento de pacotes, de código aberto e que roda em Windows, macOS e Linux.

10. Análise Exploratória de Dados

A Análise Exploratória de Dados, chamada de *EDA - Exploratory Data Analysis* em inglês, é uma etapa importante em qualquer projeto de Ciência de Dados. É o processo de investigação do conjunto de dados para descobrir padrões e anomalias, e formar hipóteses com base em nossa compreensão do conjunto de dados.

Para isso, utilizamos de conhecimentos de estatística (tema do próximo capítulo), como valores como máximo e mínimo, distribuição, média, mediana, moda, análise estatística, box-plot, histogramas. Tudo isso vai ser analisado durante a Análise Exploratória de Dados e os conceitos de estatística serão aplicados para essa finalidade.

A Figura 10.1 apresenta alguns conhecimentos importan-

tes para a realização da EDA.

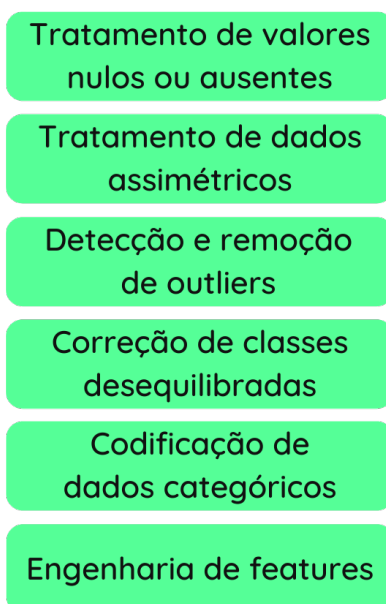


Figura 10.1: Análise Exploratória de Dados

Tratamento de valores nulos ou ausentes é o processo de analisar e tratar os dados quando há valores nulos ou ausentes, dependendo da proporção de valores nulos ou ausentes. Esse não é um processo simples em que há uma regra única que possa ser aplicada em todos os projetos. A título de exemplificação, compartilhamos uma abordagem simples e prática. Caso a variável tenha mais de 30% de valores ausentes, é necessário verificar a importância dessa feature. Se ela for importante, precisamos usar alguma técnica estatística para preencher os valores ausentes (como média, mediana ou moda). Se não for importante, pode-se excluir a feature da análise.

Tratamento de dados assimétricos envolve a análise de

dados assimétricos, muito comuns em bases de dados. A assimetria é o grau de distorção de uma distribuição normal. Basicamente, temos assimetria positiva ou assimetria negativa. Para remover a assimetria, os dados podem ser normalizados ou escalados. Isso é importante essencialmente quando esses dados forem ser utilizados para a modelagem em Machine Learning, pois alguns algoritmos não conseguem lidar de maneira efetiva com dados assimétricos.

Deteccção de outliers é o processo de identificar pontos de dados que estão distantes dos outros pontos de dados. Na estatística, um outlier é um ponto de observação que está distante de outras observações. Existem vários métodos para identificar outliers na base de dados, como o intervalo interquartil (IIQ) - uma medida de variabilidade - e o uso de boxplots - representação gráfica de dados numéricos por meio de seus quartis.

Correção de classes desequilibradas envolve avaliar e corrigir, quando necessário, a existência de uma classe que domina (com mais números de valores de dados) o conjunto de dados. Por exemplo, uma análise de dados que envolve uma variável se uma pessoa é fumante ou não fumante. Caso a proporção de não fumantes seja muito superior à de fumantes, um modelo de Machine Learning pode ser afetado pelas classes em desequilíbrio, gerando um resultado tendencioso. Recomenda-se fazer uso da Matriz de Confusão para casos como esse e analisar algumas medidas de desempenho importantes, como acurácia, precisão, *recall* (revocação) e

especificidade.

Caso queira saber mais sobre a Matriz de Confusão, tem um vídeo no nosso canal do YouTube chamado **Matriz de Confusão | Explicação e exemplos práticos** [↗](#) que explica bem o seu funcionamento e uso.

Codificação de dados categóricos é o processo de transformar dados categóricos em dados contínuos. Isso facilita a análise e o processamento desses dados em modelos de Machine Learning que só aceitam dados contínuos como *input*, como regressão linear, regressão logística, SVM, etc. Duas técnicas que são comumente utilizadas para essa finalidade são a codificação de rótulo - transformar rótulos não numéricos em rótulos numéricos - e a conversão de intervalos numéricos (que correspondem a dados categóricos) em números, usando alguma técnica estatística como a média ou a moda.

Por fim, a **Engenharia de features** (*Feature engineering*) é o processo de criação de novas features ou variáveis a partir dos dados brutos, capturando informações adicionais que não são facilmente aparentes na base de dados original. Assim, consegue-se uma melhoria na análise e no desempenho dos dados no modelo de Machine Learning. Também podemos citar outra técnica que está intrinsecamente relacionada à Engenharia de features, chamada de Seleção de features (*Feature selection* ou *Feature extraction*), que é o processo de escolher as principais features, a partir da base de dados

principal, e reduzir a dimensionalidade do problema de treinamento. Normalmente, realiza-se a Engenharia de features primeiro, para gerar features adicionais e, em seguida, a Seleção, para eliminar features desnecessárias, redundantes ou altamente correlacionadas.

Antes de finalizarmos este capítulo, ressaltamos, mais uma vez, que o trabalho de Análise Exploratória de Dados não se resume apenas ao que foi apresentado aqui, pois ela não é um processo com passos pré-definidos. Cada base de dados e cada projeto, com seu respectivo objetivo, exige uma abordagem diferente. Mas, para fins didáticos, trouxemos conhecimentos e tópicos relevantes para a grande maioria de trabalhos exploratórios e que serão úteis para o seu arsenal de habilidades como Cientista de Dados.

11. Estatística

Neste capítulo entramos numa seara onde dividimos os adultos dos adolescentes/crianças (profissionalmente falando). É aqui que muita gente se assusta, começa a roer as unhas e começa a pensar “O que que eu estou fazendo da minha vida?”.

Entender como um algoritmo funciona não é complicado, sua implantação não é complicada, mas entender os conceitos estatísticos dos algoritmos, e até mesmo antes disso, entender como uma distribuição funciona e como os dados devem ser trabalhados não é uma tarefa fácil. Mas não significa que seja ruim, certo? Nem tudo que é fácil é bom, e nem tudo que é difícil é ruim.

Inclusive, falaremos na Parte IV deste livro sobre o mer-

cado de trabalho para o Cientista de Dados. Uma das coisas que você vai ver lá (um pequeno spoiler, se nos permitem) é que é uma carreira muito valorizada e que, considerando todo o mercado, paga muito bem. Por que será? E por que tem empresas com vagas de Cientistas de Dados que não são preenchidas?

A verdade é que a grande quantidade de conhecimentos exigidos do Cientista de Dados cria uma barreira. Não é fácil aprender tanta coisa em pouco tempo. É um processo contínuo e, como falamos anteriormente, o mais importante é a constância, a regularidade, não a velocidade. Por isso, saiba de uma coisa: o que vai diferenciar o seu papel como Cientista de Dados é ter sólidos conhecimentos de estatística. Se você tem medo, não se desespere: apesar de não ser um assunto muito rápido e fácil de assimilar, é interessantíssimo e vai trazer benefícios reais para a sua vida, não apenas profissional, mas pessoal também! E se fosse muito fácil ser um Cientista de Dados, a carreira não seria tão valorizada como é.

A Figura 11.1 mostra os principais conhecimentos de estatística que você deve saber. Não é uma lista exaustiva, mas uma seleção de tópicos relevantes, alguns com um segundo nível para facilitar o seu entendimento.

Como é uma lista extensa, recomendamos conhecer os tópicos superficialmente primeiro para, então, adentrar mais detidamente em cada um deles para aprofundamento. É importante, por exemplo, entender o que é probabilidade, suas

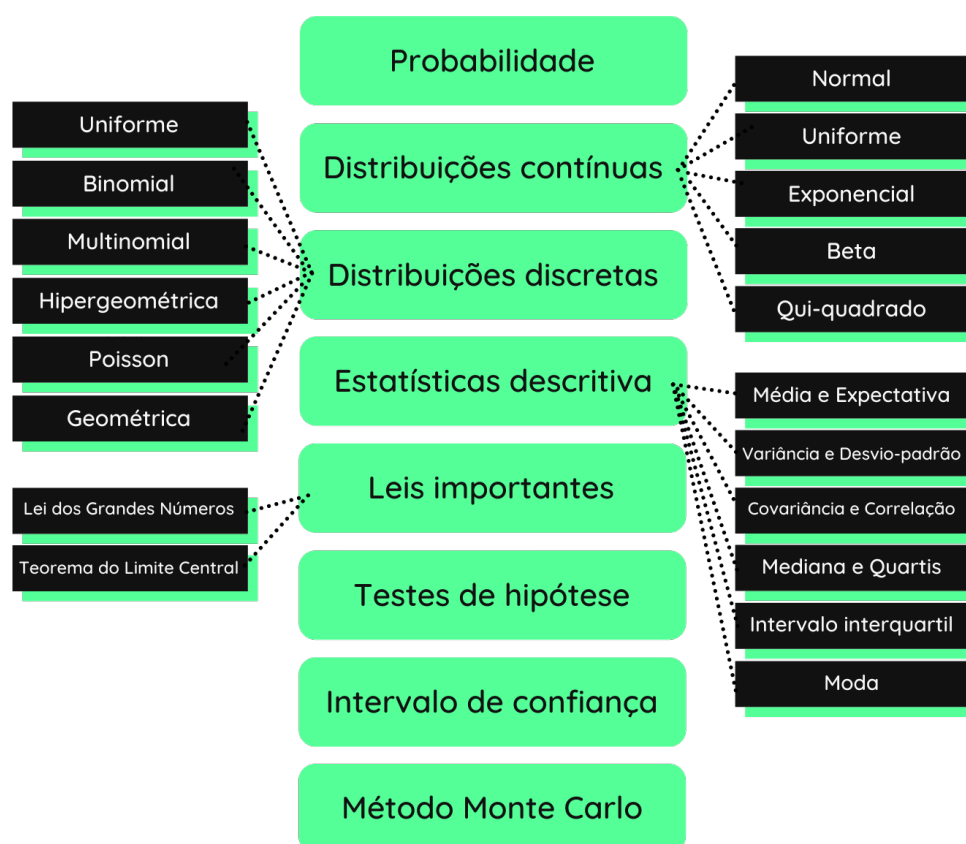


Figura 11.1: Estatística

principais leis, tipos, usos, etc. Preocupe-se em compreender a **intuição** por trás de cada técnica. Essa é uma atenção que poucas pessoas dão, pois nos acostumamos a aprender para fazer provas e avaliações (copiamos as fórmulas, substituímos os valores, encontramos o resultado e marcamos a opção correta). Na vida do Cientista de Dados isso é o menos importante! Precisamos entender os dados e decidir que técnicas estatísticas serão utilizadas, em cada caso. Nessa situação, o problema não está dividido em assuntos (como fazemos didaticamente ao apresentar um conteúdo) e você

precisa saber não só usar determinadas técnicas, mas explicar o porquê delas terem sido escolhidas.

Ora, só é possível fazer isso se você entender o que cada um desses conhecimentos significa verdadeiramente. Por exemplo, alguns algoritmos preditivos são sensíveis a distribuições que não estejam normalizadas. Assim se os dados não estiverem dentro de uma determinada distribuição, os resultados do algoritmo podem estar enviesados ou não serem precisos. Por isso, realizar transformações de distribuição são importantes.

Outra exemplo: os testes estatísticos com o p-valor determinam se determinada variável é relevante para o modelo. É importante entender o seu conceito para não cair na armadilha de usar variáveis irrelevantes e o modelo não ter a qualidade que deveria.

Mas antes de aprender a calcular o p-valor, entenda para que ele serve, qual a sua importância, qual o risco de não utilizá-lo, que problemas ele resolve, e qual a **intuição** do seu funcionamento - por que ele funciona? Isso é muito mais importante que aplicar fórmulas. Afinal, o computador fará os cálculos para nós. E uma fórmula a gente pesquisa e encontra rapidamente na internet. Concentre-se em entender, não em repetir.

12. Visualização de Dados

Muita gente acha que o conhecimento de Visualização de Dados é pouco importante. E não poderiam estar mais errados! É uma área de trabalho do Cientista de Dados tão relevante que alguns profissionais estão se especializando na arte de criar belíssimas e instrutivas visualizações.

Citação 12.0.1 *The most important thing about a chart is not its aesthetics, the technology used to create it, the kind of data visualization layout or even the data it represents. The most important thing about a chart is its impact. Impact is what a chart does.*

A coisa mais importante sobre um gráfico não é sua estética, a tecnologia usada para criá-lo, o tipo de layout de visualização de dados ou mesmo os dados que ele representa. A coisa mais importante sobre um gráfico é seu impacto. Impacto é o que um gráfico faz.

- Elijah Meeks

Além do impacto, uma coisa simples e poderosa que o gráfico traz é a simplicidade de se mostrar uma grande quantidade de dados, o que permite fornecer *insights*. São várias as suas utilidades... eles quebram a monotonia - ajudando a atenção e interesse do leitor -, facilitam o entendimento de aspectos complexos, em que muitos números estão sendo apresentados, etc.

No trabalho de Análise Exploratória de Dados, por exemplo, é comum a criação de visualizações pelos Cientistas de Dados dentro dos seus notebooks.

A Figura 12.1 mostra os principais conhecimentos relacionados à Visualização de dados que um Data Scientist precisa conhecer. Mais uma vez: é uma lista subjetiva, em que apresentamos aquilo que entendemos importante, mas que não se esgota. Da mesma maneira, não pode-se dizer que uma pessoa não é Cientista de Dados caso não tenha conhecimento de D3.js, por exemplo. O nosso objetivo é trazer para você os tópicos que consideramos relevantes para o seu papel como

Cientista de Dados, aquilo que vai agregar valor para o seu trabalho.

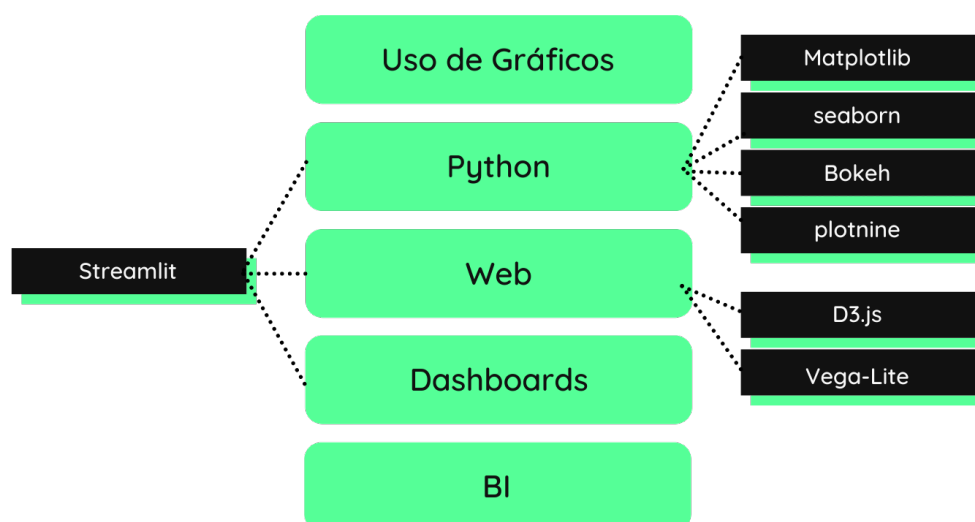


Figura 12.1: Visualização de dados

O tópico **Uso de Gráficos** é significativo, pois a escolha de um gráfico é feita não por gosto, mas por adequação ao seu objetivo. Por exemplo, o que se quer mostrar é uma comparação, uma relação, uma distribuição ou uma composição? Essa pergunta já deixa muita gente desconcertado. Às vezes, as pessoas até acertam na escolha do tipo de gráfico, mas sem saber, conscientemente, o motivo da escolha. Mas, mesmo quando isso acontece, o desconhecimento afeta outras escolhas visuais que podem facilitar ou prejudicar o objetivo do gráfico. Cada aspecto dos dados e do objetivo com o gráfico é importante para a decisão de qual visualização utilizar: Quantas variáveis a base possui? Qual a quantidade de observações? São dados categóricos ou numéricos? Os dados se alteram ao longo do tempo ou é um retrato estático da

realidade?

Com relação ao **Python**, temos diversos pacotes utilizados para a visualização de dados. O principal deles sem sombra de dúvida é o *Matplotlib*, que é uma biblioteca abrangente para a criação de visualizações estáticas, animadas e interativas em Python. Temos também o *seaborn*, que é uma biblioteca de visualização baseada no Matplotlib que fornece uma interface de alto nível para criar gráficos atraentes e informativos. O pacote *Bokeh* é uma excelente opção para visualização interativa de dados. Ao contrário do Matplotlib e seaborn, ele renderiza seus gráficos usando HTML e JavaScript. Isso o torna um excelente candidato para a elaboração de painéis e aplicativos baseados na web. Por fim, trazemos o *plotnine*, que é uma implementação de uma gramática de gráficos em Python baseada no ggplot2 (biblioteca de visualização mais famosa da linguagem R).

No tópico **Web**, temos o D3.js e o Vega-Lite como excelentes opções. O *D3.js* (ou D3) é uma biblioteca JavaScript para visualizar dados usando padrões da web, usando SVG, CSS e HTML. Já o *Vega-Lite* é uma gramática de visualização de alto nível. Ele fornece uma sintaxe JSON concisa para dar suporte à geração rápida de visualizações.

Também é importante conhecer as técnicas de criação de **Dashboards** e os conceitos e algumas ferramentas de **Business Intelligence**, como Tableau e Microsoft Power BI.

Por fim, destacamos o *Streamlit*, que é uma biblioteca Python de código aberto que facilita a criação e o compar-

tilhamento de belos aplicativos da web personalizados para Data Science e Machine Learning. Além de ser uma biblioteca excelente, o seu sucesso também reside no fato de que ele dispensa conhecimentos de desenvolvimento web para utilizá-lo. Você só precisa saber Python!

13. Machine Learning

13.1 Introdução

Machine Learning, ou Aprendizado de Máquina, é uma área da Inteligência Artificial que usa algoritmos para encontrar respostas, explorar dados e buscar soluções de negócios com técnicas de Aprendizado Supervisionado, Aprendizado Não Supervisionado e Reinforcement Learning (Aprendizado por Reforço). Usando poder computacional e por meio de pesquisa e tratamentos matemáticos, é possível encontrar resultados que apenas com a capacidade humana não seriam possíveis.

Tom Mitchell tem uma boa definição para Machine Learning:

Citação 13.1.1 *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .*

Diz-se que um programa de computador aprende com a experiência E respeitando a alguma classe de tarefas T e medida de desempenho P , se seu desempenho nas tarefas em T , conforme medido por P , melhora com a experiência E .

- Tom Mitchell

Esses programas de computador citados são os algoritmos, e esses algoritmos podem ser implementados em softwares, aplicações web ou ainda em aplicações de notebook como o Jupyter, entre outros. Eles usam uma ou várias bases de dados para entender como padrões ocorrem e o que pode ser utilizado no futuro para previsão de resultados para uma base de dados semelhante, mas com novos dados, ou ainda para entender como a base de dados se comporta e quais os padrões podem ser encontrados.

Na Figura 13.1 destacamos os conhecimentos mais importantes relacionados à Machine Learning que você precisa conhecer.

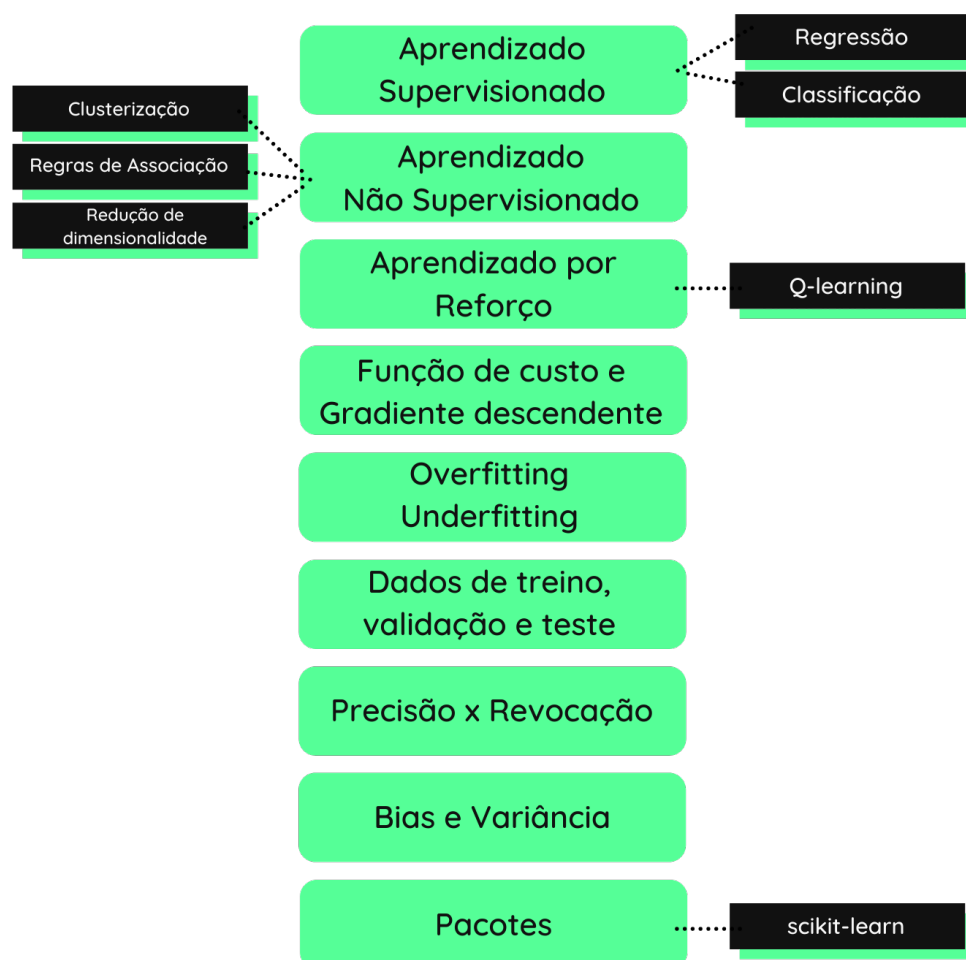


Figura 13.1: Machine Learning

Neste livro, vamos abordar mais detalhadamente sobre Aprendizado Supervisionado, Aprendizado Não Supervisionado e alguns conceitos importantes como Overfitting e Underfitting, Dados de treino, validação e teste, e Função de custo e Gradiente descendente. Vamos nessa?

13.2 Aprendizado Supervisionado

Algoritmos de Aprendizagem Supervisionada usam uma variável-alvo (também chamada de variável dependente, ou *target*) para entender como as outras variáveis da base de dados, chamadas de variáveis independentes, podem explicá-la. Em outras palavras, por meio de relações matemáticas, as variáveis que são usadas para a modelagem vão, usando operações matemáticas, criar associações com a variável que queremos prever.

Os algoritmos de análise supervisionada dividem-se em dois principais tipos: **regressão** e **classificação**.

13.2.1 Regressão

Os algoritmos de regressão são usados quando a variável-alvo (dependente) é do tipo contínua, ou seja, quando tem uma escala matemática ou não é constituída por classes (como cores, verdadeiro ou falso, etc).

É um método estatístico onde a relação final é uma fórmula matemática em que cada variável independente vai receber um peso, que é uma proporção na relação entre elas. O tipo de regressão mais utilizado é a regressão linear, mas também há algoritmos para outros tipos de regressão:

- Regressão por árvores de decisão
- Regressão por redes neurais
- Random forest

- K-Nearest Neighbours
- Support Vector Machines (SVM)

Vamos tratar rapidamente algumas das principais métricas de performance para problemas que envolvem regressão e o que elas representam.

13.2.2 Métricas de desempenho para regressão

Mean Absolute Error - MAE ou Erro Médio Absoluto

É a média do erro residual entre o que foi previsto pelo modelo e o valor real da amostra. Parece complicado, não é? Mas é um conceito simples, onde você tem a sua amostra real que vai ser usada para a criação do modelo preditivo e em cada linha da sua base tem um valor real daquela variável dependente que vai ser usada para a criação do modelo. O modelo vai prever valores para aquela variável, mas nem sempre vai acertar em cheio. Essa diferença entre o valor previsto e o valor real, a diferença matemática mesmo, vai ser somada e, no final, dividida pelo número de amostras. Aí teremos a média do erro, aqui chamado de MAE ou Erro Médio Absoluto. Quanto menor o erro, melhor, mas é sempre bom entender a base para saber para cada caso o que deve ser considerado bom ou ruim.

R quadrado

O R quadrado usa esta fórmula esquisita:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_j - \hat{y}_j)^2}{\sum_{i=1}^n (y_j - \bar{y}_j)^2}$$

Bem, nós mesmos dissemos que saber usar uma fórmula é o menos importante, lembra? Então vamos para a intuição do R quadrado: ele mostra o quanto as variáveis independentes representam a variável dependente. É a porcentagem da variação da variável dependente que é explicada por um modelo linear. Num exemplo mais prático, se o R quadrado é de 75%, isso quer dizer que o modelo explica 75% da variabilidade dos dados de resposta ao redor de sua média.

É bom estudar a fórmula para entendê-la e saber o que cada elemento da sua composição significa. Mas ter em mente o seu objetivo facilita esse entendimento.

13.2.3 Classificação

Ainda em análise supervisionada, os algoritmos de classificação, como o próprio nome já diz, referem-se a situações onde a variável-alvo é dividida em classes. O que são essas classes? São valores que não tem uma relação matemática entre si. E o que isso quer dizer? Por exemplo, num algoritmo onde o que vai ser previsto é se o cliente irá ou não pagar um

empréstimo caso o banco o conceda.

Vamos ver alguns algoritmos de classificação:

- Regressão logística
- Naïve Bayes
- Gradiente descendente estocástico
- Random forest
- K-Nearest Neighbours - KNN
- Árvores de decisão

Vamos tratar rapidamente algumas das principais métricas de performance para problemas que envolvem classificação.

13.2.4 Métricas de desempenho para classificação

Matriz de Confusão

Uma matriz de confusão (apesar desse nome engraçado) é a ferramenta usada quando usamos um algoritmo de classificação em que a variável *target* (dependente) é binária. É possível também utilizá-la para variáveis não binárias, mas sua aplicação é mais simples de entender quando aplicamos para uma *target* binária (sim ou não, falso e verdadeiro, certo e errado, etc).

A Figura 13.2 mostra uma matriz de confusão, que consiste na contagem de erros e acertos de um algoritmo de

		Predicted	
		True	False
Sample	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Figura 13.2: Matriz de Confusão

classificação:

- *True positive* (TP): foi previsto no modelo como verdadeiro e na amostra era falso - **acerto**.
- *False negative* (FN): foi previsto no modelo como negativo e na amostra era verdadeiro - **erro**.
- *False positive* (FP): foi previsto no modelo como verdadeiro e na amostra era falso - **erro**.
- *True negative* (TN): foi previsto no modelo como negativo e na amostra era negativo - **acerto**.

A relação entre esses erros e acertos e também com a quantidade geral de amostras geram métricas em porcentagens que são usadas para avaliar a qualidade dos modelos. Entre elas, estão três que são muito importantes: Acurácia, Precisão e *Recall* (Revocação).

Acurácia

A Acurácia mede a porcentagem geral de acertos do modelo, independente da classe que foi previsto, no caso de uma classificação binária pode ser sim ou não. É representada pela seguinte razão:

$$Acuracia = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisão

Precisão é a relação entre os TP e FP, como mostra a fórmula a seguir:

$$Precisao = \frac{TP}{TP + FP}$$

Isso representa tudo o que o modelo previu como positivo e a sua relação com o que realmente foi acertado, com a razão entre os acertos da classificação.

Recall (Revocação)

Outra métrica bastante importante é a *Recall*, representada da seguinte forma:

$$Recall = \frac{TP}{TP + FN}$$

Ou seja, é a relação entre o que o modelo classificou como positivo e era realmente positivo, dividido pela soma deste mesmo valor com o que era positivo na amostra e o modelo classificou como negativo. Assim, tudo o que a amostra mostrava como positivo ($TP + FN$) e sua relação com os acertos destes resultados (TP). *Recall* também pode receber o nome de *True Positive Rate*.

13.3 Aprendizado Não Supervisionado

Algoritmos de Aprendizagem Não Supervisionada usam aprendizagem de máquina para encontrar relações entre variáveis que não são de fácil percepção para o ser humano, principalmente por se tratar de bancos de dados onde há muitas variáveis e sua dimensão é muito grande, ou quando temos bases de dados que não possuem a variável dependente, apenas as variáveis independentes. Ou seja, os dados não estão rotulados. Não sabemos a resposta de nada, mas temos um conjunto de observações às quais quero entender melhor o que significam.

Para isso, os modelos de Aprendizagem Não Supervisi-

onada são capazes de trabalhar “por conta própria” para descobrir padrões e informações que não facilmente perceptíveis.

Esses algoritmos descobrem padrões ocultos ou agrupamentos de dados. Sua capacidade de descobrir semelhanças e diferenças nas informações torna a solução ideal para análise exploratória de dados, estratégias de vendas cruzadas, segmentação de clientes e reconhecimento de imagens, por exemplo.

Vamos imaginar que vamos fazer a relação entre duas variáveis de um dataset: dia da semana e número de vendas de café. É possível traçar um gráfico simples.

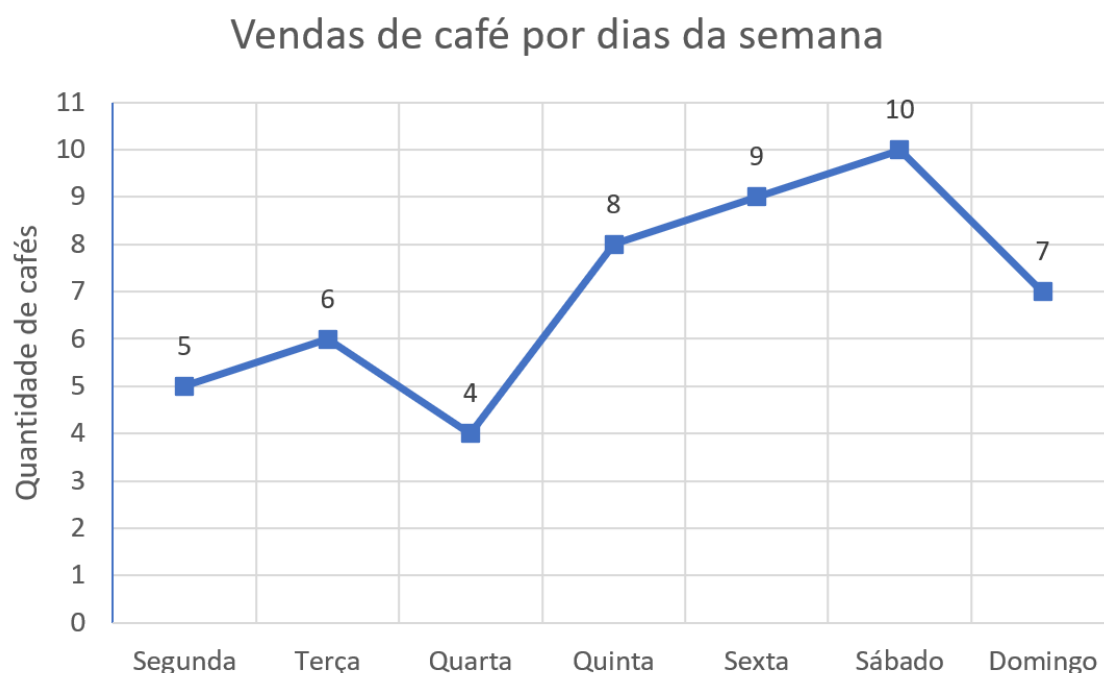


Figura 13.3: Vendas de café por dia da semana

Agora vamos imaginar uma situação diferente. Você tem mais variáveis demográficas, como salário, tamanho da casa onde vive, acesso à internet, e por aí vai. E agora ainda mais complexo, quando começamos a adicionar informações de índices de medidas corporais, como por exemplo para a obtenção de dados de riscos de doença cardíaca no futuro. É uma loucura fazermos um gráfico com tantas variáveis, certo? Há algumas formas gráficas de representar isso, mas todas elas tem as suas limitações para a inteligibilidade humana.

Nessas situações usamos Aprendizagem Não Supervisionada para encontrar padrões entre os dados. Cada linha da base de dados vai ser tratada como um *input*. Então, de acordo com as relações matemáticas entre os valores dessas linhas, e os valores de cada variável, é possível encontrar padrões entre os *inputs*.

O Aprendizado Não Supervisionado usa alguns algoritmos para tratar dessas relações e também outra aplicação muito interessante que é a redução de dimensionalidade.

Vamos ver agora alguns tipos de algoritmos e aplicações.

13.3.1 Clusterização

A principal e a mais utilizada técnica de Aprendizagem Não-Supervisionada é a Clusterização (*Clustering*). Em uma tradução direta, significa agrupar. É muito útil quando queremos direcionar campanhas para clientes que têm características semelhantes, além do sexo, lugar onde mora, etc.

É possível fazer a relação, por exemplo, com o tipo de uso do serviço, ou ainda se for uma empresa médica, quais os remédios que tomam e a dosagem, etc.

13.3.2 Detecção de anomalias

Também chamada de detecção de *outliers*, o Aprendizado Não-Supervisionado procura padrões de anomalia em uma base de dados onde não há um *target*. Exemplo disso é quando há uma base de dados onde procura-se encontrar fraude bancária, e nos registros não há registro que indique que uma transação bancária é fraudulenta ou não, o que nesse caso justifica a descrição do termo em que não há um rótulo (*label*), não é sabido se os processos foram fraudulentos.

13.3.3 Redução de dimensionalidade

O principal algoritmo usado para a Redução de dimensionalidade propósito é o PCA - *Principal Component Analysis* (Análise de Componentes Principais). Ele analisa as relações entre variáveis para encontrar, entre essas, quais as variáveis que se relacionam mais entre si dentro de uma base de dados. Essa ferramenta é muito usada na Análise Exploratória de Dados para entender como os dados se comportam na base e quais informações importantes podem ser obtidas por ele. Através de vetores e rotações vetoriais, os dados são relacionados e, de acordo com as variáveis que são agrupadas, análises mais profundas são processadas.

13.4 Conceitos importantes

Vamos agora conhecer alguns conceitos fundamentais em Machine Learning para finalizarmos este capítulo e esta parte do livro.

13.4.1 Overfitting e Underfitting

Os conceitos de **overfitting** e **underfitting** são muito importantes e devem ser muito bem compreendidos. Quando um algoritmo é posto para rodar, o que se espera é que o resultado seja o melhor possível, claro! Quando temos um algoritmo de previsão, como uma regressão linear, queremos que o que seja previsto esteja o mais próximo possível dos valores iniciais, certo? Quase isso!

A precisão de um algoritmo é comumente medida em porcentagem de acertos. Então se um algoritmo tiver 100% de acerto, não é bem isso que procuramos. Claro, quanto maior o nível de acerto, melhor, mas há um limite em que começa a acontecer o overfitting. *Overfit* significa **sobreajuste**, ou seja, o ajuste foi além do limite!

A Figura 13.4 mostra, em uma escala reduzida, como funciona o Underfitting, o Overfitting e um modelo Otimizado.

Podemos ver como, no exemplo de overfitting, a linha gerada pelo algoritmo está muito próxima dos dados. Tudo bem, isso ainda parece que deve ser o ideal de cada algoritmo, certo? Como foi falado anteriormente: sim, queremos que nosso algoritmo seja preciso e que tenha bons resultados.

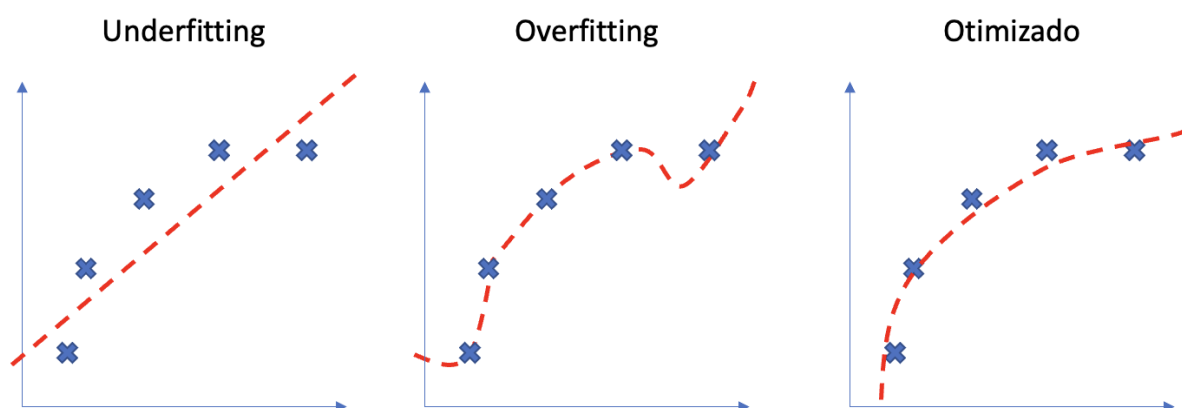


Figura 13.4: Underfitting, Overfitting e Otimizado

Mas quando acontece a modelagem, estamos fazendo um programa que vai ser reusado no futuro com uma base de dados diferentes. As colunas das tabela vão ser as mesmas, com certeza, mas cada linha da tabela vai ter valores diferentes do que os dados que foram usados para gerar o modelo.

Assim, se fizermos um algoritmo que é muito preciso na modelagem, a regra é que quando ele é posto para rodar com outra base de dados, sua precisão fica comprometida porque ele está preparado apenas para aquela base de dados de modelagem específica!

Contrariamente ao overfitting está o underfitting. *Underfit* significa **sub-ajuste**. O caso é o extremo contrário, onde o modelo não chega perto do que a própria base de dados de teste representa, fazendo com que os resultados para uma base de dados futura seja ainda pior do que no caso do overfitting.

O ideal é que se procure sempre otimizar o algoritmo

a ser implementado e que o Cientista de Dados tenha discernimento, de acordo com sua experiência, do que é um modelo otimizado e que não caia na armadilha do overfitting, sabendo quando uma otimização deixa de ser a própria otimização e passa a ser um overfitting. É preciso saber quando parar de otimizar.

13.4.2 Dados de treino, validação e teste

Quando separamos uma base de dados para modelagem, o primeiro passo, ou melhor, o passo zero - para ficar bem claro - é a divisão da base em **treino, validação e teste**.

O que isso significa? Quer dizer que quando vamos fazer uma modelagem, precisamos testar os modelos para não cairmos na armadilha que tratamos anteriormente do overfitting ou do underfitting. Para que isso seja evitado, precisamos dividir a base de dados de uma maneira que, dentro da própria base de dados, possamos realizar avaliações de performance.

Para isso, separamos a base de dados em proporções que são comuns e usadas geralmente na indústria:

- 70% treino, 15% validação, 15% teste.
- 80% treino, 10% validação, 10% teste.
- 60% treino, 20% validação, 20% teste.

Mas, algumas vezes, não temos um volume de dados grande que proporcione um *split* dos dados eficiente, quando

é o caso de base de dados com poucos registros. Para esse caso, pode ser feito de forma muito eficiente a divisão da base de dados apenas em treino e teste. Então, para validar, é usada uma técnica chamada **validação cruzada** (*cross-validation*). Para este caso, as proporções comuns de divisão são:

- 70% treino, 30% teste.
- 80% treino, 20% teste.

Em poucas linhas, o que é validação cruzada? Ela consiste em dividir a sua base de treino em partes iguais e rodar o modelo o mesmo número de vezes que a base de dados foi dividida, ou até que os resultados sejam consistentes entre as validações.

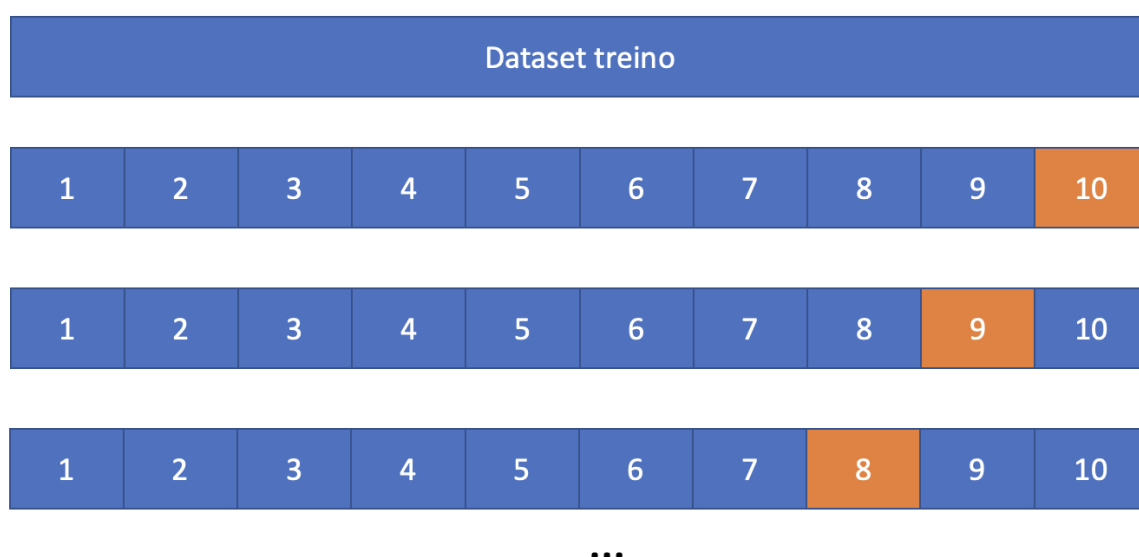


Figura 13.5: Validação cruzada

No exemplo da Figura 13.5, o dataset foi dividido em 10 partes iguais. Para a primeira rodada de validação, vão ser usados os primeiros 9 pedaços para treino e o último para validação. Na segunda rodada, os pedaços 1 a 8 e 10 para treino e o pedaço 9 para validação, e assim por diante. Os algoritmos de *cross-validation* geralmente têm um limite de iterações ajustado para que não seja necessário passar por todos os splits se assim não for preciso. Então, por exemplo, se após 4 iterações a validação for muito parecida e satisfatória, o algoritmo pode ser configurado para parar e dar como concluída a validação.

13.4.3 Função de custo e Gradiente descendente

O **Gradiente descendente** é uma técnica matemática utilizada para buscar mínimos locais em uma função desconhecida. No cálculo, o gradiente é um vetor tangente à função e aponta para o sentido e direção onde o crescimento da função é máximo.

Em aprendizagem de máquina essa função é chamada de **função de custo** ou função de erro. A função de custo geralmente é desconhecida e possui como X os coeficientes relacionados às variáveis independentes do modelo e Y o valor do erro ou custo. É com base nessa função que os algoritmos de aprendizagem de máquina ajustam seus parâmetros para chegar ao menor erro possível.

Para tal, podemos utilizar o cálculo do gradiente para que,

a partir de um valor de erro (ou custo) saibamos a direção e sentido para um desse erro vai aumentar ao máximo. Como na verdade queremos reduzir o erro basta pegarmos o sentido oposto do gradiente.

Esse cálculo é feito para cada coeficiente relacionado as variáveis independentes, para que saibamos se precisamos aumentar ou reduzir para que o erro também reduza. É quase um paralelo ao ajuste de uma persiana quebrada, a gente nunca sabe se tem que puxar ou soltar as cordas pra subir ou descer a persiana, só testando e ajustando devagar que conseguimos o resultado.

IV Mercado de trabalho para o Cientista de Dados

14 Vagas e salário 99

14.1 Brasil e Mundo

15 Outros tipos de trabalho . 105

15.1 Freelance

15.2 Consultoria

14. Vagas e salário

14.1 Brasil e Mundo

Nesta última parte do livro iremos responder a última pergunta que nos comprometemos no início: **Como é o mercado de trabalho para o Cientista de Dados?** Para começar, neste capítulo falaremos das oportunidades de trabalho no mercado e como é o salário para o Cientista de Dados no Brasil e no mundo.

Vamos começar com alguns dados interessantes, na Figura 14.1: as 50 melhores profissões para se trabalhar nos Estados Unidos em 2021, segundo o site Glassdoor.

Caso você não conheça, o Glassdoor é um dos maiores sites de vagas e recrutamento do mundo, com milhões de informações - que incluem opiniões de funcionários, informa-





50 Best Jobs in America for 2021

Best Jobs

2021

United States

Share



	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Java Developer	\$90,830	4.2/5	10,103	View Jobs
#2	Data Scientist	\$113,736	4.1/5	5,971	View Jobs
#3	Product Manager	\$121,107	3.9/5	14,515	View Jobs
#4	Enterprise Architect	\$131,361	4.0/5	10,069	View Jobs
#5	Devops Engineer	\$110,003	4.0/5	6,904	View Jobs

Figura 14.1: Glassdoor - 50 melhores profissões nos EUA (acesso em 16/07/2021)

ções salariais e relatos de entrevistas de emprego - de 900 mil empresas, além de possuir mais de 11 milhões de vagas abertas em média.

O cargo Data Scientist aparece na segunda posição dentre os 50 melhores trabalhos nos Estados Unidos em 2021. O mercado dos Estados Unidos é um bom parâmetro para avaliarmos com está o mercado mundial, já que as coisas por lá costumam nascer e acelerar antes do que aqui no Brasil, a exemplo do que mostramos no Capítulo 4 com relação às buscas do termo “Data Science” no Google.

Outro dado interessante é o salário médio do cientista de dados lá: US\$ 113.736 (dólares) no ano, o que dá em torno de US\$ 9.500 (dólares) mensais. Nada mal, não? Mais uma informação relevante: no momento em que as informações

foram capturadas, havia 5.971 vagas de emprego abertas para cientista de dados nos EUA!

Um aspecto interessante, atualmente, é que muitas empresas permitem o trabalho remoto. Assim, o Cientista de Dados pode se candidatar para vagas não só no Brasil, mas em todo o mundo!

Mas vamos ver, na Figura 14.2, quantas vagas encontramos abertas, no LinkedIn, para Cientista de Dados no Brasil?

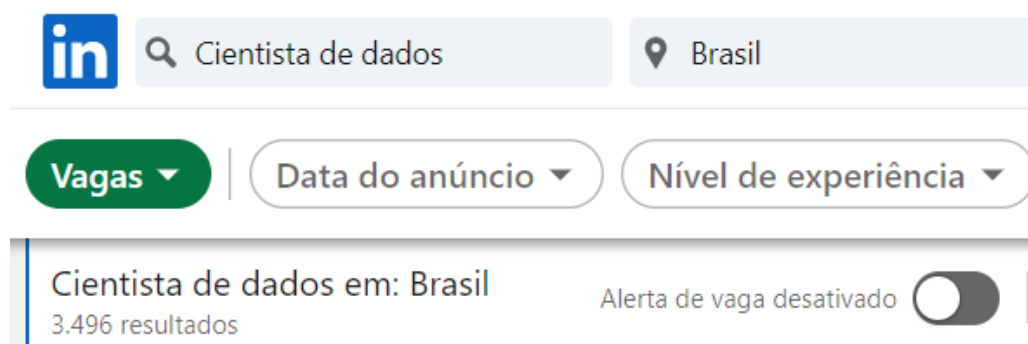


Figura 14.2: LinkedIn - Vagas para Cientista de Dados (acesso em 16/07/2021)

Eram 3.496 vagas abertas na data da nossa pesquisa (16 de julho de 2021). E isso só no LinkedIn, ok? Existem diversas plataformas de anúncio de vagas e essa quantidade não reflete toda a realidade do Brasil. Ou seja, com certeza sabemos que esse número é maior.

Tem mais uma informação que você acreditamos que você também quer saber: **qual é o salário médio de um Cientista de Dados no Brasil?** Acertamos? Então vamos ver!

A Figura 14.3 mostra essa informação pra gente e também foi retirada do site Glassdoor.



Figura 14.3: Glassdoor - Salário médio do Cientista de Dados no Brasil (acesso em 16/07/2021)

Esse é o salário médio, calculado com base em 761 salários postados na plataforma. É um bom salário! Tem Cientista de Dados ganhando menos? Sim. Tem Cientista de Dados ganhando mais? Também tem! Lembre-se que esse dado é uma média. O seu salário vai depender diretamente do seu conhecimento, habilidade e experiência! É possível ganhar bem mais, no Brasil? Totalmente! Não é a regra, mas é totalmente possível.

Vamos ver um exemplo? A Figura 14.4 mostra a informação do salário médio de um Cientista de Dados no Nubank. Mais uma vez, as informações foram retiradas do site Glassdoor.

São informações calculadas com base em 21 salários, que variam de R\$ 9.000 a R\$ 27.000, sendo R\$ 13.922 o salário médio. Veja que o menor salário para cientista de dados no Nubank é maior do que o salário médio do Brasil. Esses dados nos mostram muitas coisas... uma delas é que o Nubank

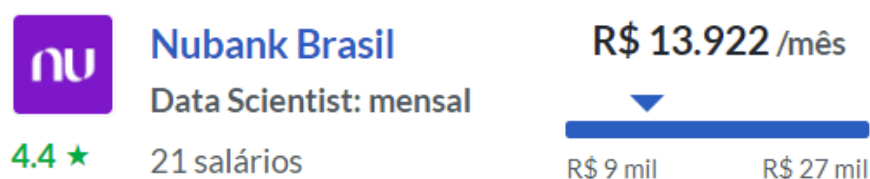


Figura 14.4: Glassdoor - Salário médio do Cientista de Dados no Nubank (acesso em 16/07/2021)

valoriza muito o Cientista de Dados. Também nos mostra que é possível ganhar R\$ 27.000 reais, caso você queira, se dedique, aprenda muito e adquira muita experiência! Na verdade, é possível ganhar até mais do que isso.

Bem, já dá pra imaginar que essa realidade de grande quantidade de vagas e bom salário não é uma característica só dos EUA e do Brasil, certo? Não vamos colocar neste livro dados de outras regiões do mundo para o capítulo não ficar muito longo. Mas, caso ainda tenha dúvida, é só pesquisar por essas informações na Europa, por exemplo, para ver que o Cientista de Dados está em alta conta na Alemanha, na Inglaterra, na França, etc.

Antes de finalizarmos esta seção, um detalhe muito importante! Na Europa, as empresas costumam usar o inglês como idioma padrão para comunicação, tanto escrita quando falada. Não é difícil de imaginar por que isso acontece. A Europa é um continente muito diverso, com muitos países e línguas diferentes. Por isso, adotar a língua inglesa facilita o intercâmbio de mão-de-obra. Ou seja, caso você queira trabalhar na Finlândia (um ótimo país, por sinal), você não

precisa aprender finlandês. Você consegue viver uma vida
perfeitamente funcional com o inglês! #ficaadica

15. Outros tipos de trabalho

Nós apresentamos nesta parte do livro, até então, dados de uma realidade do mercado de trabalho para alguém que quer se tornar empregado de uma empresa e ganhar salário. Acreditamos que essa é a realidade da maioria das pessoas! Porém, muita gente não sabe que existem outros dois caminhos muito interessantes e atrativos para se trabalhar como Cientista de Dados:

- Freelance
- Consultoria

Vamos ver um pouco sobre esse mundo também! Pode ser que seja um bom caminho para você.

15.1 Freelance

Como você já deve saber, o *freelancer* é o profissional que trabalha por conta própria (autônomo), oferecendo seus serviços profissionais, sem qualquer vínculo empregatício, e é remunerado por trabalho realizado.

O que é legal, atualmente, é que existem plataformas que fazem a intermediação entre os *freelancers*, que oferecem seu serviço, e as empresas ou profissionais que demandam trabalho. Normalmente, esse tipo de trabalho é calculado em hora trabalhada.

Quais são as vantagens de trabalhar com freelance?

- Amplia sua rede de contatos.
- Trabalha em horários flexíveis.
- Pode atuar em empresas do mundo inteiro.
- Pode ganhar em dólar ou euro.
- Adquire experiências diversas.

Mas nem tudo são flores, certo? Nos sentimos na obrigação de mostrar, também, algumas desvantagens ou, ao menos, aspectos que devem ser levados em consideração para que você possa avaliar melhor esse caminho:

- Rendimento mensal incerto.
- Não tem direitos trabalhistas.
- Excesso de trabalho, caso não saiba estimar prazos e não

seja uma pessoa organizada.

- Falta de contato pessoal.

Para ajudá-lo ainda mais, na Figura 15.1 você pode ver um dos resultados de uma publicação de 2017 do *BLS - Bureau of Labor Statistics* (Secretaria de Estatísticas Trabalhistas) dos EUA, chamada *Contingent And Alternative Employment Arrangements* (Arranjos de Trabalho Contingentes e Alternativos), que mostra que 79,1% dos consultores / freelancers independentes preferem permanecer independentes. Apenas 8,8% preferem ter um emprego tradicional. Isso é um dado interessante... não é um trabalho para todo mundo, mas quem permanece não tem vontade de mudar.

Preference	Independent contractors
Total, 16 years and over (thousands).....	10,614
Percent.....	100.0
Prefer traditional arrangement.....	8.8
Prefer alternative arrangement.....	79.1
It depends.....	7.5
Not available.....	4.5

Figura 15.1: BLS - Bureau of Labor Statistics
Contingent And Alternative Employment Arrangements, 2017

Vamos apresentar agora algumas plataformas de freelance que você pode acessar e conhecer, caso tenha interesse nesse

tipo de atuação profissional.

15.1.1 Upwork

A Upwork [↗](#) é um dos maiores *marketplaces* de trabalho do mundo, principalmente de tecnologia (é listada na NASDAQ). Apesar da grande maioria das oportunidades serem para trabalhos menores, existem empresas que procuram profissionais para contratos mais longos. A plataforma fornece uma espécie de moedas virtuais que são gastas a cada aplicação para um trabalho. Uma quantidade fixa mensal é recebida pelos freelancers, mas é possível comprar mais moedas virtuais.

FORMA DE ENTRADA (SCREENING PROCESS)

Possui um processo de entrada relativamente fácil, com poucos casos de rejeição. Essa vantagem pode se tornar uma grande desvantagem pelo fato haver uma concorrência muito grande por bons trabalhos. É um bom local para começar, mas é preciso paciência, pois conseguir um primeiro trabalho pode levar algum tempo.

PAGAMENTO

Os trabalho são quase todos vinculados ao dólar americano (USD), mas existem serviços e empregos cotados em

outras moedas. O pagamento, no entanto, é realizado em USD. A Upwork aceita retiradas por PayPal, Payoneer e Wire Transfer direto para um banco brasileiro. Cada tipo de pagamento tem uma taxa específica.

OPORTUNIDADES

Existem muitos freelancers que vivem somente de trabalhos na Upwork. Por um período, o grande Kaggle Grandmaster Mário Filho trabalhou como cientista de dados pela plataforma, além de ter trabalhado na própria Upwork.

Como já dissemos, o primeiro trabalho normalmente demora um pouco e muitos serviços pedem que o freelancer já possua um review na plataforma: um paradoxo que deve ser quebrado com muita paciência e tentativas.

Os serviços podem ser por hora trabalhada ou por empreitada. Para cada serviço o freelancer deve informar quanto quer receber por hora ou pela empreitada e enviar a oferta. Cada oferta dessas gasta as moedinhas virtuais.

Uma forma interessante de conseguir mais chances de ser contratado é caprichar na aplicação para a vaga/trabalho. Escrever bastante, dizer porque o cliente deve contratá-lo, mostrar interesse e às vezes até realizar alguns serviços “de graça” para demonstrar capacidade são algumas dicas para conseguir as primeiras oportunidades nessa plataforma.

VANTAGENS

- Muitos serviços.
- Pagamento relativamente bom (principalmente com o câmbio do dólar a R\$5,00). Uma vez vencida a barreira dos primeiros bons *reviews*, é possível ser muito bem pago pelos serviços.

DESVANTAGENS

- Conseguir o primeiro trabalho é difícil.
- Existem muitos freelancers que aceitam oportunidades por valores muito baixos, o que dificulta ainda mais conseguir um bom trabalho que pague bem no início.

LINKS ÚTEIS

- ☞ No Brasil, Mário Filho.
- ☞ Nos Estados Unidos, Josh Burns (o CARA da Upwork, sabe tudo).

15.1.2 Codementor

Bastante desconhecida no Brasil, a Codementor ☞ é uma plataforma de mentoria e freelance de tecnologia. Seu foco é exatamente como o nome sugere: unir mentores e mentora-dos.

Mais que um *marketplace* de trabalhos de freelance, mui-

tos usuários da plataforma procuram pessoas com mais experiência para ajudá-los em tarefas específicas ou mesmo no auxílio para trilhar uma carreira.

Apesar de não ser o foco, existem muitas oportunidades de freelance e vagas de longo prazo.

FORMA DE ENTRADA (SCREENING PROCESS)

O processo de entrada da Codementor é um pouco mais rígida que a Upwork, por exemplo. Não há entrevistas ou provas, mas de fato os avaliadores verificam a formação e experiência anterior no LinkedIn e o GitHub para saber se o profissional pode integrar na plataforma.

Conhecemos alguns casos de rejeição, principalmente quando não havia projetos e experiência cadastrados no LinkedIn, do contrário, não deve ser difícil passar no processo.

PAGAMENTO

Os trabalho são todos vinculados ao dólar americano (USD). A plataforma aceita retiradas por PayPal e Wise (antiga Transferwise).

OPORTUNIDADES

Uma vez dentro da plataforma, conseguir oportunidades na Codementor é muito mais fácil que na Upwork. Talvez

pelo fato de haver uma barreira de entrada, mesmo que não muito rígida. A grande maioria dos clientes que procuram serviços são estudantes, ou profissionais e startups que não tem pessoas de tecnologia no quadro. Existem muitas oportunidades para ciência de dados, principalmente de manipulação e análise de dados. PowerBI, Tableau e outras ferramentas de análise de dados possuem bastante necessidade de freelance e mentorias.

Com um pouco de trabalho e insistência, principalmente na abordagem dos clientes logo após o cadastro da oportunidade, é possível realizar os primeiros trabalhos já na primeira semana. Existe um cadastro do valor em dólares por hora que o freelancer/mentor deseja receber, mas isso pode ser alterado em qualquer momento.

Os serviços podem ser por realizados de muitas formas: mentoria, freelance, pagamento direto (na plataforma) e mais algumas outras formas. Uma ferramenta bem interessante, para cálculo de horas de mentoria, é a integração com o Zoom com cobrança automática. Ficou 15 minutos com o mentorado e acabou a chamada: é debitado automático do cliente e transferido para sua conta da Codementor.

Como mencionado, é importante estar “online” por algumas horas para aproveitar as oportunidades recém cadastradas.

VANTAGENS

- Uma vez dentro, fácil de conseguir trabalho.
- Paga relativamente bem (média de 50 USD por hora).
- Plataforma de mentoria é bem feita e fácil de utilizar.
- A cobrança automática integrada com o Zoom é uma ideia muito boa e funciona bem.

DESVANTAGENS

- Quem não tem formação ou experiência em algo de tecnologia tem dificuldades em entrar.
- O número de trabalhos disponíveis é drasticamente inferior à Upwork, por exemplo.

LINKS ÚTEIS

☞ No Brasil, sempre ele, Mário Filho.

15.1.3 Toptal

A Toptal ☞ é o Santo Graal para qualquer cientista de dados. Startup que percebeu o gap entre a necessidade de bons profissionais no vale do silício com a abundância de *meganerds* de países com câmbio desfavorável. Uma empresa americana paga facilmente 150 a 200K USD por ano por um bom profissional na Califórnia, por exemplo, enquanto existem brasileiros e *hermanos* bons de serviço que aceitariam facilmente por 70 a 80K USD ano.

Mais que uma plataforma de freelance, a Toptal tem parcerias com grandes empresas que oferecem contratos mais duradouros com um ótimo salário (*what's not to like?*). Uma vez dentro, se abre um oásis de boas oportunidades (não é exagero).

FORMA DE ENTRADA (SCREENING PROCESS)

O nome diz muito sobre a empresa também. Eles não querem profissionais nem medianos: querem os bons, os melhores. Eles se vangloriam de ter somente 3% dos processos seletivos bem sucedidos, ou seja, 97% das candidaturas são rejeitadas. Isso parece inalcançável, mas com dedicação é possível integrar o Santo Graal.

As seleções são separadas por especialidade. No nosso mundinho, eles tem a seleção de cientista de dados.

O processo seletivo da Toptal possui quatro etapas:

1. Entrevista em inglês: somente para medir nível de comunicação.
2. Prova com 3 questões de Ciência de Dados e Estatística (plataforma Codility) [↗](#).
3. Prova com 2 questões de Ciência de Dados e Estatística com um entrevistador te acompanhando a distância.
4. Um trabalho prático de Ciência de Dados.

Ufa... passado nas quatro etapas, o selecionado entra na plataforma e passa a ter acesso a oportunidades em empre-

sas do mundo todo... algumas bem famosas como Airbnb, Bridgestone, Coinbase, Duolingo, Shopify.

Para se preparar para o processo é necessário treinar muito seu inglês, além de dominar exercícios de ciência de dados e estatística. É possível treinar na plataforma utilizada pela própria Toptal: Codility [↗](#).

PAGAMENTO

Os trabalho são todos vinculados ao dólar americano (USD). A plataforma aceita retiradas por PayPal e Wire Transfer direto para a instituição financeira.

OPORTUNIDADES

O céu é o limite. Muitas vagas *full* e *part time*. Salários muito competitivos em dólar. É difícil achar os valores exatos porque os profissionais assinam um termo muito rígido de não informar dados sobre a plataforma. Um simulador da própria Toptal diz que a média gira em torno de 35 a 50 dólares a hora (no câmbio de hoje, 04/08/2021, na madrugada fria de agosto: R\$ 180 a 260 a hora). Nada mal, não é?

Diz a lenda que, com a pandemia, o número de empresas que procuraram a Toptal cresceu vertiginosamente e, pelo jeito, essa demanda veio para ficar.

VANTAGENS

- Vagas mais perenes.
- Possibilidade de conseguir empregos *part* e *full time*.
- Chances reais de receber salários de 7.000 dólares por mês.
- Uma vitrine enorme de oportunidades: dizem que o LinkedIn de quem é selecionado na Toptal não para de apitar de tantas vagas oferecidas.

DESVANTAGENS

- O processo de entrada é muito pesado, mas é possível se preparar.
- Pouco conhecimento no Brasil.

LINKS ÚTEIS

No Brasil:

☞ Túlio (DevPleno).

☞ André Hil.

No Mundo, Carlos Roso ☞ (é o cara que sabe tudo da Toptal).

15.2 Consultoria

Há, ainda, a possibilidade de se trabalhar como consultor em ciência de dados. Porém, para esse tipo de função exige-se mais experiência, já que um consultor é um profissional que fornece conselhos profissionais ou especializados.

O ideal, para quem vai ingressar na área de ciência de dados, é começar como empregado ou freelancer. Mais tarde, caso queria atuar como consultor, há também essa possibilidade, inclusive como empreendedor!

Não citaremos nenhuma empresa específica, mas faça essa pesquisa. Digite “consultoria em ciência de dados”, ou “data science consulting” e veja a quantidade de empresas e profissionais que estão atuando dessa forma. Assim como a demanda por cientistas de dados como empregados vai crescer, a demanda por freelancer e consultores também irá.



Conclusão

16	Conclusão	119
16.1	Nossos canais	
	Referências	121
	Livros	
	Artigos	
	Sites	
	Índice remissivo	123

16. Conclusão

Esperamos que a gente tenha conseguido cumprir nosso propósito de responder as três perguntas que definimos na Introdução deste livro:

1. O que é Ciência de Dados?
2. O que aprender para se tornar um Cientista de Dados?
3. Como é o mercado de trabalho para o Cientista de Dados?

Conheça a **Jornada Cientista de Dados**, uma formação completa - curso, projetos, mentoria e comunidade - criada pelo Let's Data para você se tornar um cientista de dados!

Caso queira entrar em contato conosco, envie um email

para *contato@letsdata.ai* que responderemos.

16.1 Nossos canais

Acesse nosso site e cadastre-se para receber avisos e notícias do mundo da Ciência de Dados: <https://letsdata.ai> ↗

Acompanhe a gente nas redes sociais para receber conteúdos gratuitos e de qualidade!

- YouTube ↗
- Instagram ↗
- LinkedIn ↗
- Twitter ↗

Inscreva-se também no Let's Data Podcast. Realizamos episódios ao vivo periodicamente pelo YouTube sobre Ciência de Dados, Inteligência Artificial e Machine Learning e disponibilizamos nos principais players de podcast:

- Spotify ↗
- Google Podcast ↗
- iTunes ↗
- Castbox ↗
- Deezer ↗

Referências

Livros

Análise preditiva: o poder de prever quem vai clicar, comprar, mentir ou morrer / Eric Siegel ; tradução de Wendy Campos. - Rio de Janeiro, RJ: Alta Books, 2017.

Data Science para negócios / Foster Provost, Tom Fawcett. - Rio de Janeiro, RJ: Alta Books, 2016.

Data Science from Scratch / Joel Grus. - Sebastopol, CA: O'Reilly, 2019.

Artigos

Data Scientist: The Sexiest Job of the 21st Century [↗](#)
/ Thomas H. Davenport and D.J. Patil - Harvard Business

Review, 2012.

Contingent and Alternative Employment Arrangements

↗ / Bureau of Labor Statistics. U.S Department of Labour, 2017.

The Role of Academia in Data Science Education ↗ /

Rafael A. Irizarry - Harvard Data Science Review - MIT, 2020.

Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards ↗ / by Usama Fayyad and Hamit Hamutcu - Harvard Data Science Review - MIT, 2020.

Sites

- ↗ Risk Prediction in Surgery.
- ↗ The Decision Lab - Biases.
- ↗ What is Data Science? - Metthey Brett.
- ↗ The Data Science Venn Diagram - Drew Conway.
- ↗ CRISP-DM 1.0 - Step-by-step data mining guide - SPSS)
- ↗ What is Machine Learning and types of Machine Learning - Part-1 - Chinmay Das.
- ↗ 8 Popular Regression Algorithms In Machine Learning Of 2021 - Ajay Ohri.
- ↗ 7 Types of Classification Algorithms - Rohit Garg.
- ↗ What is Machine Learning? A Definition - Expert.ai
- ↗ Best Use of Train/Val/Test Splits, with Tips for Medical Data - Rachel Draelos.
- ↗ Pandas Official Site.

- ☞ NumPy Official Site.
- ☞ scikit-learn Official Site.
- ☞ Matplotlib Official Site.
- ☞ Beautiful Soup Official Site.
- ☞ ETL (Extract, Transform, Load) - IBM.
- ☞ ETL - O que é e qual sua importância? - SAS.
- ☞ Git Official Site.
- ☞ GitHub Official Site.
- ☞ Oracle | What is NoSQL.
- ☞ MongoDB | NoSQL Explained.
- ☞ Conceitos de Data Warehouse - AWS.
- ☞ IEEE Spectrum - Top Programming Languages.
- ☞ Virtual Environments and Packages - Python.
- ☞ What Charts Do - Elijah Meeks.

Índice Remissivo

A

Ambientes Virtuais	63
Análise Exploratória de Dados .	65
Aplicação da lei	36
Aprendizado Não Supervisionado	
88	
Clusterização.....	90
Detecção de anomalias	91
Redução de dimensionalidade	
91	
Aprendizado Supervisionado...	82
Classificação	84
Métricas de desempenho para	
classificação.....	85
Métricas de desempenho para	
regressão	83
Regressão	82

C

Conceitos importantes	92
Dados de treino, validação e teste	
94	
Função de custo e Gradiente des-	
cendente.....	96
Overfitting e Underfitting...	92
Consultoria	117
CRISP-DM.....	55

D

Definições.....	22
Detecção de fraude, Segurança e	
Eficiência logística	37

E

Estatística	70
-------------------	----

ETL 52

F

Família e Vida pessoal 34

Formatos de dados 50

Freelance 106

 Codementor 110

 Toptal 113

 Upwork 108

G

GIT / Github 53

Governo, Política e Educação .. 38

H

Histórico e contexto 16

J

Jupyter Notebook 62

L

Linguagem, Pensamento e Psicologia 39

M

Machine Learning 79

Marketing, Propaganda e Internet
35

Matrizes e Fundamentos de Álgebra Linear 54

N

Noções Básicas de Bancos de Dados
48

O

Outros tipos de trabalho 105

P

Pacotes Python 61

 NumPy 62

 Pandas 61

Programação 58

Python 59

R

Recursos Humanos 40

Risco financeiro e Seguros 35

S

Saúde 36

V

Vagas e salário 99

Visualização de Dados 74

Como se tornar um Cientista de Dados?

Escrevemos este livro para responder três perguntas relevantes:

- 1) O que é Ciência de Dados?**
- 2) O que aprender para se tornar um Cientista de Dados?**
- 3) Como é o mercado de trabalho para o Cientista de Dados?**

Cada uma dessas perguntas é uma parte deste livro. Escrevemos as respostas de uma maneira simples, para que você possa ler e aprender sem se sentir sobrecarregado. Você verá que essa foi a principal motivação não só deste livro existir, mas da criação do Let's Data.