

Aula 1 - Introdução a ciência de dados com R

Profa. Yana Borges

Fevereiro de 2022

1 Introdução

1.1 O que é Ciência de Dados?

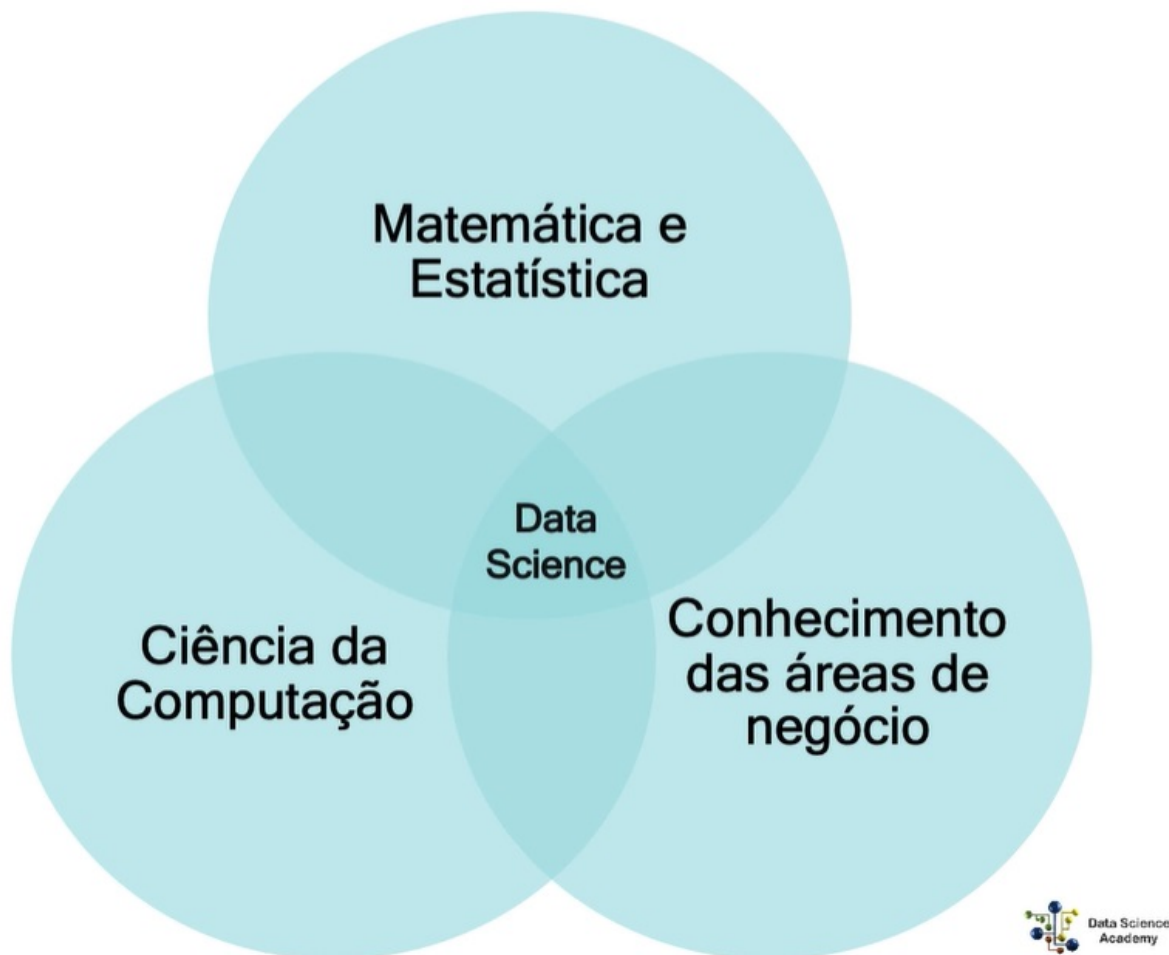
Trata-se de um termo cada vez mais utilizado para designar uma área de conhecimento voltada para o estudo e a análise de dados, onde busca-se extrair conhecimento e criar novas informações. É uma atividade interdisciplinar, que concilia principalmente duas grandes áreas: Ciência da Computação e Estatística. A Ciência de Dados vem sendo aplicada com o apoio em diferentes outras áreas de conhecimento, tais como: Medicina, Biologia, Economia, Comunicação, Ciências Políticas etc. Apesar de não ser uma área nova, o tema vem se popularizando cada vez mais, graças à explosão na produção de dados e crescente dependência dos dados para a tomada de decisão (Oliveira et al., 2018).

A Ciência de Dados é uma grande disciplina e consiste em conjunto de habilidades especializadas, tais como: estatística, matemática, programação, computação e conhecimento de negócios, além de técnicas e teorias, como análise preditiva, modelagem, mineração de dados e visualização de dados (Data Science Academy).

1.2 Áreas de conhecimento

A Ciência de Dados envolve o uso de métodos automatizados (**Ciência da Computação**) para analisar (**Matemática e Estatística**) enormes quantidades de dados a fim de extrair conhecimento (**áreas de negócio**) a partir dos dados.

- A **Matemática e Estatística** fornecem técnicas e procedimentos de cálculo, análise, correlação, transformação, limpeza e interpretação dos dados.
- A **Ciência da Computação** fornece as ferramentas usadas no processo de análise, tal como programação de computadores, armazenamento e processamento paralelo e otimização do tempo de execução dos processos de análise.
- O objetivo da Ciência de Dados é resolver problemas de negócio e domínio (**área de negócio**) é fundamental para saber interpretar os dados e os resultados do processo de análise.



1.3 Aplicações e uso de Ciência de Dados

Usando Data Science, as empresas tornaram-se inteligentes o suficiente para vender seus produtos através da análise do interesse e da capacidade de compra individual de seus clientes.

1.3.1 Busca na Internet

Quando falamos de busca, pensamos 'Google'. Certo? Mas há outros motores de busca como Yahoo, Bing, Ask, AOL, DuckDuckGo, etc. Todos estes motores de busca (incluindo o Google) fazem uso de algoritmos de Data Science para entregar o melhor resultado para a nossa consulta numa mera fração de segundos. Considerando que o Google processa mais de 20 petabytes de dados todos os dias, não fosse por Data Science, o Google não seria o 'Google' que conhecemos hoje.

Google

data science

Todas Imagens Notícias Vídeos Livros Mais Ferramentas

Aproximadamente 2.750.000.000 resultados (0,88 segundos)

Anúncio · <https://www.mbauspesalq.com/> ▾
Curso Data Science USP - Data Science EAD
 Curso Ciência de Dados USP. Professores USP. Conteúdo Atualizado. Excelência em Ensino. Aprenda Analytics, Machine Learning, Big **Data**, **Data** Mining, IoT, Deep Learning e Mais. Invista na sua Carreira. Última Chance 2021. Aulas Ao Vivo. Networking Garantido.

Conheça a Programação
 Temas Atuais e Relevantes
 Saiba Mais

Como Funciona
 Início das Aulas, Carga Horária
 Duração Total e Mais

Conheça os Professores
 Professores USP
 Saiba Mais

Depoimentos
 Quem fez, aprovou
 Depoimento de Alunos

Anúncio · <https://www.oimasterdados.com.br/> ▾
Formação em Data Science - Programa de capacitação
 Capacite-se para o mercado de **Data Analytics** e tenha oportunidade de contratação na Oi. Aprenda **Data Science** e **Data** Intelligence com especialistas do mercado. Comece agora. Chance de Contratação. Aprenda Big **Data**. Programa Inédito. Aprenda **Data Science**.

Anúncio · <https://www.udemy.com/curso-online/aprenda-hoje/> ▾
Data Science e Python - Comece a aprender hoje
 Aprenda as mais Importantes Técnicas de **Data Science** e se torne um Cientista de Dados!

<https://www.datascienceacademy.com.br/> ▾
Data Science Academy
 A Data Science Academy é um portal de ensino a distância, focado em capacitação profissional em Data Science, Big Data, Inteligência Artificial, Internet das ...

Cursos Gratuitos
 Inicie hoje mesmo! Nossos cursos gratuitos são uma pequena ...

Formação Cientista de Dados
 Data de Início ... No quarto curso, o aluno já tem uma boa visão de ...

[Mais resultados de datascienceacademy.com.br](https://www.datascienceacademy.com.br/) »

As pessoas também perguntam

Como funciona o data science? ▾


O que significa data scientist? ▾

Quanto ganha um data scientist? ▾

O que faz um profissional de data science? ▾

Feedback

<https://www.alura.com.br/> ▾ Data Science ▾
Cursos da Escola Data Science - Alura
 Cursos da Escola Data Science, Data Science. Análise exploratória com Python, Pandas, séries temporais e R. Ciência de Dados de verdade para minerar ...

Ciência de dados 
 Área de estudo

Ciência de dados é uma área interdisciplinar voltada para o estudo e a análise de dados econômicos, financeiros e sociais, estruturados e não-estruturados, que visa a extração de conhecimento, detecção de padrões e/ou obtenção de insights para possíveis tomadas de decisão. Wikipédia


Cursos ▾


Livros ▾


Estatística ▾


Objetivos ▾

Itens também pesquisados Ver mais 10

 **Análise de dados**

 **Inteligência artificial**

 **Mineração de dados**

 **Aprendiz... profunda**

Feedback

1.3.2 Propaganda digital (publicidade segmentada e ‘re-targeting’)

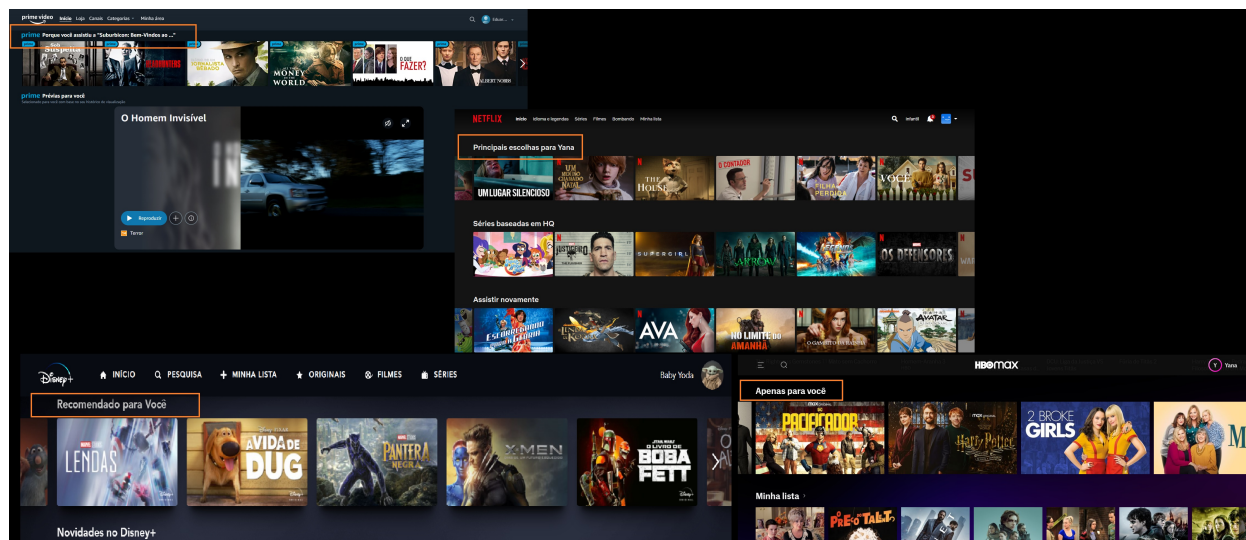
Se você pensou que a Busca seria a maior aplicação de Data Science e de Machine Learning, aqui vai um desafio – todo o espectro de marketing digital. Desde os banners exibidos nos websites até as telas digitais dos aeroportos – quase todo seu conteúdo é decidido por meio de algoritmos de Data Science.

Esta é a razão pela qual os anúncios digitais têm sido capazes de obter um CTR (‘click-through rate’) maior do que o dos anúncios tradicionais. Os anúncios digitais podem ser montados com base no comportamento passado do usuário. Esta é a razão pela qual eu vejo anúncios de treinamento em ‘analytics’ enquanto que, no mesmo lugar e ao mesmo tempo, meu amigo vê anúncios de roupas.

1.3.3 Sistemas de Recomendação

Quem poderia se esquecer das sugestões sobre produtos similares da Amazon? Elas não só nos ajudam a encontrar produtos relevantes dentre os bilhões de produtos

Muitas empresas têm passado a usar este motor/sistema para promover seus produtos e sugestões de acordo com o interesse do usuário e relevância da informação. Gigantes da Internet como a Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDB e muitas mais utilizam este sistema para melhorar a experiência do usuário. As recomendações são feitas com base em resultados de pesquisas anteriores dos usuários.



1.3.4 Reconhecimento de imagens e fala

Você sobe sua imagem e a dos amigos no Facebook e começa a receber sugestões para ‘taggear’ seus amigos. Esta função automática de sugestão de tag usa algoritmo de reconhecimento facial. Da mesma forma, ao usar whatsapp web, você escaneia um código de barras em seu navegador usando seu telefone celular. Além disso, o Google lhe dá a opção de buscar imagens que você submete. O Google usa reconhecimento de imagens e fornece resultados de pesquisa relacionados.

Alguns dos melhores exemplos de produtos de reconhecimento de voz são Google Voice, Siri, Cortana, etc. Caso você não esteja em condição de digitar uma mensagem, o recurso de reconhecimento de voz não deixa sua vida parar. Basta falar a mensagem e ela será convertida em texto.

1.3.5 Mais

Usando Data Science, os departamentos de marketing das empresas decidem quais produtos são os melhores para up selling (“melhoramento”) e cross-selling (venda cruzada) com base nos dados de comportamento de clientes. Além disso, podem prever sua participação na carteira do cliente, quais são propensos a deixarem de ser clientes, para quais clientes devem ser ofertados produtos de maior valor e muitas outras perguntas que podem ser facilmente respondidas por Data Science. Finanças (Risco de Crédito, Fraude), Recursos Humanos (quais funcionários são mais propensos a pedir demissão, o desempenho dos funcionários, decidir bônus) e muitas outras tarefas são facilmente conseguidas usando Data Science nestas disciplinas.

Jogos, websites de comparação de preços, planejamento de rotas aéreas, detecção de fraude e risco, logística de entrega, quase tudo neste planeta, que gera dados, cai sob o radar de Data Science. Veja mais no site Vooo – Insights.

1.4 Fluxo de trabalho da Ciência de Dados

Não existe apenas uma forma de estruturar e aplicar os conhecimentos da Ciência de Dados. A forma de aplicação varia bastante conforme a necessidade do projeto ou do objetivo que se busca alcançar. Neste curso, usaremos um modelo de workflow bastante utilizado, apresentado no livro R for Data Science (Hadley Wickham, 2017).

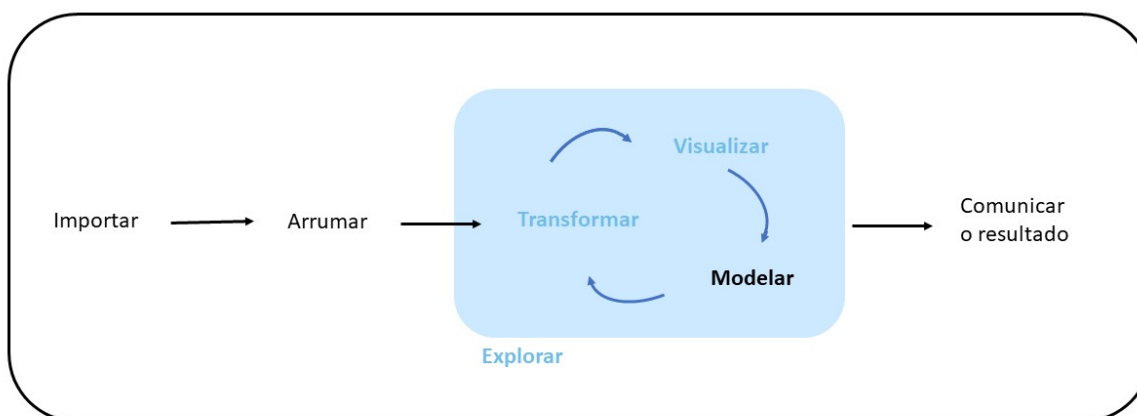


Figura 1: fluxo de trabalho básico para ciência de dados

2 Linguagens para ciência de dados

Para a aplicação dessas atividades comuns da Ciência de Dados, é necessário dominar as ferramentas corretas. Existem diversas linguagens/ferramentas: R, Python, SAS, SQL, Matlab, Stata, Aplicações de BI etc.

Cabe ao cientista de dados avaliar qual é a ferramenta mais adequada para alcançar seus objetivos.

2.1 O que é R e por que devo aprender

R é uma linguagem de programação estatística que vem passando por diversas evoluções e se tornando cada vez mais uma linguagem de amplos objetivos. Podemos entender o R também como um conjunto de pacotes e ferramentas estatísticas, munido de funções que facilitam sua utilização, desde a criação de simples rotinas até análises de dados complexas, com visualizações bem acabadas.

Seguem alguns motivos para aprender R:

- É completamente gratuito e de livre distribuição;
- Curva de aprendizado bastante amigável, sendo muito fácil de aprender;
- Enorme quantidade de tutoriais e ajuda disponíveis gratuitamente na internet;
- É excelente para criar rotinas e sistematizar tarefas repetitivas;
- Amplamente utilizado pela comunidade acadêmica e pelo mercado;
- Quantidade enorme de pacotes, para diversos tipos de necessidades;
- Ótima ferramenta para criar relatórios e gráficos.

Apenas para exemplificar sua versatilidade, este material foi todo feito em R.

2.2 R

Para fazer o download do R, vá para CRAN, a rede de distribuição do R (em inglês - *comprehensive R archive network*). O CRAN é composto por um conjunto de servidores-espelho distribuídos ao redor do mundo e é usado para distribuir R e os pacotes de R. Não tente escolher um servidor que esteja perto de você: em vez disso, use o espelho em nuvem, (<https://cloud.r-project.org> - conteúdo em inglês), que o descobre automaticamente para você.

Uma nova grande versão do R é lançada uma vez por ano, e há 2-3 versões secundárias a cada ano. É uma boa ideia atualizar regularmente. A atualização pode ser um pouco trabalhosa, especialmente para versões maiores, que exigem que você reinstale todos os seus pacotes, mas adiá-la só piora a situação.

2.3 RStudio

O RStudio é um ambiente de desenvolvimento integrado, ou IDE, para programação R. Faça o download e instale-o em <http://www.rstudio.com/download>. O RStudio é atualizado algumas vezes por ano. Quando uma nova versão estiver disponível, o RStudio informará você. É uma boa ideia atualizar regularmente para que você possa aproveitar os melhores e mais recentes recursos.

Para instalação, visite o site <https://didatica.tech/como-instalar-a-linguagem-r-e-o-rstudio/>