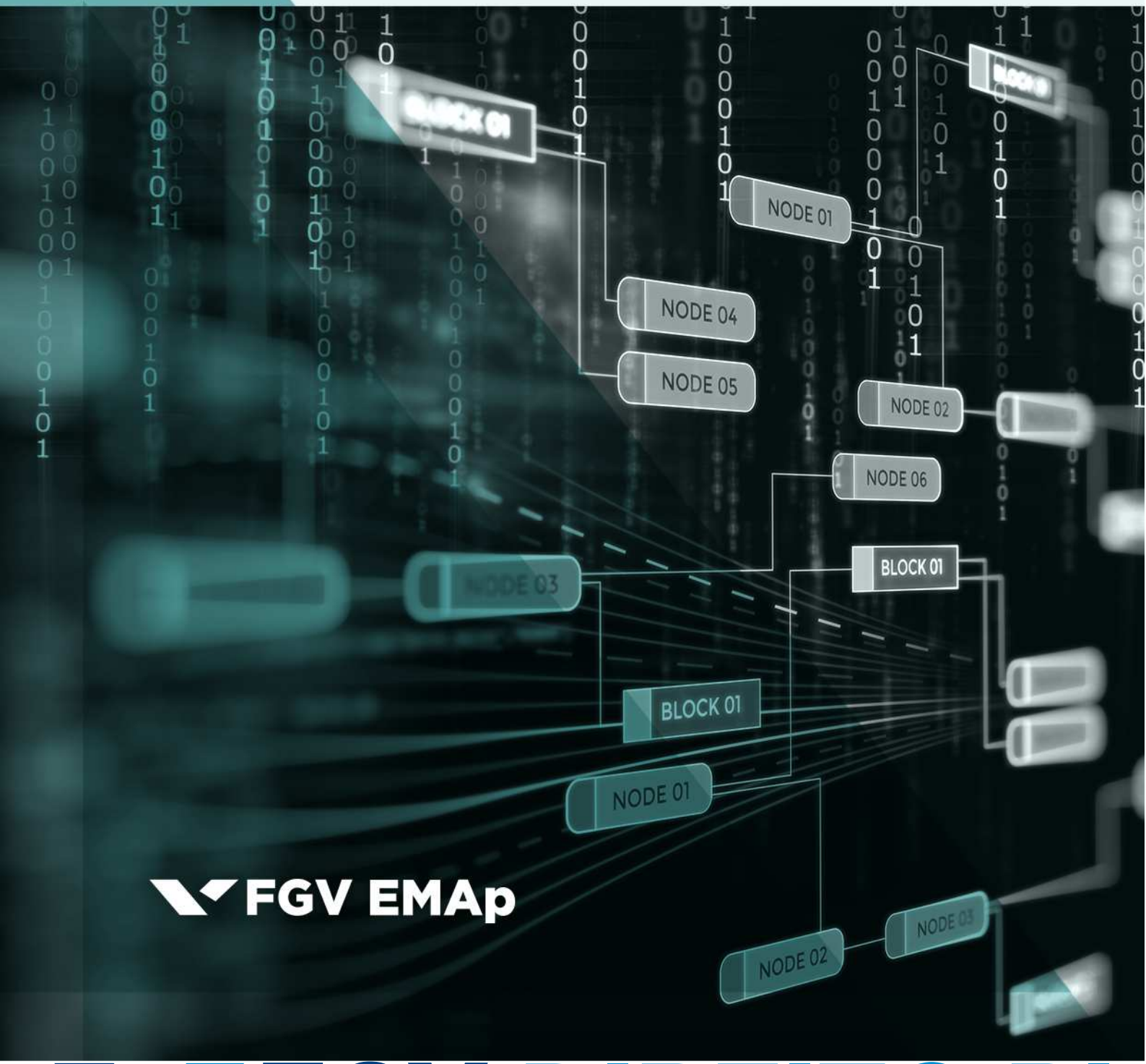
INTRODUÇÃO À CIÊNCIA DE DADOS

Jorge Poco



FGV EMAp

INTRODUÇÃO

Ciência de dados é um ramo multidisciplinar da ciência que envolve técnicas de computação, inteligência artificial, estatística e otimização, com o intuito de resolver problemas analiticamente complexos, tendo conjunto de dados como núcleo de operação. O processo de descoberta de conhecimento a partir de dados compreende um conjunto de etapas que envolvem desde a organização e limpeza da informação até a validação e o uso da informação obtida. Um cientista de dados precisa ter conhecimento das diversas modalidades de técnicas que podem ser empregadas em cada uma das etapas do processo de Ciência de Dados, sendo ainda capaz de validar resultados e respeitar aspectos éticos.

O objetivo geral deste curso é fornecer uma visão geral do que é Ciência de Dados, as suas principais etapas e o ferramental matemático e computacional tipicamente empregado em cada uma das etapas. Por sua vez, os objetivos específicos são:

- Apresentar as técnicas matemáticas e computacionais de forma simples e intuitiva, permitindo ao estudante compreender como elas funcionam, quando e em que tipo de problemas devem ser empregadas e quais são as suas limitações.
- Apresentar exemplos de aplicações reais do processo de extração e geração de conhecimento via Ciência de Dados.
- Discutir os aspectos éticos envolvidos em Ciência de Dados.

Apresentação geral da estrutura

O presente material didático é um conjunto de bibliografias recomendadas. A área de Ciência de Dados é bastante recente e abrange uma gama grande de técnicas e conceitos. O conteúdo que deve ser ministrado em um curso de Ciência de Dados é ainda uma questão em discussão, sendo que pesquisadores e profissionais da área divergem sobre quais são os conceitos realmente essenciais e como tais conceitos devem ser abordados e organizados. Além disso, Ciência de Dados tem sido alvo de interesse de estudantes e profissionais com formação muito variada, o que torna a construção de material didático sobre esse tema um desafio, pois tal material deve ser capaz de capacitar e educar tanto

estudantes de Matemática e Computação quanto profissionais das áreas de Direito e Finanças. A falta de uma metodologia educacional consolidada nesse cenário nos motivou a optar pelas bibliografias recomendadas em vez da construção de uma apostila, a qual certamente estaria ultrapassada em pouquíssimo tempo.

Dessa forma, cada unidade do curso sugere um conjunto de leituras, além de *links* para material *on-line* relacionado ao tema da unidade. As leituras principais detalham o que foi apresentado e discutido nas videoaulas.

Pré-requisito

Os textos indicados neste material didático, bem como em outros pontos da disciplina, são em inglês, uma vez que há pouco conteúdo em língua portuguesa que atenda ao crivo de seleção de informações consideradas pelos autores desta disciplina como basilares para a introdução à Ciência de Dados. Sem dúvida, há na *web* um sem-número de textos nos seus diferentes contextos e aplicações, no entanto, este curso presta-se ao serviço de contemplar os conteúdos mais atuais e mais robustos voltados à Ciência de Dados, e a exploração desses materiais, em língua inglesa, não é só necessária como também garantia de se buscar na fonte, sem traduções intermediárias.

SUMÁRIO

MÓDULO I – O QUE É CIÊNCIA DE DADOS?

Este é o nosso primeiro passo para conhecer a Ciência de Dados. Aqui, vamos definir vários conceitos que são necessários para entender os próximos módulos de estudo, seja por meio deste conjunto de bibliografias recomendadas ou pelo acesso às *webaulas* desta disciplina. Mostraremos alguns exemplos que motivam o porquê de a Ciência de Dados ser uma área de grande importância nos últimos anos. Além disso, aprenderemos sobre as habilidades que precisamos desenvolver e reforçar para nos considerarmos bons cientistas de dados.

A revolução dos dados

Nesta unidade, aprenderemos diferentes conceitos relacionados à Ciência de Dados. Um dos mais importantes é *Big Data*, e como esse fenômeno começou a mudar a maneira como os problemas são resolvidos hoje.

A partir de agora, indicaremos, por meio deste material, as publicações que abordam o conceito de *Big Data* e que estão já estão mundialmente consagradas, pois o nosso objetivo é que você encontre, na fonte, a compreensão necessária para dar sequência aos seus estudos sobre Ciência de Dados.

A general introduction to data analytics (MOREIRA *et al.,* 2018) Seções 1.1, 1.2 e 1.3

Estas seções descrevem o conceito de *Big Data*, que é inicialmente definido por três "Vs": volume, variedade e velocidade. Da mesma forma, é feita uma relação entre *Big Data* e *Data Science*. Quando um conjunto de dados é considerado *Big Data*?

The government-academia complex and big data religion

(Disponível em: <https://www.forbes.com/sites/gilpress/2014/09/09/the-government-academia-complex-and-big-data-religion/#254394262a10>)

Este artigo descreve ideias dos momentos em que o *Big Data* pode ser bom e quando não pode. Há uma ênfase no que é a religião de *Big Data*, na qual há seguidores apaixonados que acreditam pertencer a um novo movimento científico que não precisa fazer perguntas, mas apenas coletar dados e deixá-los falar por eles mesmos.

The data science handbook (CADY, 2017) Seção 13.1

Esta seção toca mais uma vez a conexão entre *Big Data* e *Data Science*, e o cuidado que deve ser tomado ao usar ferramentas de *Big Data*, pois elas ainda estão em processo de evolução.

Tomada de decisão baseada em dados

Nesta unidade, veremos como a Ciência de Dados apoia a tomada de decisões baseada em dados e, às vezes, permite tomar decisões automaticamente em grande escala.

Data science for business (PROVOST; FAWCETT, 2013): Capítulo 1, Seção *Data science, engineering, and data-driven decision making (p. 3-7)*

Esta seção descreve a relação entre Ciência de Dados e tomada de decisão com base em dados e argumenta que o objetivo final da Ciência de Dados é melhorar a tomada de decisão, uma vez que esta é geralmente a principal meta das empresas.

Conhecimentos e habilidades de um cientista de dados

Nesta unidade, veremos duas perspectivas de quais devem ser as habilidades de um cientista de dados. Essas duas perspectivas não são mutuamente exclusivas, pelo contrário, elas se complementam.

Principles of data science (OZDEMIR, 2016) Capítulo 1, Seções *What is data science? e The data science Venn diagram*

A primeira seção é uma breve introdução do que é Ciência de Dados e de algumas terminologias que serão usadas. A segunda seção apresenta as habilidades que os cientistas de dados devem. Em resumo, o diagrama descrito no texto nos mostra que a Ciência de Dados é composta de três áreas importantes: matemática/estatística, computação e conhecimento do domínio.

An introduction to data science (SALTZ; STANTON, 2018) Capítulo 0 *Introduction data science, many skills*

Este capítulo nos fornece um breve resumo do que é Ciência de Dados e quais habilidades um cientista de dados deve ter. Por exemplo, menciona que um cientista de dados deve ter a capacidade de aprender o problema, ser um bom comunicador de resultados, saber analisar dados, saber sobre visualização, raciocinar eticamente, entre outros.

Rise of the data scientist (Disponível em: <https://flowingdata.com/2009/06/04/rise-of-the-data-scientist>)

Este texto traz outra descrição das diferentes habilidades que um cientista de dados precisa ter. Foi publicado em 2009, época em que o conceito de Ciência de Dados começou a nascer e quão importante seria no futuro próximo.

Curiosidades e inquisições (perguntas) sobre dados

Nesta unidade, reforçaremos os conceitos vistos anteriormente na Ciência de Dados, mas vamos concentrar-nos em uma habilidade especial que é muito importante no início do processo da Ciência de Dados. A leitura desta unidade nos dá alguns exemplos para praticar.

The hardest thing in data science

(Disponível em: <https://buckwoody.wordpress.com/2015/12/30/the-hardest-thing-in-data-science>)

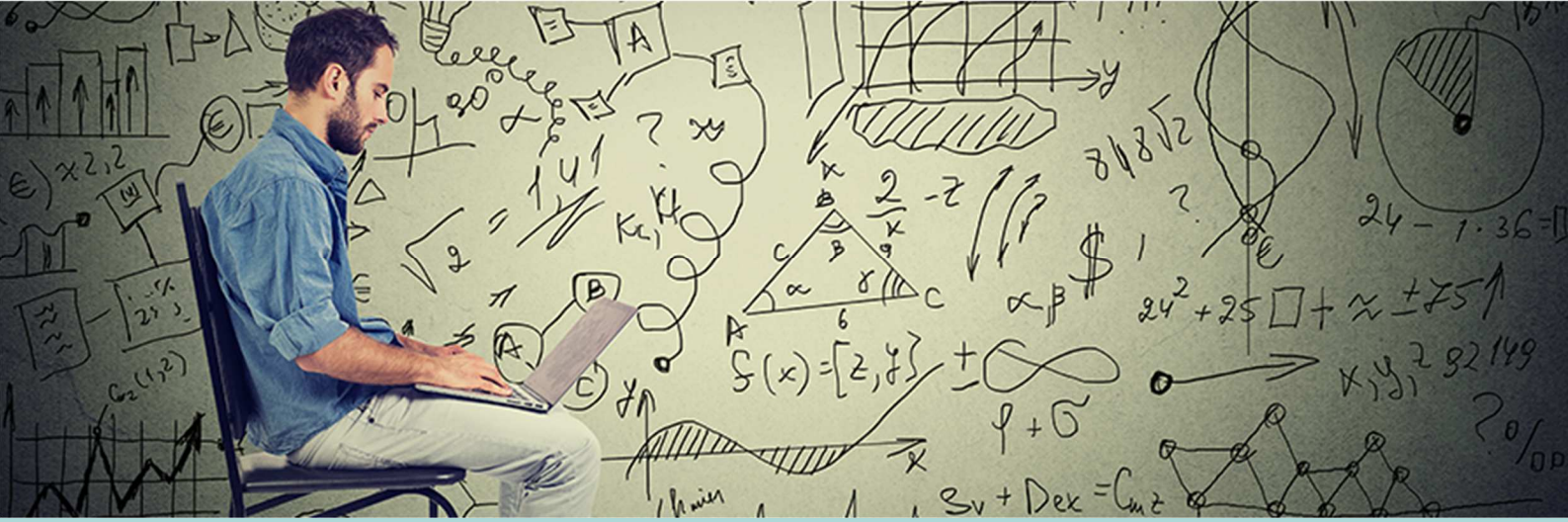
Fazer boas perguntas é tanto uma ciência quanto uma arte. Este *post* descreve esta como a tarefa mais difícil na Ciência de Dados. É por isso que entender os objetivos da empresa – ou do cliente – e as limitações dos seus dados parece ser pré-requisito fundamental para fazer perguntas interessantes.

The data science design manual (SKIENA, 2017) Seção 1.2

Apresenta três conjuntos de dados diferentes, nos quais é solicitado que use o seu potencial analítico e desenvolva questões interessantes sobre cada um dos conjuntos de dados. Essa é uma das habilidades que você precisará desenvolver e reforçar se quiser ser um bom cientista de dados.

Conclusão do módulo

Este módulo teve como principal objetivo explicar os fundamentos da Ciência de Dados, começando com a revolução de dados que gerou o movimento *Big Data*. Também indicou textos que explicam como o uso adequado de dados nos ajuda a tomar decisões com mais fundamentos, pois são baseados em fatos e dados, e não apenas em suposições. No fim deste módulo, o aluno também deve ter conhecimento de quais são as habilidades necessárias para ser considerado um cientista de dados.



MÓDULO II – PROBLEMAS E SOLUÇÕES EM CIÊNCIA DE DADOS

Neste módulo, vamos concentrar-nos nos exemplos em que a Ciência de Dados foi usada. Primeiro, raciocinamos sobre quais resultados possíveis poderiam ser gerados e como eles são gerados. Além disso, estudaremos recomendações que devemos seguir para que os nossos projetos sejam bem-sucedidos e não fracassem. Finalmente, veremos em que pontos a ciência dos dados, no seu estado atual, pode ajudar-nos a resolver os nossos problemas e quais são as suas limitações.

Ciência de Dados na vida real

A partir de agora, veremos alguns exemplos de problemas reais em que a Ciência de Dados foi usada para resolvê-los. Nós ainda não explicamos quais técnicas poderiam ser usadas em cada caso, como veremos com mais calma nos módulos seguintes. Em cada caso, tente identificar quais são os problemas que estão aparecendo para um cientista de dados ao enfrentar um problema real.

Principles of data science (OZDEMIR, 2016) Capítulo 1, Seção *Data science case studies*

Três estudos de caso sobre dados científicos e a descrição dos seus resultados. O propósito desta leitura é começar a raciocinar sobre como os resultados foram gerados. Não se preocupe se ainda é difícil responder a essa pergunta, pois nos próximos módulos você aprenderá as técnicas necessárias.

Fatores de sucesso na Ciência de Dados

Nesta unidade, veremos quais são os segredos para que o nosso projeto de Ciência de Dados seja bem-sucedido de duas perspectivas diferentes. É muito importante nesta primeira fase do nosso curso identificar o que devemos e não devemos fazer.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 7, Seção *Data science project principles*: why projects succeed or fail

Quais são os fatores necessários para que um projeto de Ciência de Dados seja bem-sucedido? Ou há um fator que devemos evitar para que o nosso projeto não falhe? Esta leitura detalha os fatores comuns que determinam o sucesso dos projetos de Ciência de Dados.

***Defining success*: four secrets of a successful data science experiment**

(Disponível em: <https://simplystatistics.org/2016/06/03/defining-success>)

Este pequeno *post* nos dá outra perspectiva de quais fatores são importantes para que um projeto de Ciência de Dados seja bem-sucedido.

Limitações na Ciência de Dados

Embora seja verdade que a Ciência de Dados está nos ajudando a resolver muitos problemas hoje em dia, ainda há muitas limitações. Por exemplo, devemos perguntar-nos se os dados que temos são suficientes para tomar uma decisão. Nesta unidade, estudaremos alguns desses casos.

Data science for business (PROVOST; FAWCETT, 2013) Capítulo 14, Seção *What data can't do*: humans in the loop, revisited

Este texto apresenta tarefas em que os seres humanos são muito melhores que os computadores e vice-versa. A Ciência de Dados usa a *expertise* de humanos com poder computacional. No entanto, existem certas limitações de que devemos estar cientes e não pensar que, porque temos dados e a perícia dos humanos, o nosso problema já está resolvido.

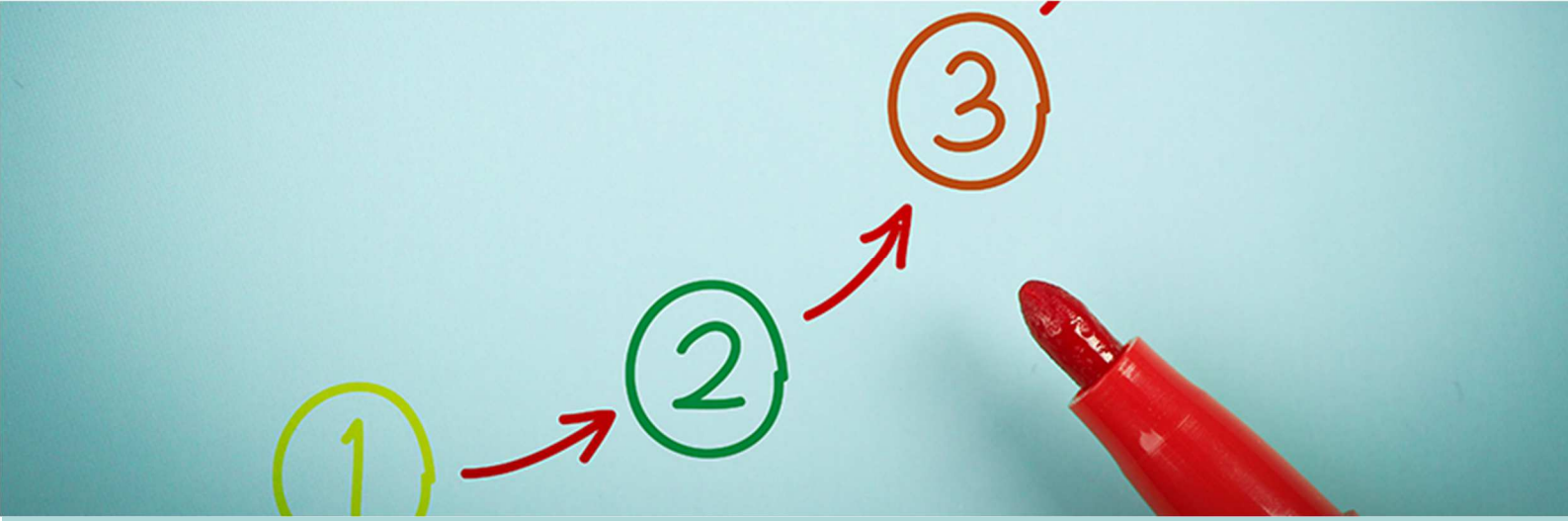
Limitations of predictive analytics: lessons for data scientists

(Disponível em: <http://www.dataversity.net/limitations-predictive-analytics-lessons-data-scientists>)

Neste texto, vemos de outra perspectiva as limitações da Ciência de Dados. O texto descreve alguns exemplos bem conhecidos em que a Ciência de Dados falhou.

Conclusão do módulo

Neste módulo, apontamos exemplos de Ciência de Dados na vida real. Embora seja verdade que até agora não conhecemos os estágios do processo de Ciência de Dados, conhecer casos reais será muito útil para entender mais facilmente seus estágios no próximo módulo. Além disso, o aluno deve ter descoberto que nem todos os problemas podem ser resolvidos com a Ciência de Dados. Não há receita que nos diga exatamente quais passos tomar para que o nosso projeto seja bem-sucedido, no entanto, neste módulo, vimos algumas recomendações que nos ajudarão a atingir as nossas metas e minimizar falhas.



MÓDULO III – CIÊNCIA DE DADOS E SUAS ETAPAS

No contexto de Ciência de Dados, os processos de extração de conhecimento e tomada de decisão são tipicamente executados seguindo um conjunto de etapas, as quais podem ser descritas como: preparação dos dados, exploração dos dados, escolha de técnicas e modelos, ajuste de modelos e, finalmente, avaliação e uso dos modelos. Neste módulo do curso, aprenderemos o porquê e a importância de cada etapa, assim como as dificuldades envolvidas em cada uma delas. O módulo inicia com um estudo sobre os possíveis tipos de dados que são usualmente encontrados em problemas reais.

Tipos de dados

Depois de estudar esta unidade, o estudante terá conhecimento sobre o que são dados estruturados e não estruturados, planilhas de dados, os seus atributos e as suas instâncias. Estudaremos ainda os diferentes tipos de dados comumente encontrados em problemas reais.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 2

Apresenta uma discussão informal sobre os diferentes tipos de dados que aparecem em problemas reais e como os tipos de dados afetam a escolha de métodos e modelos de análise. O capítulo traz ainda uma discussão sobre o que são bases de dados e como são geradas.

Statistical data type

(Disponível em: <https://en.wikipedia.org/wiki/Statistical_data_type>)

Fornece uma tabela com os principais tipos de dados, discutindo aplicações e problemas típicos onde cada tipo de dado aparece, ressaltando como cada tipo de dado deve ser tratado em termos estatísticos.

Numsense! Data science for the layman: no math added (NG; SOO, 2017) Seção 1.1

Apresenta uma discussão sucinta sobre os tipos de dados, evitando um linguajar matemático.

Experimental design and analysis (SELTMAN, 2018) Capítulo 2

Discute os tipos de dados do ponto de vista matemático, seguindo uma linguagem formal e bastante precisa em termos do rigor matemático.

Preparação dos dados

Nesta unidade, discutiremos questões relacionadas à qualidade e integridade de dados. Em particular, apresentaremos os conceitos de transformação e engenharia de atributos, além de abordarmos problemas que tipicamente aparecem em bases de dados, por exemplo, dados faltantes, inconsistentes e redundantes.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 3

Discute o problema de limpeza e tratamento de dados desde as etapas de aquisição dos dados, analisando possíveis fontes de erros. A questão do armazenamento dos dados também é abordada.

Data cleansing

(Disponível em: <https://en.wikipedia.org/wiki/Data_cleansing>)

Discute de modo informal o problema de limpeza e tratamento de dados, apresentando *softwares* existentes, limitações e desafios.

A general introduction to data analytics (MOREIRA *et al.,* 2018) Seções 4.1 a 4.4

Apresenta os problemas de limpeza e tratamento de dados de modo formal, com uma linguagem mais matemática, discutindo as ferramentas matemáticas e computacionais tipicamente empregadas no tratamento de tais problemas.

The data science design manual (SKIENA, 2017) Seção 3.3

Esta seção descreve os potenciais problemas presentes em bases de dados reais e como tais problemas podem afetar, ou até mesmo inviabilizar, o processo de ciência de dados.

An introduction to data science (SALTZ; STANTON, 2018) Capítulos 1 e 2

Apresenta o problema de limpeza e tratamento de dados de modo informal, evitando um linguajar matemático.

Explorando e analisando os dados

Nesta unidade, discutiremos a importância de explorar os dados antes de iniciar o processo de modelagem do problema. Vamos discutir duas ferramentas básicas bastante empregadas nessa tarefa: estatística descritiva e visualização. Ao final, o estudante terá uma visão clara de como explorar uma base de dados a fim de entender as propriedades e particularidades dos seus dos atributos.

A general introduction to data analytics (MOREIRA *et al.*, 2018) Capítulo 2

Fornece noções básicas de ferramentas estatísticas tipicamente empregadas em ciência de dados, mostrando como tais ferramentas podem ser empregadas para fins de exploração e conhecimento do comportamento dos dados.

Exploratory data analysis

(Disponível em: <https://en.wikipedia.org/wiki/Exploratory_data_analysis>)

Traz um resumo sobre a etapa de exploração e análise de dados, adotando o ponto de vista estatístico, o que requer alguma familiaridade com conceitos básicos de estatística para acompanhar a narrativa.

Experimental design and analysis (SELTMAN, 2018) Capítulo 4

Apresenta com bom rigor matemático – estatístico – o problema de analisar e explorar dados, requerendo alguma familiaridade com conceitos básicos de estatística por parte do leitor.

The data science design manual (SKIENA, 2017) Capítulo 6

Descreve técnicas e recursos de visualização que podem ser empregados no processo de exploração dos dados, visando principalmente a compreender como os dados estão distribuídos nos atributos.

Selecionando métodos e ajustando modelos

Nesta unidade, discutiremos como construir um modelo que "aprenda" o comportamento dos dados, permitindo identificar grupos de instâncias semelhantes ou então realizar predições a partir das informações contidas nos dados. Ao final desta unidade, o estudante terá conhecimento da diferença entre métodos descritivos e preditivos e quais classes de problemas esses métodos buscam resolver.

Data science, The MIT Press **(**KELLEHER; TIERNEY, 20180 Capítulo 5

Apresenta uma discussão informal sobre a questão de identificar o tipo de problema, as técnicas a serem empregadas a cada tipo de problema e os resultados que podem ser obtidos a partir de cada técnica.

Moving into data science as a career**: mathematical models**

(Disponível em: <https://towardsdatascience.com/moving-into-data-science-as-a-career-mathematical-models-e13f30690b00>)

Discute o problema de modelagem matemática em ciência de dados a partir da perspectiva das habilidades matemáticas e computacionais dos profissionais que atuam na área.

Numsense**! Data science for the layman: no math added** (NG; SOO, 2017) Seção 1.2 e 1.3

Apresenta o problema de como definir modelos matemáticos utilizando um linguajar não matemático, sendo uma discussão pouco profunda.

The data science design manual (SKIENA, 2017) Seções 7.1 a 7.3

Fornece uma discussão mais técnica sobre os tipos de problemas e técnicas relacionadas a cada um deles, apresentando uma taxonomia que permite claramente identificar e classificar problemas e métodos.

Avaliando métodos e modelos

Nesta unidade, discutiremos a questão da avaliação de métodos descritivos e preditivos. Em particular, estudaremos os diferentes tipos de métricas que podem ser utilizados e as suas particularidades com relação ao tipo de dado e método. Discutiremos ainda, no caso de modelos preditivos, como organizar os dados a fim de não incorrermos em avaliações errôneas ou equivocadas.

A general introduction to data analytics (MOREIRA *et al.*, 2018) Seção 5.2

Fornece uma visão geral do problema de avaliar a qualidade dos métodos e dos modelos empregados no processo de extração e de geração de conhecimento.

Evaluating a machine learning model

(Disponível em: <https://www.jeremyjordan.me/evaluating-a-machine-learning-model>)

Discute de forma clara e ilustrativa os métodos utilizados para avaliar modelos matemáticos utilizados em ciência de dados, discutindo métricas e recursos computacionais.

Important model evaluation error metrics everyone should know

(Disponível em: <https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics>)

Apresenta de forma objetiva, porém pouco formal, a questão de avaliar a qualidade dos modelos matemáticos utilizados em Ciência de Dados, discutindo as estratégias de avaliação mais utilizadas.

***Numsense*! Data science for the layman: no math added** (NG; SOO, 2017) Seção 1.4

Apresenta o problema de avaliar modelos matemáticos utilizando uma linguagem não matemática.

***Data mining*: practical machine learning tools and techniques** (WITTEN et al., 2016) Capítulo 5

Esta leitura é bastante técnica e fornece uma base sólida de conhecimentos sobre os procedimentos empregados para avaliar métodos e modelos tipicamente empregados em ciência de dados.

Conclusão do módulo

Neste módulo do curso, foram indicadas as leituras para o estudo das principais etapas do processo de extração e geração de conhecimento, a partir de dados. As leituras e os estudos realizados devem ter deixado claro que bases de dados reais apresentam inconsistências, dados faltantes e atributos redundantes, demandando processamento específico antes de serem utilizadas. O aluno deve ter adquirido ainda uma visão dos métodos empregados na exploração dos diferentes tipos de dados, mostrando a importância de ferramentas estatísticas e de visualização nesse contexto. Finalmente, o aluno deve ter compreendido a diferença entre métodos descritivos e preditivos, o tipo de problema que cada um desses métodos pode tratar e a importância de avaliar a eficácia de tais métodos.

MÓDULO IV – MÉTODOS MATEMÁTICOS E COMPUTACIONAIS

Neste módulo, estudaremos ferramentas matemáticas e computacionais comumente empregadas nas diversas etapas do processo de extração e geração de conhecimento em Ciência de Dados. O foco deste módulo não é um detalhamento rigoroso das ferramentas matemáticas e computacionais, mas, sim, apresentar os conceitos envolvidos de forma simples e intuitiva, de modo que o estudante seja capaz de compreender o que tais ferramentas são capazes de realizar e quais são as suas limitações.

Técnicas para tratamento e transformação de dados

Depois de estudar a unidade 4.1, o estudante terá conhecimento de como tratar dados faltantes, como detectar informações discrepantes – os chamados *outliers* – e como realizar transformações de escala, muito utilizadas em problemas reais. A importância e o impacto das transformações de escala nos métodos descritivos e preditivos também serão discutidos e ilustrados.

A general introduction to data analytics (MOREIRA *et al.*, 2018) Seções 4.1 a 4.4

Descreve os principais problemas presentes em bases de dados reais, fornecendo uma discussão simplificada de como tratar tais problemas.

Dealing with missing data: key assumptions and methods for applied analysis (SOLEY-BORI, 2013) Seções 1, 2, 3 e 4.1

Discute o problema específico de dados faltantes e por que são tão frequentes em aplicações reais. Traz ainda uma discussão clara sobre as principais estratégias para contornar o problema de dados faltantes, discutindo as vantagens e desvantagens de cada uma delas.

Imputation: statistics

(Disponível em: <https://en.wikipedia.org/wiki/Imputation_(statistics)>)

Discute a questão de inserir informações para remover dados faltantes, apresentando de forma resumida diferentes técnicas matemáticas e computacionais tipicamente empregadas nesse contexto.

The data science design manual (SKIENA, 2017) Seção 3.3 e 4.3

A seção 4.3 discute a questão de normalização dos dados e as suas implicações. O tema é abordado com um linguajar matemático bastante formal.

Outlier analysis (AGGARWAL, 2017) Capítulo 1, Seção 1.1 a 1.3

Este é um livro todo dedicado ao problema de detecção de *outliers*. As seções 1.1 a 1.3 descrevem o problema de forma geral e apresentam estratégias simples, porém bastante empregadas, de como abordar tal problema.

Análise de componentes principais

Nesta unidade, estudaremos uma das mais importantes técnicas de análise e transformação de dados, a chamada Análise de Componentes Principais – do inglês *Principal Component Analysis* (PCA). A técnica PCA será apresentada de modo que o estudante possa compreender como ela funciona, qual o tipo de análise que ela proporciona e quais são as suas limitações.

A tutorial on principal components analysis (SMITH, 2002) Seção 3

Este artigo discute a técnica PCA do ponto de vista estatístico. O interessante é que traz toda a fundamentação matemática necessária para a compreensão da dedução matemática do método. Alunos com pouca familiaridade com conceitos básicos de estatística devem ler os capítulos 1 e 2.

Principal component analysis

(Disponível em: <http://setosa.io/ev/principal-component-analysis>)

Este *site* fornece uma ferramenta on-line por meio da qual se pode interagir com o conjunto de dados e analisar o efeito da interação no cálculo das componentes principais.

A one-stop shop for principal component analysis

(Disponível em: <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>)

Apresenta o conceito de Análise de Componentes Principais com um linguajar informal e não matemático. Fornece ainda uma lista de recursos computacionais e material *on-line* relacionado com o tema.

***Numsense*! Data science for the layman: no math added** (NG; SOO, 2017) Capítulo 3

Apresenta o conceito de Análise de Componentes Principais com um linguajar informal e não matemático, buscando fornecer uma descrição intuitiva.

A tutorial on principal component analysis (SHLENS, 2014)

Este artigo fornece uma visão intuitiva do método PCA, apresentando em seguida um detalhamento matemático de como o método funciona. Estudantes menos familiarizados com noções de álgebra linear podem concentrar-se nas seções I, II e III.A.

Técnicas de agrupamento

Técnicas de agrupamento são métodos descritivos que buscam encontrar conjuntos de objetos semelhantes, sendo extremamente úteis para revelar padrões que caracterizam conjuntos específicos de objetos. Nesta unidade, estudaremos duas das principais técnicas de agrupamento, o método Kmeans e agrupamento hierárquico, analisando as suas diferenças, vantagens e limitações.

A general introduction to data analytics (MOREIRA *et al.,* 2018) Capítulo 5, Seções 51 e 5.3

Traz uma discussão mais detalhada sobre a questão da escolha da métrica para diferentes tipos de dados e o impacto nos resultados de agrupamentos, apresentando exemplos.

K-means clustering

(Disponível em: <https://en.wikipedia.org/wiki/K-means_clustering>)

Aborda a técnica K-means de forma bastante detalhada, discutindo a sua formulação matemática e os aspectos computacionais. Traz ainda uma lista de métodos que são variantes da técnica K-means.

Hierarchical clustering

(Disponível em: <https://en.wikipedia.org/wiki/Hierarchical_clustering>)

Apresenta o método de agrupamento hierárquico forma intuitiva, ilustrando as principais etapas do processo. Traz ainda uma lista de softwares existentes para o cálculo de agrupamento hierárquico.

***Numsense*! Data science for the layman: no math added (NG; SOO, 2017) Capítulo 2**

Discute a técnica K-means com um linguajar informal e não matemático, buscando fornecer uma descrição intuitiva.

Introduction to data mining (TAN et al. 2005) Capítulo 8, Seções 8.1 a 8.3

Apresenta de forma bastante rigorosa as técnicas K-means e agrupamento hierárquico, enfatizando aspectos matemático e computacionais. Fornece ainda uma discussão detalhada das vantagens e desvantagens de cada uma das técnicas.

The data science design manual (SKIENA, 2017) Capítulo 10, Seção 10.5

Discute as técnicas de agrupamento K-means e agrupamento hierárquico de forma clara e objetiva, apresentando ainda uma breve descrição de como avaliar técnicas de agrupamento.

Modelos de regressão

Modelos de predição, calculados via técnicas de regressão, são métodos preditivos muito utilizados para inferir valores numéricos a partir de atributos também numéricos. Nesta unidade, estudaremos uma classe particular de técnicas de regressão voltadas para a geração dos chamados modelos lineares. Aprenderemos o que são tais modelos, como interpretá-los e como avaliar a sua eficácia.

A general introduction to data analytics (MOREIRA *et al.*, 2018) Capítulo 8, Seções 8.1, 8.2.1, 8.2.2 e 8.2.3

Fornece uma discussão detalhada sobre como avaliar modelos de predição, com ênfase na questão de dividir dados de treinamento em treinamento e teste. Apresenta ainda uma discussão importante sobre Bias e Variância.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 4, p. 97-120

Apresenta o problema de regressão linear com um linguajar não matemático, fornecendo uma descrição intuitiva e acompanhada de diversos exemplos.

Linear regression

(Disponível em: <https://en.wikipedia.org/wiki/Linear_regression>)

Apresenta de forma bastante detalhada e completa o problema de regressão linear, discutindo a questão de como interpretar o resultado da regressão e como generalizar o problema para outros tipos de regressão.

***Numsense*! Data science for the layman: no math added** (NG; SOO, 2017) Capítulo 6

Discute o problema de regressão com um linguajar informal e não matemático, buscando fornecer uma descrição intuitiva.

The data science design manual (SKIENA, 2017) Capítulo 9, Seções 9.1, 9.2 e 9.5-9.5.1

Apresenta os principais conceitos relacionados com regressão linear para fins de predição, fornecendo uma discussão sobre como analisar os modelos gerados e a sua precisão.

Modelos de classificação

Modelos preditivos de classificação buscam classificar objetos a partir de informações contidas nos seus atributos. Nesta unidade, estudaremos um conjunto de técnicas de classificação largamente empregadas em problemas de Ciência de Dados. Aprenderemos como tais modelos funcionam, quais são as suas limitações e como avaliar a sua eficácia.

Existe um grande número de técnicas de classificação, e várias delas são largamente empregadas em problemas reais. A fim de apresentar as técnicas com um bom nível de detalhe, dividimos esta unidade em duas partes. A primeira parte inicia discutindo uma técnica bastante simples, vizinhos mais próximos, abordando em seguida os métodos de regressão logística e Naive Bayes, muito empregados em problemas reais. A segunda parte desta unidade aborda as técnicas SVM e Árvores de Decisão, que são consideradas mais sofisticadas. O conceito de comitês, *bootstrap* e *boosting* também são abordados na segunda parte desta unidade.

Parte I

A general introduction to data analytics (MOREIRA *et al.,* 2018) Capítulo 9

Discute técnicas de classificação baseadas em regressão, Naive Bayes e vizinhos mais próximos, detalhando, como avaliar a precisão de tais técnicas.

Logistic regression

(Disponível em: <https://en.wikipedia.org/wiki/Logistic_regression>)

Apresenta de forma detalhada a técnica de regressão logística, com uma discussão sobre como interpretar o modelo gerado. O conteúdo discute ainda a questão de ponderar, de forma distinta, diferentes classes.

Naive Bayes classifier

(Disponível em: <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>)

Apresenta a técnica de Naive Bayes no contexto de dados contínuos, discutindo as distribuições de probabilidade que são tipicamente utilizadas como modelos para os atributos.

***Numsense*! Data science for the layman: no math added** (NG; SOO, 2017) Capítulo 7

Discute a técnica de vizinhos mais próximos com um linguajar informal e não matemático, buscando fornecer uma descrição intuitiva.

Introduction to data mining (TAN et al. 2005) Capítulo 5, Seções 5.2 e 5.3

Discute as técnicas de vizinhos mais próximos e Naive Bayes de forma bastante detalhada, apresentando a fundamentação matemática que suporta tais técnicas.

The data science design manual (SKIENA, 2017) Capítulo 9, Seção 9.6 e 9.7; Capítulo 10, Seção 10.2; Capítulo 11, Seção 11.1

As seções 9.6 e 9.7 discutem como métodos de regressão podem ser utilizados também para classificação, apontando as dificuldades e limitações. A seção 10.2 discute a técnica de vizinhos mais próximos para fins de classificação, e a seção 11.1 apresenta a técnica de Naive Bayes.

Parte II

A general introduction to data analytics (MOREIRA *et al.,* 2018) Capítulo 10, Seção 10.1 e 10.2.2

Fornece uma introdução às Árvores de Decisão e SVM, discutindo ainda como elas podem ser utilizadas para realizar predição em vez de classificação.

Decision trees in machine learning

(Disponível em: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>)

Discute de forma resumida as árvores de decisão, com um linguajar não matemático. Apresenta uma lista das vantagens e desvantagens das árvores de decisão.

A complete tutorial on tree based modeling from scratch (in R & Python)

(Disponível em: <https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python>)

Fornece uma discussão bastante detalhada sobre árvores de decisão, apresentando exemplos de métricas utilizadas no particionamento. Discute ainda os conceitos de "poda", comitês, *bootstrap* e *boosting*.

Support Vector Machine (SVM) Tutorial

(Disponível em: <https://blog.statsbot.co/support-vector-machines-tutorial-c1618e635e93>)

Discute a técnica SVM com um linguajar não matemático, apresentado diversos exemplos e ilustrações do conceito de Kernel e como ele funciona.

***Numsense*! Data science for the layman: no math added** (NG; SOO, 2017) Capítulos 8, 9 e 10

Discute a técnica SVM com um linguajar informal e não matemático, buscando fornecer uma descrição intuitiva.

Introduction to data mining (TAN et al. 2005) Capítulo 5, Seções 5.5 e 5.6; Capítulo 4, Seção 4.3

Apresenta de forma bastante detalhada, e com rigor matemático, as técnicas SVM e Árvores de Decisão. O texto traz ainda uma discussão sobre comitês, *bootstrap* e *boosting*, fornecendo um conjunto e exemplos e comparações.

The data science design manual (SKIENA, 2017) Capítulo 11, Seções 11.2 a 11.4

Descreve técnicas de classificação baseadas em árvores de decisão, discutindo estratégias de *boosting* e comitês. Na seção 11.4, a técnica SVM é apresentada de forma simples e intuitiva, discutindo a questão do uso de *kernels* para o caso não linear.

Conclusão do módulo

Este módulo do curso abordou as principais ferramentas matemáticas e computacionais empregadas nas diversas etapas do processo de Ciência de Dados. As leituras e os estudos realizados apresentaram ferramentas essenciais para tratar dados faltantes, detectar objetos discrepantes – *outliers* –, transformar e normalizar atributos. O estudante deve ter compreendido a importância de tais ferramentas para viabilizar as demais etapas do processo de extração e geração de conhecimento. Os estudos devem também ter tornado clara a importância da técnica PCA como ferramenta de análise e transformação de dados. O estudante deve ter adquirido ainda noções de como funcionam e quais são as vantagens e desvantagens de duas das principais técnicas de agrupamento existentes, K-means e Agrupamento Hierárquico. As leituras e os exercícios sobre técnicas de predição e classificação devem ter sido capazes de fornecer ao aluno uma base de conhecimento sobre como tais técnicas funcionam, quais são as suas principais diferenças, vantagens e limitações. O aluno deve ter compreendido ainda os conceitos de comitê, *bootstrap* e *boosting*, importantes no contexto de predição e classificação.

MÓDULO V – EXEMPLOS REAIS

Neste módulo, estudaremos em mais detalhes exemplos reais da ciência de dados. Nos módulos anteriores, já vimos quais técnicas poderiam ser usadas em cada estágio. Nos exemplos a seguir, explicaremos as decisões tomadas passo a passo e analisaremos os resultados obtidos.

Estrutura de um projeto de Ciência de Dados

Nesta primeira unidade, faremos uma breve revisão das diferentes etapas do processo da Ciência de Dados. Nós já vimos esse processo no módulo 3, mas vamos revisá-lo brevemente aqui para poder alinhar os nossos exemplos.

A general introduction to data analytics (MOREIRA *et al.*, 2018) Seção 1.7

Como qualquer projeto, precisamos de um plano. Nesta leitura, a metodologia CRISP-DM é usada. Você verá que, em essência, é muito semelhante ao que foi apresentado no módulo 3. No entanto, para facilitar a leitura do primeiro estudo de caso na próxima unidade, seguiremos as etapas dessa metodologia.

Case study 1 – Câncer de mama em Wisconsin

Esta unidade apresenta o nosso primeiro estudo de caso e a sua solução passo a passo. Neste exemplo, usaremos um algoritmo de *clustering* que já estudamos no módulo 4.

A general introduction to data analytics (MOREIRA *et al.,* 2018) Seção 1.6.1 e Capítulo 7

A seção 1.6.1 descreve o problema do câncer de mama e o objetivo de detectar diferentes padrões de tumores de mama que podem ser usados em diagnósticos futuros. O capítulo 7 mostra, passo a passo, como é possível usar a Ciência de Dados para resolver esse problema.

Case study 2 – Prevendo preços de ações baseados em mídias sociais

Esta unidade apresenta o nosso segundo estudo de caso e a sua solução passo a passo.

Principles of data science (OZDEMIR 2016) Capítulo 13, Seção Case study 1 – Predicting stock prices based on social media"

Neste segundo estudo de caso, tentamos prever o preço das ações usando o sentimento das redes sociais. Para resolver o problema, nenhuma técnica de aprendizado de máquina é usada, apenas com a análise exploratória dos dados podemos alcançar o nosso objetivo.

Conclusão do módulo

Neste módulo, vimos em mais detalhes exemplos com as suas soluções usando algumas das técnicas descritas nos módulos anteriores. As leituras mostram passo a passo o processo da Ciência de Dados e as saídas geradas em cada etapa. Da mesma forma, foi explicado por que certas decisões foram tomadas e por que certas técnicas foram usadas. Além disso, outro exemplo foi descrito em detalhes no vídeo.

Neste ponto do curso, espera-se que os alunos já tenham a capacidade de pensar sobre os seus próprios problemas e como eles podem ser resolvidos usando a Ciência de Dados. No entanto, não entramos em detalhes de implementação em uma linguagem de programa – por exemplo, Python –, uma vez que isso requer um maior conhecimento de programação, que pode ser estudado em cursos subsequentes.

MÓDULO VI – QUESTÕES ÉTICAS EM CIÊNCIA DE DADOS

Finalmente, chegamos ao último módulo do nosso curso. Vimos vários conceitos, técnicas e ideias que são agora a base do nosso conhecimento em Ciência de Dados. Com tudo aprendido, poderemos seguir um curso mais profundo sobre cada uma das técnicas ensinadas.

Neste último módulo, vamos ver alguns conceitos muito importantes. Nesta era do *Big Data*, temos o poder de acessar muitos dados, no entanto, isso também traz uma grande responsabilidade de saber como agir eticamente com o único objetivo de tornar o mundo melhor. No entanto, existem alguns pontos que podem parecer óbvios, mas é bom conhecê-los. As próximas três unidades lidam mais de perto com cada uma delas.

Privacidade e segurança

A Ciência de Dados pode ajudar a tornar o mundo muito melhor, mas também pode prejudicar a sociedade ou os indivíduos. Nas leituras abaixo, descreveremos alguns aspectos éticos e de privacidade que todo cientista de dados deve conhecer. Por exemplo, devemos ter cuidado na forma como comunicamos os nossos resultados, manter a segurança de dados confidenciais e cuidar da privacidade em dados agregados, entre outros.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 6 *Privacy and ethics* (p. 182-205)

Este artigo aborda a segurança e os aspectos éticos dos dados. Como sabemos, a Ciência de Dados trabalha com dados que podem conter informações sensíveis. Por exemplo, o caso de uma cadeia de supermercados que criou um modelo para atribuir uma pontuação de gravidez a cada cliente é apresentado. Com base nesses modelos, uma menina começou a receber uma série de *e-*

mails com produtos para gravidez, razão pela qual o pai registrou uma queixa. No entanto, nos dias seguintes, foi anunciado que a filha estava grávida, mas ela não quis divulgá-lo. Muitos problemas como esses podem acontecer quando trabalhamos com dados confidenciais, e nós, como cientistas de dados, precisamos saber como lidar com isso.

The data science design manual (SKIENA, 2017) Seção 12.7 Societal and ethical implications

Esta seção curta descreve os problemas éticos com *Big Data*.

Calling bullshit

Nesta unidade, vamos definir *bullshit* no contexto de *Big Data*. Além disso, apresentaremos alguns exemplos que ajudam a entender claramente esse conceito. O objetivo é que, uma vez terminadas as leituras, elas consigam reconhecer *bullshit* onde as encontrarem.

On bullshit (FRANKFURT, 1986)

Este documento descreve o que é *bullshit* além de alguns conceitos e categorias de *bullshit*. Para nós, *bullshit* é linguagem, números estatísticos, discursos e outras representações de dados destinados a persuadir as pessoas com um desrespeito flagrante pela verdade e pela coerência lógica.

Deeper into bullshit (COHEN, 2002)

Este artigo estuda mais profundamente o problema da besteira. Em particular, ele analisa e critica o artigo original do documento. A abordagem dada neste documento é mais orientada para o trabalho acadêmico.

Próxima geração de cientistas de dados

Finalmente, nesta unidade, veremos quais características queremos que os futuros cientistas de dados tenham, descrevendo claramente os aspectos relacionados às habilidades que precisam ser desenvolvidas.

***Doing data science*: straight talk from the frontline**. (O'NEIL; SCHUTT, 2014) Capítulo 16, Seção *What are next-gen data scientists*?

Quais características um cientista de dados da próxima geração deve ter? Nas leituras indicadas nos módulos anteriores, vimos quais são as habilidades atuais necessárias; no entanto, o que mais queremos? Esta leitura aborda algumas características necessárias, como a atitude ética mencionada acima, e o que pensamos mais profundamente sobre o *design* e o processo para tornar o mundo um lugar melhor, e não pior.

Conclusão do módulo

No final deste módulo, o aluno deve ser capaz de identificar os diferentes problemas éticos que surgem quando se trabalha em Ciência de Dados. Os dados nos dão poder, mas também exigem muita responsabilidade. O aluno deve ser capaz de reconhecer e evitar *bullshit*, da mesma forma, explicar por que é considerado assim. Finalmente, este módulo mostra a perspectiva futura da ciência de dados e que esperamos dos próximos cientistas de dados.

CONLUSÃO

Finalmente, chegamos ao fim do curso Introdução à Ciência de Dados. O conteúdo deste curso foi projetado para pesquisadores e profissionais de diferentes áreas; por isso, começamos o nosso estudo a partir de conceitos básicos, mas necessários. Os módulos 1 e 2 nos deram uma perspectiva muito ampla dessa área, e começamos a ver problemas em casos reais. Nos módulos 3 e 4, abordamos alguns detalhes sobre os aspectos matemáticos e computacionais necessários para iniciar a análise dos dados. Embora seja verdade que não entramos em detalhes de implementação, uma vez que isso requer um nível mais avançado de programação, o conteúdo apresentado é um primeiro passo para incentivar os alunos a aprender programação e implementar as técnicas e algoritmos descritos aqui. No módulo 5, colocamos o conhecimento em prática e vimos o resultado em cada etapa do processo da ciência de dados. Finalmente, no módulo 6, aprendemos os aspectos éticos que acompanham o trabalho com dados confidenciais e privados.

Esperamos que todos os conceitos e técnicas ensinados ao longo da disciplina os motivem a aprofundar o conteúdo de cada um dos módulos, além de colocar em prática o seu espírito cientista de dados e começar a trabalhar em um projeto com *data science*.

BIBLIOGRAFIA

AGGARWAL, C. C. *Outlier Analysis*. Springer, 2017.

CADY, Field. *The data science handbook*. Hoboken, NJ: John Wiley & Sons, 2017.

COHEN, Gerald A. Deeper into bullshit. *Contours of agency*: essays on themes from Harry Frankfurt, p. 321-339, 2002. Disponível em: <http://phil480.weebly.com/uploads/8/1/2/6/8126749/cohen_and_frankfurt_2002.pdf>.

FRANKFURT, Harry. *On bullshit*, 1986. Disponível em: <http://www2.csudh.edu/ccauthen/576f12/frankfurt__harry_-_on_bullshit.pdf>.

KELLEHER, J. D.; TIERNEY, B. *Data science, The MIT Press*, 2018.

MOREIRA, J. M.; CARVALHO, A.; HORVÁTH, T. *A general introduction to data analytics.* Hoboken: Wiley, 2018.

NG, A.; SOO, K. *Numsense*! Data science for the layman: no math added. Annalyn Ng & Kenneth Soo, 2017.

O'NEIL, Cathy; SCHUTT, Rachel. *Doing data science*: straight talk from the frontline. Sebastopol, California: O'Reilly Media, Inc., 2013.

OZDEMIR, Sinan. *Principles of data science*. Packt Publishing Ltd., 2016.

PROVOST, Foster; FAWCETT, Tom. *Data science for business*: what you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.

SALTZ, J. S.; STANTON, J. An introduction to data science. *SAGE Publications*, 2018.

SELTMAN, H. J. *Experimental design and analysis*, 2018. Disponível em: <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>.

SHLENS, J. *A tutorial on principal component analysis*, 2014. Disponível em: <https://arxiv.org/abs/1404.1100>.

SKIENA, S. The data science design manual, Springer, 2017.

SMITH, L. *A tutorial on principal components analysis*. Technical Report OUCS-2002-12, 2002. Disponível em: <http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf>.

SOLEY-BORI, M. *Dealing with missing data*: key assumptions and methods for applied analysis, Technical Report n. 4, Boston University, 2013. Disponível em: <https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf>.

TAN, P.-N.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. *Introduction to data mining*, Pearson, 2005. Disponível em: <https://www-users.cs.umn.edu/~kumar001/dmbook/ch7_clustering.pdf>.

WITTEN, Ian H. et al. *Data mining*: practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann, 2016.

PROFESSOR-AUTOR

Jorge Poco, Professor Associado da Escola de Matemática Aplicada na Fundação Getulio Vargas. Graduado em Engenharia de Sistemas pela Universidad Nacional de San Agustín, Peru. Mestre em Ciência da Computação pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, Brasil. Doutor em Ciência da Computação pela Polytechnic School of Engineering da New York University, USA. Possui artigos publicados em periódicos e conferências de grande impacto internacional, por exemplo, IEEE Transaction on Visualization and Computer Graphics e Computer Graphics Forum. Recebeu o Prêmio de Excelência em Pesquisa na Universidad Católica San Pablo e um Certificado de Excelência da New York University.

FGV EMAp