Sociedade Brasileira de Educação Matemática

RIPEM

Ciência de dados para cidadãos informados: Aprendizagem na interseção de alfabetização em dados, estatística e justiça social

Joachim Engel
Ludwigsburg University of Education
Alemanha
✉ engel@ph-ludwigsburg.de
0000-0002-8251-6670

Laura Martignon
Ludwigsburg University of Education
Alemanha
✉ martignon@ph-ludwigsburg.de
0000-0002-0318-2327

2238-0345 ISSN

10.37001/ripem.v14i3.3816 doi

Recebido • 01/03/2024
Aprovado • 13/05/2024
Publicado • 20/08/2024

Editor • Gilberto Januario

Resumo: A ciência de dados, como ciência prática, foi concebida para tratar de problemas tangíveis nas ciências, na tecnologia e na sociedade. Educar os alunos em ciência de dados vai além do ensino sobre algoritmos, habilidades de manipulação de conjuntos de dados, seleção e aplicação de análises adequadas e criação e interpretação de representações visuais de dados. Envolve também o aumento de uma compreensão crítica de como os dados são produzidos e como podem ser usados para fins específicos, incluindo a função do contexto na interpretação dos dados. Enfatiza o desenvolvimento de uma consciência para a ética dos dados e da consideração das implicações para a política e a sociedade, quando são usados algoritmos poderosos. A participação na democracia, na sociedade digital e de dados de hoje, exige o desenvolvimento de uma série de habilidades transversais que precisam ser fomentadas nas instituições educacionais por meio de pedagogias criticamente orientadas que entrelaçam habilidades e práticas de dados técnicos com alfabetização estatística e midiática. Com base em uma análise das tendências e necessidades de proteger os valores democráticos em uma sociedade com dados e nas próprias reflexões sobre as práticas de ensino, este artigo apresenta recomendações sobre a criação de cursos de ciência de dados para desenvolver cidadãos informados.

Palavras-chave: Alfabetização estatística. Educação em ciência de dados. Ética de dados. Cidadania.

Data science for informed citizen: Learning at the intersection of data literacy, statistics and social justice

Abstract: Data science as a practical science has been conceived to address tangible problems in science, technology and society. Educating students in data science goes beyond teaching about algorithms, skills of manipulating data sets, selecting and applying appropriate analyses, and creating and interpreting visual representations of data. It also involves raising a critical understanding of how data are produced and how they can be used for particular purposes, including the role of context in interpreting data. It emphasizes developing an awareness for data ethics, and considering the implications for policy and society when powerful algorithms are used. Participation in democracy, in today's digital and datafied society requires the development of a series of transversal skills which need to be fostered in educational institutions through critically oriented pedagogies that interweave technical data skills and practices together with statistical and media literacies. Based on an analysis of trends and needs in open societies and on own reflections of teaching practices this paper gives recommendations on designing data science courses to develop informed citizen.

Keywords: Statistical Literacy. Data Science Education. Data Ethics. Citizenship

Ciencia de datos para ciudadanos informados: Aprendizaje en la intersección de la alfabetización informática, la estadística y la justicia social

Resumen: La ciencia de datos como ciencia práctica se ha concebido para abordar problemas tangibles de la ciencia, la tecnología y la sociedad. Educar a los estudiantes en la ciencia de datos va más allá de enseñar algoritmos, habilidades para manipular conjuntos de datos, seleccionar y aplicar análisis adecuados y crear e interpretar representaciones visuales de los datos. También implica suscitar una comprensión crítica de cómo se producen los datos y cómo pueden utilizarse para fines concretos, incluyendo el papel del contexto en la interpretación de los datos. Hace hincapié en el desarrollo de una conciencia ética de los datos, en la consideración de las implicaciones para la política y la sociedad del uso de potentes algoritmos. La participación en la democracia, en la sociedad digital y cargada de datos de hoy en día, requiere el desarrollo de una serie de competencias transversales que deben fomentarse en las instituciones educativas a través de pedagogías de orientación crítica que entrelacen las habilidades y prácticas técnicas de datos junto con la alfabetización estadística y mediática**.** Basándose en un análisis de las tendencias y necesidades de las sociedades abiertas y en reflexiones propias sobre las prácticas docentes, este documento ofrece recomendaciones sobre el diseño de cursos de ciencia de datos para formar ciudadanos informados.

Palabras clave: Alfabetización estadística. Educación en ciencia de datos. Ética de los datos. Ciudadanía

1 Introdução

O cenário das informações está mudando drasticamente na era digital devido à crescente disponibilidade de informações pela Internet, ao uso generalizado de tecnologias digitais, à abundância de dados e ao fácil acesso a ferramentas de análise de dados. mídia digital e a disponibilidade de dados de escopo e magnitude absolutamente ilimitados mudam nosso acesso à informação de forma radical nosso acesso às informações. As fontes de dados emergentes fornecem novos tipos de evidências, provocam novos tipos de perguntas, possibilitam novos tipos de respostas e moldam as maneiras pelas quais as evidências são usadas para influenciar a tomada de decisões na vida privada, profissional e pública. Em um mundo cada vez mais orientado por dados, as mudanças sociais, societárias e tecnológicas exigem novas competências. Essa expansão afeta não apenas o mundo profissional, mas todos nós. A inovação, o progresso social e o bem-estar de nossa sociedade civil exigem que as pessoas nos campos da ciência, dos negócios, da política e da sociedade saibam como avaliar e entender os dados para desenvolver uma compreensão informada de nosso mundo e enfrentar os desafios urgentes da sociedade com percepções empíricas e argumentos sólidos baseados em dados.

Ao mesmo tempo, o Big Data, com suas possibilidades de vigilância, manipulação e controle, apresenta sérios problemas para a democracia e a liberdade (consulte, por exemplo, Helbing et al., 2017). A capacidade de avaliar a credibilidade das informações e de suas fontes nunca foi tão importante. O World Risk Report[1] , publicado pela Fundação Fórum Econômico Mundial da Suíça em janeiro de 2024, considera a desinformação como a maior ameaça à humanidade nos próximos dois anos, à frente de eventos climáticos extremos, polarização social e conflitos armados.

[1] https://www.weforum.org/publications/global-risks-report-2024/

Algoritmos baseados em dados são usados para traçar o perfil de membros da sociedade e tomar decisões cruciais que provavelmente afetam de forma desproporcional aqueles com menos privilégios e recursos à sua disposição (O Neill, 2016). O modelo da Amazon para classificar pedidos de emprego[2], por exemplo, demonstrou ser contra as mulheres. Os problemas do Facebook foram expostos pela primeira vez pelo escândalo da Cambridge Analytica, e a empresa continua enfrentando muitas questões éticas. Cathy O'Neil, em seu livro Weapons of Math Destruction (Armas de Destruição Matemática) (O'Neill, 2016), aponta os perigos e as injustiças do uso de modelos algorítmicos para determinar a pontuação de crédito, o preço das apólices de seguro, se alguém pode ser colocado em liberdade condicional ou até mesmo quais crimes a polícia deve investigar. Não aprender a entender, analisar e contestar dados fará com que os cidadãos fiquem em uma posição de desvantagem informacional cada vez maior em relação aos agentes sociopolíticos e comerciais. Consequentemente, a educação para a alfabetização em dados precisa abordar uma visão ampla dos dados tanto como fatores sociais e técnicos. Consequentemente, a educação em alfabetização de dados e a ciência de dados não pode ser reduzida ao aprendizado do domínio técnico sobre algoritmos, o gerenciamento de big data e computação.

Com todas as promessas da ciência estatística de criar um mundo melhor (assim é o slogan do International Statistical Institute), há sérias preocupações éticas na medida que mais e mais atividades humanas são transcritas em dados, quantificadas e analisadas (Van Es e Schäfer, 2017). As decisões tomadas por empresas e órgãos governamentais são cada vez mais orientadas por dados e algoritmos, enquanto os processos pelos quais os dados são gerados, comunicados e representados, não são necessariamente transparentes nem desprovidos de efeitos negativos (O'Neil, 2016). As pessoas geralmente não sabem por que, como ou, até mesmo, que os dados sobre si mesmas estão sendo coletados, analisados e "compartilhados" com outras partes (Dalton et al., 2016). Em uma sociedade cada vez mais informatizada, os dados geralmente recebem o status de fato objetivo, apesar de sua natureza construída, parcial eenviesada. .

Com base na revisão da literatura, em uma análise das necessidades de sustentar estruturas democráticas e na reflexão das próprias práticas de ensino, este artigo tem como objetivo fornecer orientação sobre como projetar a educação em ciência de dados para o cidadão informado por meio do pensamento crítico e da capacitação para a participação cívica. A seguir, descrevemos na Seção 2 a necessidade de a alfabetização em dados fazer parte da educação geral em qualquer nível educacional e seus desafios para o novo campo emergente da educação em ciência de dados (Seção 3) antes de explicarmos na Seção 4 nosso conceito de implementação de alguns elementos da ciência de dados e apresentarmos nossos objetivos em aulas para alunos que se preparam para ser professores do ensino médio. Para esse grupo, não se pressupõe nenhuma formação matemática ou tecnológica específica além do ensino médio, de modo que o conceito e os objetivos podem ser aplicados a qualquer outro grupo de cidadãos instruídos e informados. Nós nos concentramos em como a ciência de dados usa algoritmos do „machine learning“ como uma das suas metodologias para analisar dados, fazer previsões e automatizar processos de tomada de decisões. Por fim, na Seção 5, resumimos a necessidade de uma abordagem abrangente que enfatize as reflexões sobre o impacto social dos aplicativos de ciência de dados.

2 Alfabetização em dados - desafios para os educadores do século 21

[2] https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html

A ciência de dados, como uma ciência prática, foi concebida para tratar de problemas tangíveis na ciência, na tecnologia e na sociedade. Educar os alunos em ciência de dados vai além de ensinar sobre algoritmos, habilidades de manipulação de conjuntos de dados, seleção e aplicação de análises apropriadas e criação e interpretação de representações visuais de dados. Envolve também o aumento de uma compreensão crítica de como os dados são produzidos e como podem ser usados para fins específicos, incluindo a função do contexto na interpretação dos dados. Enfatiza o desenvolvimento de uma consciência para a ética dos dados e a consideração das implicações para a política e a sociedade quando algoritmos poderosos são usados. *Portanto, usar dados reais não é suficiente, precisamos ensinar ciência de dados para abordar problemas reais!* Os algoritmos não são o objetivo da ciência de dados, eles são uma ferramenta importante. Como comparação: A física não tem a ver com cálculo, mas com a compreensão dos fenômenos naturais. O cálculo é uma ferramenta para a física. O ensino de ciência de dados para cidadãos informados e, por exemplo, futuros professores de matemática, requer conteúdos diferentes de um curso de ciência de dados para graduados em ciência da computação, engenharia de dados ou estatística.

O cidadão capacitado na era da digitalização precisa de habilidades para navegar em uma superabundância de dados a fim de tomar decisões informadas - na vida cotidiana, bem como em vários níveis sociais e políticos. "A alfabetização em dados abrange as habilidades em dados que são importantes para todas as pessoas em um mundo moldado pela digitalização. É uma parte indispensável da educação geral", de acordo com a Data Literacy Charter 2021 iniciada pela „Stifterverband“ - uma iniciativa de empresas e fundações na Alemanha dedicada a promover melhorias na educação, ciência e inovação - e endossada por várias associações e indivíduos (Schüller et al., 2021). No centro da alfabetização em dados estão as habilidades para coletar, gerenciar, avaliar e adquirir novos conhecimentos a partir dos dados de forma crítica. Isso inclui explicitamente habilidades para avaliar criticamente os dados e seu impacto na interação social e política.

Para articular o conhecimento, apoiar posições com evidências, testar suposições ou avaliar probabilidades em situações de risco e incerteza, e a capacidade de explorar e entender os dados é essencial. Para que as pessoas obtenham informações e avaliem sua qualidade para formar opiniões informadas e independentes, a alfabetização midiática também é necessária. Justamente porque as pessoas são repetidamente chamadas em suas vidas públicas e privadas a tomar decisões que vão além da sua esfera real de conhecimento e experiência, é importante que elas aprendam a obter informações confiáveis, fazer perguntas para entender melhor os fatos e tomar decisões baseadas em fatos (Lengnink et al., 2014). A alfabetização digital é um objetivo central da preparação dos jovens para a era digital (OECD, 2019). A alfabetização em dados inclui a capacidade de coletar, gerenciar, avaliar e aplicar dados de forma crítica. Ela é fundamental para transformar sistematicamente os dados em conhecimento. O Data Literacy Charter enfatiza que a alfabetização em dados fortalece o julgamento, a autodeterminação e o senso de responsabilidade e promove a participação social e econômica de todas as pessoas em um mundo moldado pela digitalização (Schüller et al., 2021). A alfabetização em dados serve para promover a maturidade em um mundo digitalizado moderno e, portanto, é importante para todas as pessoas, não apenas para os especialistas. O objetivo da educação em alfabetização de dados é que cada indivíduo e nossa sociedade como um todo, lidem com os dados de maneira consciente e ética. A alfabetização em dados possibilita ações bem-sucedidas e sustentáveis, baseadas em evidências e que levam em consideração a incerteza e as mudanças em nosso ambiente de vida.

A alfabetização em dados precisa fazer parte da missão educacional geral das escolas no século XXI e diz respeito à educação matemática de maneira especial, além da educação em

ciência da computação e ciências sociais (Wolfram, 2020; Messy Data Coalition, 2020).

Para que os alunos "apreciem e analisem criticamente a incorporação social e a construção dos dados" (Richterich, 2018), é necessário trabalhar com dados reais que demonstrem problemas autênticos concretos usando fontes de dados abertamente acessíveis (Engel, 2017; Ridgway, 2015). Os dados abertos podem ser analisados pelos alunos, capacitando-os a formular, perguntar e investigar questões socialmente candentes e, assim, exercer seus direitos como cidadãos (Ridgway, 2021).

Por fim, enfrentar os desafios de uma sociedade orientada por dados não é apenas uma questão de projeto curricular ou de responsabilidade individual do professor, mas requer estratégias e políticas institucionais para apoiar os professores no desenvolvimento da alfabetização em dados. Isso requer um plano coerente de mudança sistêmica. Em alguns contextos, isso pode começar com a integração de elementos de alfabetização de dados e estatística em cursos tradicionais. Em outros contextos, para ensinar a transmitir a alfabetizacao tradicional, pode ser adequado utilizar simplesmente conjuntos de dados autênticos e em grande escala relacionados com problemas sociais.

3 O surgimento da educação em ciência de dados

Embora vivamos em uma era de inundação de dados, o ensino de estatística no ensino médio e fora dele está muito focado em um paradigma do século XX, desenvolvido em uma época em que os dados vinham de estudos planejados, o software era caro e o objetivo de um estudo estatístico era calcular um valor p ou encontrar um intervalo de confiança. Hoje, os dados estão universalmente presentes e o software de alta qualidade é barato ou até mesmo gratuito. Muitas barreiras que impediam os alunos do ensino médio de analisar dados no passado não existem mais.

A educação em ciência de dados deve ir além de uma combinação de recursos numéricos, estatísticos e técnicos, incluir pensamento crítico, cidadania e promover habilidades para avaliar, analisar e interpretar dados. Os programas educacionais devem ir além dos recursos de dados e incluir o pensamento crítico, promover habilidades para avaliar, analisar e interpretar dados e seu significado para a política e a sociedade. A alfabetização em dados precisa fazer parte da missão educacional geral das escolas no século XXI e diz respeito à educação matemática de maneira especial, além da educação em ciência da computação e ciências sociais

A participação na democracia, na sociedade digital e de dados de hoje, exige o desenvolvimento de uma série de habilidades transversais, que precisam ser fomentadas nas instituições educacionais por meio de pedagogias criticamente orientadas que entrelaçam as habilidades e práticas técnicas de dados com a alfabetização em informação e mídia. A educação em ciência de dados deve ir além de uma combinação de recursos numéricos, estatísticos e técnicos, incluir pensamento crítico, cidadania e promover habilidades para avaliar, analisar e interpretar dados. Os programas educacionais devem ir além dos recursos de dados e incluir o pensamento crítico (Van Es e Schäfer, 2017), promover habilidades para avaliar, analisar e interpretar dados e seu significado para a política e a sociedade (ProCivicStat Partners, 2018; Schield, 2004). Essa abordagem pode capacitar os alunos a questionar a ética, as estruturas e a economia do uso de dados e, fundamentalmente, a aparente inevitabilidade da vigilância e da datafição de todos os aspectos da vida cotidiana (Atenas et al., 2020).

Atualmente, existem vários conceitos, propostas e experiências para introduzir elementos de ciência de dados na sala de aula nos níveis de ensino fundamental e médio. O

Projeto Internacional de Ciência de Dados nas Escolas (IDSSP)[3], uma colaboração internacional de estatísticos, cientistas da computação e educadores, desenvolveu um currículo para apresentar a ciência de dados aos alunos nos dois últimos anos do ensino médio, bem como um currículo para capacitar os professores sobre como ensinar ciência de dados. Outras iniciativas inovadoras recentes são, por exemplo, ProDaBi[4] (Biehler & Fleischer, 2021) e Mobilize IDS[5] (Gould, 2021).

As abordagens da ciência de dados em sala de aula muitas vezes podem ser vinculadas, em termos de conteúdo, à abordagem de questões sociais que estão na mente de muitos, como mudanças climáticas, eventos pandêmicos, igualdade de renda etc. (Engel 2017). Organizações apartidárias compilaram uma grande quantidade de informações que estão disponíveis publicamente na Internet para qualquer pessoa usar como informação e discussão - desde o trabalho das Nações Unidas sobre os Objetivos de Desenvolvimento Sustentável para medir o progresso social até os escritórios de estatísticas nacionais que coletam informações sobre emprego, renda e migração, passando por organizações não governamentais que monitoram as mudanças climáticas ou a saúde dos cidadãos (Ridgway 2015). Plataformas como Gapminder ou Our World of Data fornecem acesso de baixo custo para monitorar a situação do mundo, desde o desenvolvimento humano e a felicidade global ("World Happiness Report") até as taxas de infecção por COVID-19 e as mudanças climáticas.

4 Ciência de dados para cidadãos informados: Objetivos gerais

Nesta seção, discutimos um relato pessoal do que acreditamos que todo cidadão educado e informado do século XXI deve saber sobre „machine learning“ e tomada de decisão automatizada. Nossas afirmações são discutíveis, e o leitor pode achar que alguns objetivos estão faltando, pode enfatizar outros mais do que nós, ou discordar em maior grau. No entanto, os conteúdos a seguir foram implementados e avaliados em uma série de aulas para futuros professores do ensino médio em matemática e estudos sociais. Não se pressupõe nenhuma preparacao matemática nem tecnológica específica além do ensino médio, portanto, o conceito e as metas podem se aplicar a qualquer grupo de cidadãos educados e informados. A ciência de dados é um campo interdisciplinar que usa métodos científicos, processos, algoritmos e sistemas para extrair conhecimento e percepções dos dados com técnicas de estatística, análise de dados e „machine learning“ (ML). O „machine learning“ é usado na ciência de dados para fazer previsões ou classificar dados. Ele inclui algoritmos como árvores de decisão, „random forests“, redes neurais e muitos outros que identificam automaticamente padrões e tomam decisões com base em dados.

4.1 Aumentar a conscientização sobre onde, na vida cotidiana, encontramos produtos de ciência de dados

Os produtos de ciência de dados e, especialmente, os algoritmos de „machine learning“tornaram-se parte integrante de nossa vida cotidiana, muitas vezes de maneiras que muitos de nós podem não reconhecer imediatamente. Aqui estão alguns exemplos de alguns domínios da vida cotidiana em que os produtos de aprendizado de máquina são encontrados

- Smartphones e dispositivos: Nossos smartphones são um centro de aplicativos de„machine learning“. Desde assistentes de voz, como Siri, Google Assistant e Alexa, até texto preditivo, tradução de idiomas e recursos de autocorreção em teclados e

[3] http://www.idssp.org
[4] https://www.prodabi.de/en/
[5] https://www.introdatascience.org

reconhecimento facial para desbloqueio de dispositivos, todos são alimentados por algoritmos de„machine learning“..

- Comércio eletrônico e compras on-line: Sites como Amazon e eBay usam o „machine learning“ para recomendar produtos. Esses sistemas analisam o comportamento de compras anteriores, o histórico de pesquisas e o que outros usuários semelhantes compraram ou visualizaram.
- Detecção de fraudes: Os algoritmos de „machine learning“ ajudam a detectar transações fraudulentas, identificando padrões que podem indicar atividade fraudulenta.
- Bancos e finanças: Os modelos de „machine learning“ ajudam a avaliar a capacidade de crédito de uma pessoa, analisando grandes quantidades de dados e padrões financeiros.
- Educação: Aplicativos e plataformas educacionais usam o „machine learning“ para adaptar o conteúdo ao ritmo e ao estilo de aprendizado de cada aluno, fornecendo exercícios e feedback personalizados.

Esses são apenas alguns exemplos que ilustram como as tecnologias de „machine learning“ aprimoram e facilitam muitos aspectos de nossas rotinas diárias, muitas vezes tornando os serviços mais eficientes, personalizados e fáceis de usar. Temos certeza de que seus alunos terão mais exemplos. Eles também sabem onde seus registros pessoais são rastreados e armazenados e como seus dados estão sendo usados?

4.2 Ensinar a conscientização sobre a qualidade e a adequação dos dados

Os dados - a base empírica para decisões informadas por evidências e para a criação de conhecimento - são certamente preferíveis a anedotas, pensamentos positivos, superstição, preconceito ou ideologia. No entanto, os dados em si não são fatos nem verdade. Alguns autores consideram os dados como modelos da realidade. Os dados não fornecem representações objetivas do mundo. Eles podem surgir de forma oportunista ou como resultado de decisões conscientes tomadas por alguém para pesquisar um determinado tópico. Em geral, os dados são coletados a um custo alto, para um propósito específico e com um projeto de pesquisa especificamente escolhido. Eles medem variáveis manifestas de uma maneira específica. Eles são a base para a construção de variáveis latentes com base em algum tipo de modelo com um conceito específico em mente. Em um nível mais complexo, pode-se perguntar por que determinadas medidas foram escolhidas, por quem e para que fins. A medição está sempre ligada a alguma teoria do fenômeno que está sendo estudado. Um economista pode usar uma conceituação diferente do termo pobreza em seu trabalho do que um sociólogo. No exemplo das ciências naturais, a massa, o comprimento e o tempo foram escolhidos como medidas não por serem "óbvios", mas porque sua medição permite que previsões precisas sobre o mundo físico sejam feitas e usadas. O bem-estar das nações foi medido pelo produto interno bruto por pessoa, ou PIB, até que essa medida monolítica foi questionada por Amartya Sen e substituída pelo Índice de Desenvolvimento Humano (IDH), mais abrangente.

A coleta de dados não é uma atividade de lazer, mas é um trabalho laborioso, às vezes tedioso, que geralmente exige muito esforço e recursos financeiros. Ela atende aos interesses de alguém, e é legítimo questionar de quem é esse interesse. Por que esses dados foram coletados? Os dados coletados implicitamente contam uma história. De quem é essa história? E de quem não é essa história?

Perguntas críticas ou reflexivas sobre os métodos usados em pesquisas podem incluir (mas não se limitam a):

- As medidas (por exemplo, questionários) estão bem definidas? As medidas são robustas e apropriadas para os fins para os quais estão sendo usadas?

- Os metadados (ou seja, explicações detalhadas de como as variáveis foram definidas, características da amostra etc.) estão disponíveis?
- Os procedimentos de amostragem foram adequados? Quem está faltando nos dados coletados? (por exemplo, medir como os cidadãos se sentem em relação a um determinado tópico analisando os fluxos de mídia social não é uma amostra de não usuários).

Muitos estudos nas ciências sociais estão preocupados com teorias de causalidade; a causalidade está associada a desafios filosóficos difíceis que vão muito além de mantras simples como "correlação não implica causalidade". Entretanto, quando os dados são provenientes de estudos observacionais, pesquisas ou dados de arquivo, e não de estudos experimentais, pode ser difícil determinar uma identificação confiável de relações de causa e efeito.

Além do conhecimento técnico sobre os processos de geração de dados, é importante que os indivíduos sejam capazes de fazer perguntas críticas para avaliar a credibilidade e a validade de quaisquer dados, descobertas ou conclusões que encontrarem, tanto em bases técnicas quanto lógicas - até mesmo dados ou relatórios de fontes presumivelmente confiáveis, como órgãos oficiais de estatística. É importante examinar, a partir de uma perspectiva crítica, as narrativas e interpretações dos dados e as conclusões tiradas a partir deles, por exemplo:

- Qual é a qualidade das evidências apresentadas em um artigo da mídia ou em uma alegação para apoiar afirmações sobre políticas ou ações necessárias (por exemplo, em relação a leis de reciclagem, igualdade salarial ou vacinação)?
- Quão razoáveis são as projeções e quão apropriados são os modelos estatísticos subjacentes e as suposições que foram aplicadas para analisar dados sobre questões importantes (por exemplo, sobre a progressão do aquecimento global ou a taxa de propagação de infecções como a pandemia do coronavírus COVID-19)?
- Quando são feitas afirmações sobre uma correlação entre variáveis (por exemplo, tabagismo e risco de morte), supõe-se que as relações sejam lineares, e a pergunta é se elas são realmente assim (ou talvez curvilíneas)? Mais importante ainda, se forem assumidos processos causais ou relações de causa e efeito, existem contas rivais plausíveis, covariáveis ou fatores intervenientes inexplorados que poderiam afetar os resultados? Geralmente, há várias maneiras igualmente válidas de descrever um fenômeno social.
- As conclusões são consistentes com outras evidências disponíveis? Quando são feitas propostas para políticas sociais, pode-se perguntar se a identificação do problema foi feita adequadamente e se foram usados dados relevantes.

Osborne e Pimentel (2019) fornecem uma heurística usando uma árvore de decisão rápida e econômica que até mesmo os não especialistas ("pessoas de fora competentes") podem usar para avaliar a credibilidade dos argumentos baseados na ciência. Eles sugerem três perguntas simples:

1. A fonte dessas informações é confiável? As evidências de credibilidade são dadas por: Não há conflitos de interesse, a fonte é reconhecida, a análise do tópico é imparcial
2. A fonte tem a experiência necessária para atestar a afirmação? As evidências de conhecimento especializado e experiência são fornecidas por meio do histórico, da reputação entre os colegas, das credenciais e do contexto institucional e por meio da experiência profissional relevante.
3. Existe um consenso entre os especialistas científicos relevantes?

Se a resposta a todas as três perguntas for Sim, então aceite a alegação de acordo com sua heurística. Se a resposta apenas à pergunta 3 for NÃO, enquanto 1) e 2) forem respondidas

positivamente, pergunte sobre a natureza da discordância. Qual é a opinião dos especialistas mais conceituados? Que variedade de descobertas são consideradas plausíveis? E quais são os riscos de estar errado?

Todos devem adotar uma atitude questionadora e saber quais perguntas fazer sobre a natureza, as limitações ou a credibilidade de diferentes fontes de dados, mensagens estatísticas e conclusões. Entretanto, adotar uma postura crítica ao avaliar as evidências não significa ser crítico a todo custo. Em vez disso, a crítica consiste em adotar a atitude de um cético imparcial que está disposto a aceitar uma apresentação, mas precisa ser convencido pelas evidências. Em situações em que os dados são apresentados de forma enganosa, os alunos devem ser incentivados a reapresentá-los de forma mais apropriada; em situações em que os dados são duvidosos (ou falsificados), eles devem ser incentivados a encontrar dados relevantes em fontes confiáveis.

4.3 Ensinar a conscientização sobre os vieses das decisões de máquina

Os algoritmos de aprendizado de máquina podem apresentar muitos vieses por diversos motivos, muitas vezes refletindo problemas nos dados em que são treinados, no projeto do próprio algoritmo ou nos contextos sociais e históricos mais amplos em que são desenvolvidos e implantados. Aqui estão vários fatores importantes que contribuem para o viés no aprendizado de máquina.

- Dados de treinamento enviesados : Qualquer viés no conjunto de treinamento será amplificado no conjunto de teste, levando a decisões enviesadas . Se os dados usados para treinar um modelo de aprendizado de máquina contiverem vieses - seja por sub-representação ou super-representação de determinados grupos, seja por vieses históricos presentes nos dados -, o modelo provavelmente aprenderá e perpetuará esses vieses.
  Em outubro de 2019, pesquisadores descobriram que um algoritmo usado em mais de 200 milhões de pessoas em hospitais dos EUA para prever quais pacientes provavelmente precisariam de cuidados médicos adicionais favorecia muito os pacientes brancos em detrimento dos negros. Embora a raça em si não fosse uma variável usada nesse algoritmo, outra variável altamente correlacionada à raça era o histórico de custos de assistência médica. O raciocínio foi que o custo resume quantas necessidades de saúde uma determinada pessoa tem. Por vários motivos, os pacientes negros incorreram, em média, em custos de assistência médica mais baixos do que os pacientes brancos com as mesmas condições.
- Viés algorítmico: O design do próprio algoritmo pode introduzir um viés. Alguns algoritmos podem ser mais propensos a ampliar as tendências presentes nos dados de treinamento. Por exemplo, os algoritmos que penalizam fortemente os outliers podem não ter um bom desempenho para grupos minoritários que estão sub-representados nos dados de treinamento.
  Para prever o desempenho acadêmico de alunos de escolas, os órgãos reguladores de qualificação nacional no Reino Unido criaram fórmulas para atribuir notas de exames previstas com base na previsão dos professores - o algoritmo atribuiu notas mais baixas aos alunos de escolas públicas e notas melhores (ainda melhores do que a previsão dos professores) aos alunos de escolas independentes menores (Smith, 2020).
- Contexto histórico e social: Os contextos sociais, históricos e culturais nos quais os dados são gerados geralmente contêm vieses. Como os modelos de aprendizado de máquina aprendem com dados anteriores, eles podem inadvertidamente aprender e perpetuar esses preconceitos sociais. As ferramentas de reconhecimento facial que utilizam algoritmos de „machine learning“ estão sendo usadas para a aplicação da lei

no Brasil, causando um debate acalorado sobre se esses algoritmos têm preconceitos racistas e propagam as desigualdades existentes[6] (Silva, 2022).

Sem dúvida, o exemplo mais notável de viés de ML é o algoritmo COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) usado nos sistemas judiciais dos EUA para prever a probabilidade de um réu se tornar reincidente. Devido aos dados usados, ao modelo escolhido e ao processo de criação do algoritmo em geral, o modelo previu duas vezes mais falsos positivos para reincidência de infratores negros (45%) do que de infratores brancos (23%) (Shin, 2020).

Abordar o viés no „machine learning“ é um grande desafio que exige atenção cuidadosa a todo o ciclo de vida do desenvolvimento do modelo, desde a coleta e a preparação dos dados até a implantação e o monitoramento. As estratégias para atenuar o viés incluem o uso de conjuntos de dados mais diversificados e representativos, a aplicação de técnicas de redução de viés durante o treinamento do modelo e o monitoramento e a atualização contínuos dos modelos para garantir resultados justos e equitativos. Os dados usados precisam representar "o que deveria ser" e não "o que é". Caso contrário, como no algoritmo de contratação da Amazon, o risco é alto de sub-representar e causar discriminação contra um determinado grupo de pessoas. A validade dos algoritmos precisa ser avaliada quando aplicada a vários grupos sociais. Além disso, a implementação precisa de algum tipo de governança de dados obrigatória e aplicada para garantir uma prática que seja ética em relação aos valores de uma sociedade livre (Shin, 2020).

4.4 Ensinar a conscientização sobre o impacto dos produtos de ciência de dados na sociedade

Criar uma consciência crítica sobre o impacto do aprendizado de máquina na sociedade é fundamental para formar cidadãos informados que possam contribuir para o desenvolvimento, a implantação e a governança éticos dessas tecnologias.

Nas escolas, podemos integrar discussões sobre os impactos sociais da tecnologia, incluindo o aprendizado de máquina, no currículo em vários níveis educacionais. Isso pode variar de simples lições sobre alfabetização digital em escolas de ensino fundamental a debates mais complexos sobre ética em cursos de ensino médio e universitários. A ética no aprendizado de máquina engloba dimensões jurídicas, políticas, sociais e econômicas. Portanto, é adequada uma abordagem interdisciplinar que enfatize a cooperação com outros campos, como ética, filosofia, sociologia ou ciência política. Essa abordagem garante que os alunos não apenas aprendam como os algoritmos de „machine learning“ funcionam, mas também entendam suas implicações sociais mais amplas, incluindo as dimensões econômicas, sociais e éticas

Uma maneira eficaz é envolver os alunos em projetos em que eles tenham a oportunidade de projetar, implementar ou criticar sistemas de „machine learning“ com considerações éticas em mente. Isso pode incluir tarefas como a criação de um modelo de ML que leve em conta possíveis vieses ou o desenvolvimento de diretrizes para a implantação ética de sistemas deML.

Os tópicos que são acessíveis aos alunos da escola ou da universidade incluem preocupações com privacidade, vigilância, parcialidade e justiça, autonomia e o futuro do trabalho. Uma preocupação específica refere-se à opacidade (geralmente chamada „de caixas pretas") da maioria dos algoritmos de„machine learning“ . Delegar a tomada de decisões a uma máquina, especialmente em áreas críticas como a justiça criminal ou decisões médicas de vida

[6] https://brazilian.report/podcast/2024/01/17/algorithms-ai-facial-recognition-racist/

ou morte, levanta questões éticas e morais graves. Embora poucas pessoas possam se importar com a forma como uma máquina traduz um documento de um idioma para outro (como este artigo, em uma primeira abordagem, foi traduzido do inglês para o português por DeepL[7]), desde que a tradução seja precisa, as decisões que envolvem direitos humanos pessoais nunca devem ser deixadas para a máquina.

4.5 Ensine alguns fundamentos tecnológicos sobre o aprendizado de máquina

Embora os algoritmos sejam a ferramenta e não o objetivo da ciência de dados, para apreciar a natureza específica do aprendizado de máquina, até mesmo os alunos que não estão se formando em ciência da computação precisam aprender um pouco da tecnologia sobre como uma máquina pode fazer algo que chamamos de "aprendizado". Uma boa introdução são as regras de decisão automatizadas, representadas por árvores de classificação (Breiman et al., 1984). As árvores são intuitivas, simples de aplicar, fáceis de entender e fornecem resultados facilmente interpretáveis. As árvores de decisão criadas algoritmicamente a partir de dados de treinamento são ferramentas simples, mas poderosas, capazes de alcançar alta precisão em muitas tarefas e, ao mesmo tempo, altamente interpretáveis. O "conhecimento" aprendido por uma árvore de decisão aparece como uma estrutura hierárquica, um plano para decisões. Essa estrutura mantém e exibe o conhecimento de tal forma que mesmo os não especialistas podem aplicá-lo imediatamente.

Os algoritmos baseados em árvores são um método importante de „machine learning“ que dá suporte à tomada de decisões, por exemplo, em medicina, finanças, políticas públicas e muito mais. As árvores abrem as portas para tópicos mais avançados de ciência de dados e ML (por exemplo, Random Forests , ensacamento e reforço, bem como conceitos fundamentais, como conjuntos de treinamento e sobreajuste). No entanto, em vez de começar com um algoritmo de computador que produz árvores ideais, sugerimos que os alunos primeiro construam suas próprias árvores, um nó por vez, para explorar como elas funcionam e quão bem. Esse processo de construção própria é mais transparente do que o uso de algoritmos como o CART (conforme implementado, por exemplo, no pacote `part` or `tree` em R ou `sklearn` em Python). Acreditamos que isso ajudará os alunos não apenas a compreender os fundamentos das árvores, mas também a entender melhor os algoritmos de construção de árvores quando eles os encontrarem.

Isso pode começar completamente desconectado. Em uma atividade prática, os alunos recebem dados em cartões. Cada cartão tem altura, peso e várias outras medidas, inclusive algumas irrelevantes (cor dos olhos) para um jogador profissional de handebol ou futebol. Os alunos trabalham para descobrir como prever o esporte com base nos outros atributos. Eles perceberam que poderiam lançar seu algoritmo como uma árvore de classificação. Por fim, obtiveram cartões não vistos anteriormente para testar a árvore; isso levou, entre outras coisas, à descoberta do sobreajuste: um fenômeno em que o uso de variáveis irrelevantes resultou em uma árvore excelente, talvez até perfeita, para os dados de treinamento, mas que ficou pior com os novos dados de teste.

Em uma próxima etapa, usamos a plataforma de análise de dados on-line comum (CODAP) (Finzer, 2019) e seu plug-in ARBOR (Erickson, 2019). O CODAP é um pacote gratuito, de código aberto e baseado na Web. Ele foi projetado especialmente para o aprendizado de análise de dados introdutória; os alunos usam principalmente a seleção, os controles na tela e o arrastar para realizar suas tarefas, em vez de escrever e executar códigos.

[7] https://www.deepl.com/translator

Em vez de se basear em algoritmos, o ARBOR exige que o usuário faça escolhas sucessivas sobre qual variável dividir, como dividir e quando parar de aumentar a árvore (Erickson e Engel, 2023). Ao usar o ARBOR, os alunos podem compreender o que os algoritmos realizam e, talvez ainda mais importante, compreender a natureza das próprias árvores. O ARBOR não tem algoritmos automatizados que calculam divisões ideais e árvores de tamanho correto. O usuário, por meio de movimentos de arrastar e soltar, decide passo a passo quais variáveis usar para divisões consecutivas, como especificar a divisão e quando encerrar o crescimento da árvore. As taxas de classificação incorreta que avaliam a qualidade da divisão escolhida são imediatamente informadas, o que permite a comparação com divisões alternativas. O objetivo do ARBOR não é a derivação de uma árvore ideal, mas permitir que o usuário explore a flexibilidade do método de árvore e as consequências de várias divisões, para, assim, obter apreciação das árvores como um método automatizado de aprendizagem a partir de dados.

Somente depois de algumas atividades com árvores "cultivadas manualmente", os alunos passam a usar algoritmos e pacotes poderosos para árvores de classificação implementados em R ou Python. Lá, com algum código fornecido pelo instrutor, os alunos deixam o algoritmo construir árvores aplicadas a alguns problemas do mundo real que são ideais de acordo com alguns critérios. Diferentemente de alguns métodos mais avançados de„machine learning“ , as árvores de classificação são transparentes, mas estão atrás de seus concorrentes modernos em termos de eficiência e precisão. No entanto, a partir das árvores de classificação (simples), é apenas um pequeno passo até o conceito de Random Forests. Por meio de problemas concretos e com o fornecimento do código R apropriado, os alunos percebem que os métodos avançados, como o de Random Forests, superam as árvores de classificação simples em relação a alguns critérios externos, como a taxa de classificação correta, mas pelo prêmio de uma regra de decisão completamente opaca.

5 Conclusão

5.1 Motivação, definição do problema e contexto

A análise de dados deve ser motivada por um objetivo. E deve estar inserida em um contexto claro no qual deve ser aplicada e informada. Um bom aplicativo de ciência de dados resolve um problema bem definido ou responde a uma pergunta específica. Esse é o trabalho árduo que precisa ser feito antes de aplicar as ferramentas automatizadas. E é uma das coisas mais difíceis para os alunos aprenderem e internalizarem na escola e na universidade.

5.2 Procedência dos dados e metadados

A análise mais sofisticada não tem valor se for baseada em dados fracos ou questionáveis. O contexto do problema a ser abordado é fundamental para avaliar a relevância e a qualidade necessárias dos dados. A análise de dados não deve ser realizada às cegas, aplicada a dados inadequados ou repletos de erros e lacunas. Os alunos devem aprender a documentar suas fontes de dados e sua origem. E, o que é mais importante, a ser céticos quanto à confiabilidade de seus dados.

5.3 Interação homem-máquina e decisões

A análise deve ser uma colaboração entre analistas humanos e algoritmos de computador, com os algoritmos servindo como ferramentas operadas por humanos. É o analista humano que pode se adaptar às circunstâncias variáveis, reconhecer as limitações do modelo, entender as restrições do conjunto de dados, avaliar e corrigir, excluir ou considerar valores excepcionais e desviantes e entender as possíveis consequências não intencionais de um modelo

que otimiza um critério.

5.4 Ética

Cada vez mais, as consequências éticas da análise da ciência de dados estão sendo reveladas. Não devemos confiar so´em algoritmos e precisamos treinar nossos alunos para pensar e agir de forma ética e aplicar esses princípios em seu trabalho. Os alunos devem aprender a perguntar por que uma análise está sendo realizada e a considerar as consequências éticas da resposta. Embora muitos no campo da ciência de dados vejam os modelos como objetivos e imparciais, O'Neil (p. 21) define os modelos como "opiniões embutidas na matemática". Embora a matemática dê ao modelo a aparência de objetividade, na realidade alguém criou o modelo e decidiu quais dados usar, quais variáveis incluir, qual forma de modelo usar e assim por diante. Um modelo é, na verdade, uma opinião que reflete tanto o viés do modelador quanto o viés dos próprios dados. Aqueles que estão estudando ciência de dados precisam ser sensibilizados para essas questões éticas e treinados para evitar preconceitos e discriminação nos modelos.

5.5 Solução de problemas

Também precisamos ensinar habilidades técnicas, como programação, algoritmos de aprendizado de máquina e tópicos de Big Data. Mas esse não deve ser o foco de um currículo de ciência de dados, assim como o cálculo não deve ser o foco de um currículo de física. Essas são ferramentas, e os alunos devem ser bons nelas, mas primeiro eles precisam aprender por que e como usá-las. A medida final do sucesso é resolver o problema em questão, fornecendo soluções sustentáveis que tenham um impacto tangível.

Os alunos devem aprender logo no início de sua formação que NÃO se trata de ferramentas! As ferramentas de ciência de dados, por mais poderosas que sejam, são um "como", não o "o quê". Em última análise, não se trata de conhecer e usar bem as ferramentas, mas sim de encontrar e usar soluções sustentáveis para problemas difíceis. Caso contrário, não deveríamos nos surpreender se as mentes mais brilhantes que treinamos usassem seu poder cerebral principalmente para incentivar outras pessoas a clicar em determinados anúncios, em vez de usar seu conhecimento para resolver problemas sociais e societários urgentes. Traduzido com www.DeepL.com/Translator (versão gratuita)

Referências

Atenas, J., Havemann, L., & Timmermann, C. (2020). Critical literacies for a datafied society: academic development and curriculum design in higher education. *Research in Learning Technology*. 28: 2468. https://doi.org.10.25304/rlt.v28.2468

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. Teaching Statistics, 43(S1), S133-S142. https://doi.org/https://doi.org/10.1111/test.12279

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: a dialog on data and space. *Big Data and Society*. 3 (1), 1–9. https://doi.org/10.1177/2053951716648346

Engel, J. (2017). Statistical literacy for active citizenship: a call for data science education. *Statistics Education Research Journal 16(1), 44-49* https://doi.org/10.52041/serj.v16i1.213

Erickson, T., & Engel, J. (2023). What goes before the CART. Introducing classification trees

with ARBOR and CODAP. *Teaching Statistics*, 45, S104–S113.

Finzer, W. (2019). Common Online Data Analysis Platform (CODAP). Concord: The Concord Consortium.

Gould, R. (2021). Towards data-scientific thinking. *Teaching Statistics*. 43, 11-22.

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. & Zwitter, A. (2017). Digitale Demokratie oder Datendidaktatur. In: C. Könneker (Ed.), *Unsere digitale Zukunft.* https://doi.org/10.1007/978-3-662-53836-4_1

Lengnink, K., Meyerhöfer, W. & Vohns, A. (2013). Mathematische Bildung als staatsbürgerliche Erziehung? *Der Mathematikunterricht* 59 (4), 2-7.

Messy Data Coalition. (2020). Catalyzing K-12 data education: A coalition statement. https://messydata.org/statement.pdf

OECD (2019). *OECD Skills Outlook: Thriving in a Digital World*. OECD Publishing, Paris. https://doi.org/10.1787/df80bc12-en

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality & Threatens Democracy.* Crown Publishing Group.

Osborne, J. & Pimentel, D. (2019). Science, misinformation, and the role of education. *Science*, Vol 378, Issue 6617, 246-248. https://www.science.org/doi/10.1126/science.abq8093

ProCivicStatPartners (2018). Engaging civic statistics: a call for action and recommendations. A product of the procivicstat project. http://iase-web.org/islp/pcs

Richterich, A. (2018) *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, London. https://doi.org/10.16997/book14

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review* https://doi.org/10.1111/insr.12110/full

Ridgway, J. (2022, Ed.). *Statistics for empowerment and social engagement: teaching Civic Statistics to develop informed citizens*. Springer.

Schield, M. (2004). Information Literacy, Statistical Literacy and Data Literacy. IASSIST Quarterly 28(2), 7-14. https://doi.org/10.29173/iq790

Schüller, K., Koch, H. & Rampelt, F. (2021). Data-Literacy Charta. https://www.stifterverband.org/data-literacy-charter

Shin, T. (2020). Real-life Examples of Discriminating Artificial Intelligence. *Towards Data Science* https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Silva, T. (2022). *Racismo Algorítmico: inteligência artificial e discriminação nas redes digitais.* Sesc Edições SP

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI & SOCIETY*, 35(4), 1077–1078. https://doi.org/10.1007/s00146-020-01054-3

Van Es, K. & Schäfer, M. T. (Eds). (2017) *The Datafied Society: Studying Culture through Data*. Amsterdam University Press. http://library.oapen.org/handle/20.500.12657/31843

Wolfram, C. (2020). *The math(s) fix: An education blueprint for the AI age.* Wolfram Media.