



INOVAÇÃO AO ENSINO DE CIÊNCIAS EXATAS: AVALIAÇÃO DO CURRÍCULO "INTRODUCTION TO DATA SCIENCE"

Luiza de Paula Santos Moura¹

Cedma Ranielly Santos Firminimo²

Karla Patrícia Santos Oliveira Rodríguez Esquerre³

Resumo

Iniciativas que tem como objetivo buscar fomentar ciência e tecnologia principalmente para estudantes do ensino básico da rede pública brasileira são de extrema importância. Neste contexto, a ciência de dados tem sido utilizada como principal ferramenta para a formação científica destes estudantes em iniciativas ao redor do mundo. Uma das mais conhecidas é o Introduction to Data Science (IDS), da Universidade da Califórnia Los Angeles, que utiliza, como ferramenta de apoio, a programação em R. Neste artigo será revisado o seu currículo juntamente com a viabilidade de sua aplicação para estudantes brasileiros, bem como o que é proposto quanto a introdução ao R. Isso é feito a partir do estudo íntegro do documento bem como a avaliação dele a partir dos seus dados textuais. O currículo, apesar de ser bom, precisa ser adaptado para se adequar à Base Nacional Comum Curricular.

Palavras-chave: Ciência de Dados, Práticas educativas, Programação em R, Currículo.

Abstract

Initiatives that aim to promote science and technology, especially for elementary school students in the Brazilian public school system, are extremely important. In this context, data science has been used as the main tool for the scientific training of these students in initiatives around the world. One of the best known is the Introduction to Data Science (IDS), from the University of California Los Angeles, which uses programming in R as a support tool. This article will review its curriculum together with the viability of its application for Brazilian students, as well as what is proposed regarding the introduction to R. This is done by studying the document in its entirety and evaluating it based on its textual data. Although the curriculum is good, it needs to be adapted to suit the Common National Curriculum Base.

Keywords: Educational practices, Introduction to R programming, curriculum.

Introdução

As consequências de se viver em uma sociedade globalizada e tecnológica estão presentes nas ações e pensamentos dos estudantes que estão frequentando as escolas brasileiras. Como a própria matemática está formalizada por um exercício essencialmente humano, ela está sujeita a ter sua prática alinhada inerentemente a aspectos políticos, repleta de questões de dominação e poder, como comumente estão submetidas as

¹ Universidade Federal da Bahia, luiza.paula@ufba.br

² Universidade Federal da Bahia, cedma.santos@gmail.com

³ Universidade Federal da Bahia, karlaesquerre@ufba.br



dinâmicas da humanidade. Deste modo, é importante reconhecer que o ensino e a aprendizagem da matemática não são atividades politicamente neutras e que é preciso transformá-las com práticas socialmente mais justas (Gutiérrez, 2013). Não dar atenção à identidade e ao poder leva ao fenômeno em que a matemática não seja um domínio universal, de forma que não se atinjam os potenciais de justiça social e equidade.

O uso de ciência de dados, ainda em idade escolar, possibilita ao estudante reconhecer a relevância da matemática, à medida que realiza as suas próprias indagações, investigações e descobertas por meio da busca de padrões ou de tendências. Oportunizar este tipo de percepção aos estudantes, a fim de que eles possam avaliar estatísticas publicadas nos veículos midiáticos, principalmente as sobre a esfera pública, torna-se um desafio quando não há base sólida que inclui conhecimentos em matemática básica, estatística e familiaridade com a representação visual ou numérica de dados (Engel, 2017). Levando isso em consideração, são identificadas algumas iniciativas que realizam a concepção de um currículo que pode ser aplicado para ensinar, de forma lúdica, ciência de dados a jovens estudantes em idade escolar. Uma delas é o *Introduction to Data Science* (IDS), aplicado nos Estados Unidos. Ele foi desenvolvido na Universidade da Califórnia em Los Angeles (UCLA) e tem como autor principal Robert Gould, professor no departamento de estatística na mesma instituição. Para aplicá-lo, a instituição deve fazer uma parceria com a equipe do IDS com a duração de 2 anos.

Objetivo

É objetivo deste artigo realizar a revisão do material didático fornecido pela UCLA sobre o currículo do curso de Introdução à Ciência de Dados em R (*Introduction to Data Science - IDS*), com a descrição das práticas para que professores possam aplicá-lo.

Material e Método

O material do IDS é disponibilizado abertamente em um livro orientador. Neste artigo, as instruções são analisadas tomando partida de que professor, assim como os estudantes, não conheciam previamente como programar em R. Também há práticas voltadas a conteúdo sobre estatística e probabilidade, uma área temática presente na matemática prevista na Base Nacional Comum Curricular (BNCC) para os ensinos fundamental e médio. Será verificado se as descrições contidas no material são compatíveis com as habilidades na área relativa da estatística e probabilidade, no caso do ensino fundamental I e II, e as relativas à área de matemática e suas tecnologias, no caso do ensino médio. Para isso, foi realizado um levantamento das palavras mais frequentes de ambos os textos, utilizando algoritmos de mineração de texto, a fim de conferir a compatibilidade de aplicação do material da iniciativa quanto ao currículo brasileiro.

Resultados e Discussão

De uma forma geral, pode-se retratar o foco e as habilidades levadas em consideração pelo currículo IDS por meio da visualização das palavras mais frequentes no documento. Para isso, foi realizada uma mineração de texto do documento orientador em formato PDF. Além das palavras comuns em materiais didáticos como “class” e “students”, há a presença forte de palavras que são relativas a práticas investigativas como “survey” e “explain” e outras relativas à ciência de dados como “model” e “prediction”. Já no vocabulário da BNCC, também há a presença de palavras que remetem à investigação, como “pesquisa” e “resolver”, bem como as envolvendo ciência de dados como “dados”, “tabelas” e “gráficos”.



Figura 1 - Palavras mais frequentes no documento orientador do IDS, à esquerda, e nas habilidades previstas na BNCC, à direita.

Fonte: Elaboração Própria, 2024.

É importante ressaltar que apesar de haver alguma semelhança relacionadas a vocabulário, a BNCC e o currículo do IDS não abordam os mesmos conteúdos de forma integral, o que pode ser percebido a partir da leitura completa dos documentos e que há limitações em termos semânticos dos vocábulos no método usado neste artigo.

Nas orientações do IDS ao professor estão descritos os objetivos do encontro, os materiais necessários, o vocabulário que deve ser adquirido e descrição da atividade. Também são dispostos os questionamentos que devem ser feitos aos estudantes junto com respostas que podem surgir. Este currículo é dividido em 4 unidades. Cada unidade possui diferentes eixos temáticos, onde há a proposição de diferentes atividades que estão de acordo com o objetivo de cada encontro. Eles estão dispostos na tabela 1:

Tabela 1 – Conteúdos abordados no currículo do Introduction to Data Data Science.

Unidade	Eixo Temático	Conteúdo abordado
Unidade 1	<i>Data Are All Around</i>	Explorando o mundo com dados, organização de dados, ciclo de dados, variabilidade
	<i>Visualizing Data</i>	<i>Dotplot</i> , histograma, distribuição, introdução a R, tipos de variáveis, gráfico de barras, importando dados no R
	<i>Would You Look at the</i>	Limpeza de dados, tabelas de frequência



Time?

Unidade 2	<i>What Is Your True Color?</i>	Agrupamentos, medidas de tendência central e de dispersão, <i>boxplot</i> , estatística descritiva no R
	<i>How Likely Is It</i>	Simulações no R, probabilidade
	<i>Are You Stressing or Chilling?</i>	Leitura crítica de dados, agrupamento de dados no R.
	<i>What's Normal?</i>	Distribuição normal, desvio padrão, <i>z-score</i> , simulação no R

Unidade 3	<i>Testing, Testing...1, 2, 3...</i>	Leitura Crítica, experimentos, correlação e causalidade
	<i>Would You Look at That?</i>	Estudos observacionais e experimentais, fatores de confusão
	<i>Are You Asking Me?</i>	Amostragem intervalos de confiança, bootstrapping
	<i>What's the Trigger?</i>	Coleta de dados, perguntas de pesquisa
	<i>Web Pages</i>	Dados na internet, extração de dados na internet

Unidade 4	<i>Predictions and Models</i>	Desvio padrão e erro médio absoluto, previsão de dados univariados e bivariados, ajuste de uma reta, regressão linear simples, coeficiente de correlação
	<i>Piecing it Together</i>	Regressão linear múltipla, regressão não-linear
	<i>The Growth of Landfills</i>	Modelagem de problemas reais, ajuste de modelos
	<i>Decisions, Decisions</i>	Árvores de decisão, treino e teste, classificação
	<i>Ties that Bind</i>	<i>Clustering, k-means</i>

Fonte: GOULD et al, 2023

Os encontros, no currículo, podem ser de uma das três categorias: *lesson*, *lab* e *practicum*. Os encontros *lesson* possuem como objetivo discutir conceitualmente conteúdos de estatística e a ciência de dados com exercícios que envolvem o direcionamento dos estudantes por meio de práticas ilustrativas. O *lab* é destinado a encontros em que o estudante discute e faz a prática de utilizar o aprendizado obtido anteriormente nos encontros da categoria *lesson* com análise de dados usando a programação com R. Já a *practicum*, possui objetivo de que o estudante produza um projeto de análise de dados utilizando os conceitos que aprenderam em cada unidade usando bancos de dados que ou são produzidos durante a unidade com informações concebidas pelos próprios estudantes, ou que já foram utilizados em algum momento na unidade.

Em um primeiro estudo dos estudantes usando R é feita uma exploração dos painéis no *Rstudio*. Neste mesmo encontro são introduzidos também alguns comandos para manipulação e visualização de *data frames*. Eles são feitos diretamente no console, o que,



dependendo da situação, pode dificultar ou facilitar o aprendizado do estudante. Vale lembrar que o *RScript* foi introduzido somente na unidade 2, em um momento que as funções se tornam muito complexas e não podem ser executadas facilmente no consol. No material é explicado como deve ser a sintaxe do código de maneira adequada.

Algo a pontuar sobre o ensino introdutório da programação em R no currículo é a falta de utilização de funções provenientes de bibliotecas, o que, em um momento inicial, pode facilitar, mas limita em termos do que a utilização de biblioteca como, por exemplo, a *dplyr*, proporciona ao simplificar a lógica do código quando comparado a formular usando apenas as funções nativas. O mesmo comentário pode ser feito nas próximas práticas de confecção de figuras para visualização de dados.

Conclusão⁴

O currículo do *IDS*, de maneira geral, oferece práticas ilustrativas e envolventes em se tratando do aprendizado de estatística e ciência de dados e da própria programação em R. A forma de dispor as informações no livro orientador em geral é boa, pois possui além dos procedimentos, discussões que podem ser feitas além das respostas recorrentes de uma pergunta ou discussão. Quanto à aplicabilidade do currículo ao estudante brasileiro, é necessário que haja adaptações para estar de acordo com a área da estatística e probabilidade dos ensinos fundamental e médio ou que as atividades sejam feitas de forma extracurricular já que não está de acordo de forma íntegra com o previsto pela BNCC.

Referências

- GUTIÉRREZ, Rochelle. **The Sociopolitical Turn in Mathematics Education**. v. 44, n. 1, p. 37-68. Journal for Research in Mathematics Education. Jan 2013. <<https://doi.org/10.5951/jresematheduc.44.1.0037>>
- ENGEL, Joachim. **Statistical literacy for active citizenship: A call for data science education**. v. 16, n. 1, p. 44-49. Statistics Education Research Journal. May 2017. <<https://doi.org/10.52041/serj.v16i1.213>>
- BRASIL. Ministério da Educação. **Base Nacional Comum Curricular**. Brasília, 2018. <<http://download.basenacionalcomum.mec.gov.br/>>
- GOULD, Robert; MACHADO, Suyen; JOHNSON, Terri Ana; MOLYNEUX, James. Introduction to Data Science. v. 6. 400 p. Mar 2023. <<https://www.mobilizingcs.org/>>
- OOMS, J. (2023). *_pdfutils: Text Extraction, Rendering and Converting of PDF Documents_*. R package version 3.4.0, <<https://cran.r-project.org/web/packages/pdfutils/index.html>>.
- QUEIROZ, G.; FAY, C.; HVITFELDT, E.; KEYES, O.; MISRA, K.; MASTNY T.; ERICKSON, J.; ROBINSON, D.; SILGE, J. (2024). *_tidytext: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools_*. R package version 0.4.2, <<https://cran.r-project.org/web/packages/tidytext/index.html>>.
- FELLOWS, I. (2022). *_wordcloud: Word Clouds_*. R package version 2.6, <<https://cran.r-project.org/web/packages/wordcloud/index.html>>.
- FEINERER, I.; HORNIK, K. (2024). *tm: Text Mining Package_*. R package version 0.7-13, <<https://cran.r-project.org/web/packages/tm/index.html>>

Anexo

O código R está disponível em: <https://github.com/luizapaulamoura/ser_viii_luiza>

⁴ Agradecimentos: Os autores agradecem o apoio financeiro do CNPq|PIBIC, Projeto Ciência de Dados e Engenharia: Formação Científica Ética e Justiça Social, do CNPq|MCTIC, Proc. 409481/2021-1, Práticas Investigativas, Ciência de Dados e Inteligência Artificial na Educação Pública e da Fundação Itaú para Educação e Cultura, Projeto Ciência de Dados na Educação Pública.