Anais dos Trabalhos de Conclusão de Curso
Pós-Graduação em Computação Aplicada à Educação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Versão Preliminar

A importância da expectativa para a adoção de currículos baseados em competências em cursos livres de ciência de dados

Ivan Ramos Pagnossin[1], S. Isotani[2], B. E. Penteado[3]

Resumo

Apresentamos um estudo sobre a aprendizagem de ciencia de dados em cursos li-vres. Mostramos que há uma diferença mensurável na aprendizagem de componentes relacionadas com ferramentas, quando comparada com a aprendizagem de componentes abstratas, como princípios e métodos. Especulamos que essa diferença tenha origem na expectativa do aluno em conseguir um posto de trabalho, cujos requisitos relacionam-se explicitamente com as componentes concretas. Argumentamos que esse fenômeno dificulta a adoção de currículos baseados em habilidades e competem- cias e demonstramos que a satisfação do aluno com o curso, a relevância percebida por ele sobre os tópicos abordados e o ritmo sobre sua aprendizagem contribuem com no máximo aproximadamente 20% da aprendizagem.

Abstract

This work presents a study about learning in data science courses. We present a meaningful difference exists between the learning of concrete topics, usually related to tools, when compared to the learning of abstract topics, like methods and principles. We speculate this difference is due to the student's expectancy of getting a working position as a data scientist, whose requirements explicitly match concrete topics. We argue this phenomenon hinders the adoption of competency-based curricula. Finally, we show that students' satisfaction, perceived relevance of each topic and perceived rhythm account for, at most, approximately 20% of students learning.

1 Pós-graduando em ComputaçãoAplicada a` Educação, USP, irpagnossin@usp.br
2 Seiji Isotani, USP, sisotani@icmc.usp.br
3 Bruno Elias Penteado, USP, brunopenteado@usp.br

Cite as: Pagnossin, I. & Isotani, I. & Penteado, B . (2020). A importância da expectativa para a adoção de currículos baseados em competências em cursos livres de ciência de dados. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

1. Introdução

O interesse pela ciência de dados tem crescido nos últimos anos, assim como a quantidade de vagas de trabalho nessa área. Consequentemente, a demanda e a oferta por cursos nessa área seguem a mesma tendência. Porém, a definição de um currículo global de ciência de dados ainda não existe, haja vista a interdisciplinaridade dessa área.

Concomitantemente, a adoção da Educação baseada em competências como meio para desenvolver os quatro pilares da Educação do século XXI (aprender a conhecer, aprender a fazer, aprender a conviver e aprender a ser) e também um fenômeno global. Nesse contexto, a *International Association of Business Analytics Certification* (IABAC) tem construído um arcabouço unificado e globalizado de conhecimentos, habilidades e competências para a atuação do cientista de dados.

Apesar disso, o discurso das empresas que procuram cientistas de dados no Brasil guarda uma relação explícita com conhecimentos apenas, em especial sobre ferramentas e algoritmos, sugerindo um distanciamento de qualquer proposta baseada em competências. Isso pode dificultar, especialmente em cursos livres, o foco nas competências, pois o candidato a cientista de dados, ao confrontar os requisitos dessas vagas com as ofertas de cursos, vê mais claramente a relação com aqueles cursos que focam sua comunicação (propaganda) nesses conhecimentos.

Estamos particularmente interessados no contexto dos cursos livres, pois e nele que o primeiro autor esta inserido profissionalmente. Assim, tomamos como problema de pesquisa a adoção da aprendizagem baseada em competências em cursos livres de ciência de dados. O objetivo geral e viabilizar essa empreitada e, para isso,
começamos propondo um objetivo específico: averiguar se o fenômeno explicado no parágrafo anterior de fato existe.

Para isso, propusemos três hipóteses: a primeira delas afirma que a aprendizagem dos alunos em componentes explicitamente relacionadas com ferramentas e maior do que em componentes mais abstratas. A segunda hipótese e similar, mas substitui “ferramentas” por “algoritmos”.

A ideia subjacente a essas hipóteses e que as componentes abstratas, que são igualmente importantes para a composição de habilidades, são menos relevantes aos olhos dos alunos devido `a ênfase que as ferramentas e algoritmos tem em anúncios de vagas de trabalho.

Finalmente, a terceira hipótese afirma que podemos utilizar a relevância de cada componente, conforme percebida pelo aluno, o ritmo das aulas e a satisfação geral do aluno com o curso como preditores da aprendizagem. Ou seja, podemos
prever a aprendizagem com base nesses parâmetros. A ideia aqui e que, se ao menos uma das duas primeiras hipóteses for verdadeira, podemos esperar alguma correlação com a relevância de cada componente, conforme percebida pelo aluno, e avaliar sua importância em comparação com o ritmo do curso e a satisfação geral do aluno. Isso porque, mesmo que haja o efeito esperado, sua magnitude pode ser desprezível quando comparada com outros fatores.

Nos testamos essas hipóteses por meio de análise quantitativa sobre as respostas a pesquisas de levantamento apresentados aos alunos no final de cada aula. Começamos aprofundando o cenário apresentado nos parágrafos anteriores: primeiramente, abordamos a evolução da ciência de dados e da demanda por postos de trabalho e cursos nessa área (Seção 2). Em seguida, voltamos nossa atenção para Educação baseada em competências e sua influência em propostas globalizadas de currículos de ciência de dados (Seção 3). Depois, apresentamos detalhes do cenário de aplicação deste trabalho e a motivação para ele (Seção 4). Seguimos com as referências teóricas (Seção 5), detalhes da metodologia de desenvolvimento deste trabalho e análise (Seção 6) para, então, apresentarmos e discutirmos os resultados (Seção 7). Concluímos (Seção 8) resumindo os resultados, enfatizando as limitações a apresentando possíveis prosseguimentos.

2. Ciência de dados

De acordo com o *National Institute of Standards and Technology* (NIST), ciência de dados (*data science*, DS) refere-se `a atividade de extrair conhecimento acionável de conjuntos de dados brutos utilizando processos de exploração ou formulação e teste de hipóteses [NIST Big Data Public Working Group 2015, p. 7]. Ela incorpora princípios, tcênicas e mtodos de diversas áreas, como ciências da computação, matemática e estatística, além de domínio da área de aplicação.

Essa área tem ganhado notoriedade desde o início deste século devido ao crescimento exponencial na geração de dados, conhecido genericamente como *Big Data*, devido principalmente ao advento da Web 2.0, por volta de 2005, e dos dispositivos móveis, em 2007. Desde então e também devido ao aumento na capacidade computacional, novas tcênicas de análise, a maioria computacionais, tem sido empregue. Exemplos notáveis são as redes neurais e a inferência bayesiana.

Nos últimos anos, inúmeras ferramentas robustas de computação e análise de dados, disponibilizadas livremente, tem tornado a ciência de dados cada vez mais acessível [Hayes 2019]. Exemplos são as linguagens de programação Scala [Bugnion 2016], Python [Nagpal e Gabrani 2019], R [James et al. 2013] e, mais recentemente, Julia [McNicholas e Tait 2019]. Isso sem mencionar a abordagem Au-toML, que oferece interfaces simples para o emprego de aprendizagem de máquina sem a necessidade de programação [He, Zhao e Chu 2020].

A facilidade de acesso e a alta demanda por cientistas de dados levou, então, `a oferta de cursos de ciência de dados [Hassan e Liu 2019]: no Brasil, algumas universidades, como a Univesp, a USP e a FIAP, já oferecem cursos de graduação e pós-graduação. Mas também e possível estudar gratuitamente pela Internet, nas plataformas MOOC (*Massive Open Online Courses*) como Coursera (coursera.org), EdX (edx.org) *etc*.

Ha também cursos livres nessa área que ganham força devido ao reconhecimento de que as universidades tradicionais já não suprem a necessidade de mão de obra qualificada em várias áreas [Zulauf 2006]. Esses cursos oferecem uma formação rápida, que varia em torno de um semestre, e visam colocar o candidato no mercado de trabalho independentemente da sua área de formação. Consulte, por exemplo, os

Figura 2.1. Anúncios de cursos livres de ciência de dados.

cursos de *Data Analytics* e *Data Science* da Digital House (digitalhouse.com/br) e da Tera (somostera.com).

Porém, como esses cursos visam o mercado de trabalho, em geral a divulgação deles enfatiza não os conhecimentos, competências e habilidades que um cientista de dados precisa, mas sim as ferramentas e algoritmos utilizados por eles e que são frequentemente mencionados em vagas de emprego.

A Figura 2.1 ilustra isso: ela apresenta a propaganda de dois cursos, oferecidos na rede social Instagram. Em ambos o foco e linguagens de programação (Python, Scala e R), ferramentas de *Big Data* (Spark, Hadoop *etc*), aplicativos (Tableau) e bibliotecas (Pandas, SciPy *etc*).

Embora o domínio dessas ferramentas seja necessário, ele não e suficiente para um cientista de dados de fato produzir conhecimento acionável. Para isso são necessários, por exemplo, mtodos de pesquisa, engenharia de dados, inferência estatística, dentre outros [CF-DS-Release 2019, p. 15].

> Argumentamos, então, que a ênfase em ferramentas, linguagens *etc* na *divulgação* desses cursos esmorece, aos olhos dos alunos, a importância das habilidades e competências, haja vista que sua relação com as vagas de emprego não são tão explícitas como as linguagens e ferramentas.

3. Competências e habilidades

No Brasil e no mundo, a proposta de substituir os currículos tradicionalmente baseados em conteúdo por aqueles baseados em competências e habilidades tem ganhado força. Por exemplo, no Brasil a Base Nacional Comum Curricular (BNCC) [Base Nacional Comum Curricular], homologada em 2018, “estabelece conhecimentos, competências e habilidades que se espera que todos os estudantes desenvolvam ao longo da escolaridade básica”.

Essa abordagem tem sido adotada no ensino básico [Marques 2017] e superior. De fato, ela visa a empregabilidade [Butova 2015], que é também o objetivo dos cursos livres anteriormente mencionados.

Perfil/Conhecimentos Necessários:

- Raciocínio lógico
- Lógica de Programação
- Ter conhecimento em Ferramentas de Big Data (**Hadoop, Hive, Spark, R, Python, etc.**)
- Boa Comunicação
- Bom Relacionamento

LOCAL: São Paulo - Zona Sul

Requisitos

- Ambiente Unix;
- Experiência com Python, Scala, R ou Java;
- Processamento de dados distribuídos (Spark ou Hadoop);
- AWS com S3, EC2, EMR, Redshift, DynamoDB, Kinesis;
- Conhecimentos analíticos, estatísticos e de linguagens de programação;
- Disponibilidade para viagens, pois irá atuar com o time técnico de São Francisco e Singapura e eventualmente será necessário viajar para outras unidades;
- Inglês Fluente, diferencial para espanhol;
- Curso superior na área de tecnologia/ exatas.

Habilidades Obrigatórias

- Python
- AWS

Habilidades Desejáveis

- Hadoop
- Spark

Figura 3.2. Duas vagas de cientista de dados, extraídas do LinkedIn.

Por exemplo, segundo o *EDISON Data Science Framework*, algumas competências de um cientista de dados são [CF-DS-Release 2019]:

> "Usar de engenharia (geral e de software) para pesquisar, projetar, desenvolver e implementar novos instrumentos e aplicações para a coleta de dados, armazenamento, análise e visualização".
>
> "Utilizar eficientemente uma variedade de técnicas de análise de dados, como aprendizagem de máquina, mineração de dados, analises prescritivas e preditivas, para análise complexa de dados durante todo o ciclo de vida dos dados"
>
> Ademais, exemplos de habilidades são:
>
> "Usar aprendizagem de máquina, tecnologia, algoritmos e ferramentas" "Projetar experimentos, desenvolver e implementar processos de coleta de dados"

Entretanto, as empresas frequentemente listam ferramentas e algoritmos como requisitos as` vagas de trabalho que publicam, ao invés das competências e habilidades necessárias para o exercício da ciência de dados. A Figura 3.2 ilustra duas vagas dessa Rea, a obtidas na plataforma LinkedIn. Nela podemos ver a menção as` mesmas ferramentas e algoritmos oferecidos pelos cursos livres.

4. Cenário e motivação

Em vista do exposto, o autor deste trabalho, que atua como coordenador na instituição educacional Digital House (digitalhouse.com/br), que por sua vez oferece cursos livres de DS e *Data Analytics* (DA), gostaria de adotar a aprendizagem baseada em competências nos cursos de ciência de dados e de *Data Analytics* (DA) sob sua responsabilidade.

Neste trabalho nos analisamos os resultados de aprendizagem nos cursos de DA e DS oferecidos pela instituição educacional mencionada, com o intuito de identificar uma possível concorrência entre os currículos baseados em competências e os incentivos do mercado de trabalho, conforme exposto nas hipóteses deste trabalho (Seção 7.1).

Conforme a definição de ciência de dados do NIST, os cursos de DA e DS

Tabela 4.1. Comparação dos cursos de DA e DS.

Curso	Objetivo	Cargo horária (horas)	Curacao (semanas)	Ferramenta
DA	Inteligência de mercado	140	14	Aplicativos
DS	Produto de dados	196	19	Linguagem de programação

podem ser classificados como de ciência de dados. Porém, eles guardam semelhanças e diferenças entre si:

> ***Data Analytics* (DA):** visa a inteligência de mercado (*Business Intelligence*, BI), isto e, a aplicação da ciência de dados para obter conhecimento acionável que suporte decisões estratégicas para um empreendimento. Esse curso tem carga horária de 140 horas e dura 14 semanas. Outra característica desse curso e que ele se baseia em aplicativos como Power BI, Tableau, MySQL *etc*.
>
> ***Data Science* (DS):** visa o desenvolvimento de "produtos de dados", isto e, *softwares* que automaticamente obtém conhecimento acionável para oferecê-lo aos clientes. Esse curso tem carga horária de 196 horas, dura 19 semanas e baseia-se na linguagem de programação Python e suas extensões para análise de dados.

A Tabela 4.1 resume o que foi exposto imediatamente acima, comparando os dois cursos. Note que, atualmente, os cursos em questão *não* são baseados em competências. De fato, esse e nosso objetivo geral.

5. Referencial teórico

Nesta seção apresentamos as referências teóricas que suportam o desenvolvimento deste trabalho: começamos abordando a aprendizagem baseada em competências (Seção 5.1), com o intuito de justificar sua escolha como base para propostas de cursos de ciência de dados. Em seguida, tratamos da teoria do valor da expectativa (Seção 5.2), que utilizamos para conjecturar a razão pela qual observamos os resultados deste trabalho. Finalmente, mencionamos a aprendizagem significativa (Seção 5.3) pois, conforme nossas conclusões, pode ser utilizada para intervir num dos resultados observados.

5.1. Aprendizagem baseada em competências

Uma competência e uma coleção de habilidades e conhecimentos para realizar uma tarefa. Alternativamente, a BNCC [Base Nacional Comum Curricular] define competência como "a mobilização de conhecimentos (conceitos e procedimentos), habilidades (praticas, cognitivas e socioemocionais), atitudes e valores para resolver demandas complexas da vida cotidiana, do pleno exercício da cidadania e do mundo do trabalho".

A aprendizagem baseada em competências e aquela que define objetivos de aprendizagem em termos dessas competências, que devem ser mensuráveis. Realmente, a "educação baseada em competências atualmente foca em resultados de aprendizagem e aborda o que os alunos devem aprender a *fazer*" [Butova 2015] (tra-

dução e grifo nossos).

Segundo [Butova 2015], a ideia original surgiu em 1965 com o filósofo e linguista americano Noam Chomsky, que enfatizou, no contexto linguístico, a diferença entre conhecer um idioma e saber aplicá-lo. Mais tarde essa diferença foi extrapolada para outras áreas, como a pedagogia e a filosofia.

No Brasil a adoção dessa metodologia no Ensino Básico foi oficializada por meio da BNCC, homologada em 2018, que define os conhecimentos, competências e habilidades mínimos que todos os cidadãos brasileiros devem desenvolver antes do Ensino Superior.

Embora a BNCC limite-se a` Educação Básica, existe a preocupação de utilizar também essa abordagem no Ensino Superior, particularmente em resposta ao mercado de trabalho. Mais do que isso, a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco) defendem que a Educação para o século XXI deve desenvolver "competências globais"; os quatro pilares da Educação moderna:
(1) aprender a conhecer, (2) aprender a fazer, (3) aprender a conviver e (4) aprender a ser.

De fato, a Educação baseada em competências recorrentemente citada como um meio para resolver a crescente disparidade entre as necessidades do mercado de trabalho e o que as universidades oferecem aos seus estudantes [Zulauf 2006]:

> "[No] mercado e os ambientes de trabalho, o ensino superior vem sofrendo crescente pressão para desenvolver a empregabilidade dos estudantes e tornar-se mais relevante no que diz respeito as` necessidades dos empregadores."

Concomitantemente a esse movimento, inúmeras iniciativas *ad hoc* de criar um currículo para o ensino da ciência de dados, tais como [Hassan e Liu 2019], [Anderson et al. 2014] e [Cheng e VanDeGrift 2019] tem surgido; e começam a aparecer iniciativas de unificação [Raj et al. 2019].

Uma das que merece menção e a *EDISON Data Science Framework*, desenvolvida pela IABAC, que prove uma base para a definição da profissão de cientista de dados, bem como componentes relacionados como educação, treinamento, papéis, dentre outros. Ela define três especificações:

Data Science Competence Framework (CF-DS) e o núcleo da especificação, que inclui as competências necessárias para o cientista de dados atuar no mercado de trabalho e na academia ao longo de toda sua carreira.

Data Science Body of Knowledge (DS-BoK) define áreas de conhecimento para a construção de currículos de ciência de dados que comportem as competências identificadas no CF-DS [Demchenko, Belloum e Wiktorski 2017].

Data Science Model Curriculum (MC-DS) define objetivos de aprendizagem consonantes com a CF-DS e unidades de aprendizagem associadas `as unidades de conhecimento definidas no DS-BoK.

Assim, defendemos que os currículos de cursos de ciência de dados devam ser baseados em competências e habilidades. Porém, o mercado impõe dificuldade

a essa empreitada pois, ao enfatizar ferramentas nas vagas de emprego, incentiva o candidato a cientista de dados a procurar cursos que lhe deem esse *conhecimento*. Procuramos fundamentar essa suposição com base na teoria do valor da expectativa, na próxima seção.

5.2. Teoria do valor da expectativa

Em 1964, Victor H. Vroom desenvolveu a sua teoria comportamentalista do valor da expectativa [Petri], uma teoria da motivação, segundo a qual "a escolha, persistência e desempenho de indivíduos pode ser explicada por sua crença sobre quão bem ele executará uma atividade e quanto ele valoriza essa atividade". A teoria propõe ainda que a motivação depende de três fatores:

Resultado ou recompensa esperados, chamado de valência;

Percepção de intensidade da relação entre o desempenho requerido e o resultado (instrumentalidade);

Percepção do vínculo existente entre o esforço requerido e o desempenho subsequente (expectativa).

Conforme demonstraremos nos resultados deste trabalho (Seção 7), a aprendizagem percebida pelo aluno em aulas explicitamente relacionadas com ferramentas

e maior, dentro de um nível de significância de 95%, que àquela em outras aulas. Aplicamos a teoria do valor da expectativa para conjecturar que essa diferença e devida a uma maior intensidade do vínculo entre as aulas de ferramentas e as vagas de emprego (valência), levando a motivação intrínseca que promove a aprendizagem nessas aulas.

Além disso, opondo verdadeira essa conjectura, podemos ainda utilizá-la para promover a aprendizagem nas demais aulas. Para isso, propomos (1) intensificar, pelo discurso, o vínculo entre os requisitos do mercado e os objetivos das aulas, e/ou

(2) desenvolver essas aulas utilizando ferramentas, se possível, o que se baseia na aprendizagem significativa, objeto da próxima seção.

5.3. Aprendizagem significativa

Aprendizagem significativa [Pelizzari et al. 2002] e a concepção cognitivista de ensino e aprendizagem proposta pelo psicólogo americano David Ausubel em 1963. O autor afirma que o fator isolado mais relevante para a aprendizagem e o conhecimento prévio do aluno. Ou seja, a aprendizagem ocorre quando novas informações se ancoram em conceitos ou proposições relevantes pré-existentes.

Desse modo, podemos ancorar a aprendizagem de componentes abstratas, isto e, que tem um vínculo fraco com a valência, na aprendizagem das componentes cujo vínculo e forte (nos quais verificamos maior aprendizagem percebida).

6. Metodologia

Nos utilizamos mineração de dados educacionais [Baker, Isotani e Carvalho 2011] para realizar analises quantitativas sobre os dados gerados pela interação contínua dos alunos com os sistemas de informação da instituição educacional. Ou seja, os

dados aqui utilizados não foram produzidos em resposta a uma pesquisa previamente planejada.

As analises foram realizadas utilizando *softwares* gratuitos: Python 3.8 com a interface JupyterLab e as bibliotecas statsmodel (modelagem estatística), scikit-learn (*machine learning*), dentre outras.

A análise e resultados deste trabalho foram desenvolvidos a partir de um conjunto de dados armazenado num arquivo no formato CSV (*comma-separated values*). Esse arquivo, por sua vez, foi construído aglomerando as respostas dos alunos a inúmeros formulários (Google Forms), um por aula, turma e curso, utilizando a linguagem de *script* Google Apps Script. Os dados originais contem identificação dos alunos e, por isso, o arquivo disponibilizado para reprodutibilidade deste trabalho (apêndice A) passou por uma etapa em que apenas removemos a identificação dos alunos, trocando por "Aluno 1", "Aluno 2" *etc*.

6.1. Formulários

No final de cada uma das aulas dos cursos de DA e DS, os professores divulgaram aos alunos, diretamente e pelo canal de comunicações da turma (grupo do WhatsApp), o link para um formulário do Google Forms (único para aquela aula). Os alunos eram incentivados, mas não obrigados, a responder. Cada formulário apresentava as seguintes questões:

1. Email
2. Nome
3. "Area: quanto voce ja sabia sobre topico?". Escala Likert de 1 a 5. Chamemos de $q$antes para referência futura.
4. "Area: E agora depois da aula quanto voce sabe sobre topico?". Escala Likert de 1 a 5. Chamemos de q depois.
5. "Relevância: O quanto você acha que o conteúdo abordado e relevante para a sua formação? ". Escala Likert de 1 ("pouco relevante") a 5 ("muito relevante").
6. "Ritmo: Como você classifica o ritmo da aula de hoje?". Escala Likert de 1 ("muito lento") a 5 ("muito rápido").
7. "Satisfação: O quanto você está satisfeito@ com a aula de hoje?". Escala Likert de 1 ("pouco satisfeito@") a 10 ("muito satisfeito@")
8. "Se quiser fornecer algum feedback ou comentário específico use o espaço abaixo (não obrigatório)". Discursiva. Não utilizada neste trabalho.

As questões 3 e 4 podem ser repetidas até 4 vezes (ou seja, cada aula pode apresentar até 4 tópicos), em que Area representa a microcomponente do curso (tecnologia, negócios ou estatística) e tópio e um texto escrito livremente pelo professor, que aborda um tópio da aula. Exemplos de tópio são: "Aplicabilidade da regressão logística", "Fundamentos do MySQL" *etc.* (ao todo são 985 valores distintos no conjunto de dados original).

6.2. Variáveis

A variável-alvo (dependente) deste trabalho e a aprendizagem, cuja medida operacional a foi definida como o valor numérico da questão 4 subtraído do da questão

3:

$$a := d_{\text{depois}} - q_{\text{antes}} \quad (1)$$

Por exemplo, se o aluno respondeu $q_{\text{antes}} = 2$ para a questão 3 (quanto "sabe" antes da aula) e $d_{\text{depois}} = 5$ para a questão 4 (depois), então a aprendizagem foi de $5 — 2 = 3$. Desse modo, essa variável assume valores inteiros no intervalo $[-4, 4]$ e, conforme [Harpe 2015], pode ser interpretada como numérica.

As variáveis independentes são:

Relevância (questão 5). Variável categórica com cardinalidade 5.
Ritmo (6). Variável categórica com cardinalidade 5.
Satisfação (7). Variável numérica discreta.
topico . Variável categórica (texto) com cardinalidade 985.

A natureza de cada variável acima (numérica ou categórica) foi decidida com base nas recomendações de [Harpe 2015].

6.3. Coleta dos dados

Os formulários foram apresentados entre 8/abr e 7/dez/2019 para 6 turmas de DA e 4 turmas de DS, totalizando até 255 alunos e gerando 1878 observações para DA e 1763 para DS.

7. Resultados e discussão

Antes de começarmos a apresentação dos resultados, análise e discussão, revisitamos e detalhamos as hipóteses propostas (Seção 7.1). Em seguida discutimos o pré-processamento dos dados e alguns cuidados tomados (Seção 7.2). Finalmente, analisamos e discutimos os resultados associados a cada uma das hipóteses, em sequência (Seções 7.3 a 7.5).

7.1. Hipóteses

As hipóteses propostas para este trabalho foram:

Hipótese 1: a aprendizagem dos alunos de DA e DS e maior nas aulas que abordam explicitamente as ferramentas (*e.g.*, Google Analytics para DA e Python para DS), quando comparadas as aulas mais abstratas, sobre princípios, tcênicas e mtodos (*e.g.*, arquitetura de dados para DA e princípio de funcionamento dos algoritmos de agrupamento para DS).

Hipótese 2: a aprendizagem dos alunos de DS (não de DA) e maior nas aulas que abordam algoritmos explicitamente (*e.g.*, "MeanShift e DBSCAN") do que naquelas com ttópicos mais abstratos (*e.g.*, "o funcionamento de um neurônio").

Hipótese 3: a relevância de cada tópico, o ritmo da aula e satisfação do aluno com o curso num dado instante de tempo são preditores que também afetam a aprendizagem.

Tabela 7.2. As primeiras linhas do conjunto de dados utilizados neste trabalho.

	turma	date	student	satisfaction	pace	relevance	component	learn
0	S0219	2019-04-18 00:00:00+03:00	Aluno 25	7.0	3.0	4.0	aplicações de clusterização de dados	1.0
1	S0219	2019-04-18 00:00:00+03:00	Aluno 25	7.0	3.0	4.0	MeanShift e DBSCAN	2.0
2	S0219	2019-04-18 00:00:00+03:00	Aluno 47	8.0	4.0	5.0	aplicações de clusterização de dados	1.0
3	S0219	2019-04-18 00:00:00+03:00	Aluno 47	8.0	4.0	5.0	MeanShift e DBSCAN	2.0
4	S0219	2019-04-18 00:00:00+03:00	Aluno 32	6.0	4.0	4.0	aplicações de clusterização de dados	1.0
5	S0219	2019-04-18 00:00:00+03:00	Aluno 32	6.0	4.0	4.0	MeanShift e DBSCAN	1.0

7.2. Pré-processamento

A primeira etapa na análise de dados foi transformar o arquivo dos dados para o padrão *tidy-data* [Wickham 2014], onde cada linha representa a resposta de um aluno para cada um dos até 4 ttópicos abordados numa aula. A Tabela 7.2 ilustra as primeiras linhas desse arquivo.

Note que as duas primeiras linhas representam as respostas do "Aluno 25" ao formulário aplicado na aula de 18/abr/2019 da turma S2019. Nessa aula dois ttópicos foram abordados: "aplicações de aplicações de dados" e "MeanShift e DBSCAN". No primeiro deles a aprendizagem foi de 1 unidade (última coluna) e, no segundo, de 2 unidades.

Em vista de cada formulário poder apresentar até quatro tópicos (dois neste exemplo), as informações de satisfação (*satisfativo*), ritmo (*pace*) e relevância (*relevante*) estão duplicados. Isso significa que, ao analisarmos essas variáveis, precisamos remover essas duplicidades. Fizemos isso através de uma amostragem que selecionava, aleatoriamente, *uma* linha para cada dupla (curso, turma, aula, aluno). Por exemplo, se selecionássemos a primeira linha, certamente ignoraríamos a segunda linha.

Esse arquivo foi posteriormente transformado:

Criamos uma coluna para identificar o curso (*course*), com valores DA ou DS, a partir da turma: aquelas que iniciam com "A" são de DA; as com "S", são DS.

A turma foi trocada por um valor numérico sequencial arbitrário.

As colunas de satisfação, ritmo e relevância foram normalizados no intervalo $[0, 10]$, apenas por simplicidade.

Criamos a coluna *component*, que mapeia o tópico para uma hierarquia *ad-hoc* de componentes de cada curso. Por exemplo, o tópico "Construção e execução de queries" foi mapeado para o componente "SQL/**Ferramenta**" (DA), enquanto "ARIMA. SARIMAX e Prophet" foi mapeado para "Séries temporais/**Algoritmo**" (DS). O mapeamento foi feito manualmente para os 985 tópicos presentes.

Criamos a coluna *tool* a partir da coluna *component* para identificar se aquela resposta se refere a um componente de ferramenta. O exemplo de mapeamento

Tabela 7.3. O conjunto de dados, pronto para a análise.

	course	turma	date	student	topicos	component	relevance	learn	satisfaction	pace	tool	algorithm
0	DS	6	2019-04-18	25	aplicações de clusterização de dados	Agrupamento	4	1	6	5	False	False
1	DS	6	2019-04-18	25	MeanShift e DBSCAN	Agrupamento/Algoritmo/ML	4	2	6	5	False	True
2	DS	6	2019-04-18	47	aplicações de clusterização de dados	Agrupamento	5	1	7	7	False	False
3	DS	6	2019-04-18	47	MeanShift e DBSCAN	Agrupamento/Algoritmo/ML	5	2	7	7	False	True
4	DS	6	2019-04-18	32	aplicações de clusterização de dados	Agrupamento	4	1	5	7	False	False

acima torna evidente como essa identificação foi feita. Variável categórica booleana: verdadeira para aulas sobre ferramentas.

Criamos a coluna *algorithm* a partir da coluna *component* para identificar se aquela resposta se refere a um componente de algoritmo. Variável categórica booleana: verdadeira para aulas sobre algoritmos.

A Tabela 7.3 ilustra o conjunto de dados após esse pré-processamento e já pronto para a análise.

7.3. Hipótese 1

Primeiramente vamos avaliar a hipótese 1. A estratégia e simples: para cada um dos cursos (DA e DS), vamos segmentar o conjunto de dados em dois subconjuntos, uma das aulas referentes a ferramentas; o outro com o complemento. Em seguida, aplicamos um teste de hipótese estatístico para avaliar a hipótese nula de que a media μ_f da população de todas as aulas referentes a ferramentas e igual a` media $\mu_{\neg f}$ da população de todas as demais aulas. Ou seja,

$$H_0 : \mu_f = \mu_{\neg f}$$
$$H_a : \mu_f > \mu_{\neg f}$$

O conjunto de dados pre-processado já contém a informação de que um dado componente se refere a uma ferramenta ou não (coluna *tool* na Figura 7.3). Assim, podemos segregar o conjunto de dados segundo esse critério.

Note que a hipótese 1 baseia-se na suposição de que podemos calcular a

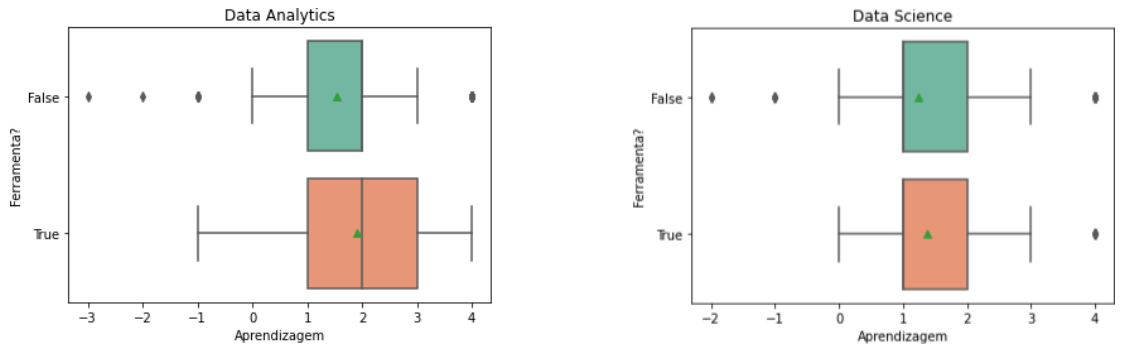

Figura 7.3. Distribuição da aprendizagem nos cursos de DA (esquerda) e DS (direita) para as aulas de ferramentas (true no eixo vertical) e as demais. Os triângulos verdes indicam a media amostral.

Tabela 7.4. Tamanho da amostra (#), média (e erro padrão) e desvio-padrão de cada subconjunto de DA (esquerda) e DS (direita).

Data Analytics			
Ferramenta?	#	$\bar{a}$	s_a
Sim	408	$1,9\pm0,1$	1,1
Não	1470	$1,55\pm0,05$	1,1

Data Science			
Ferramenta?	#	$\bar{a}$	s_a
Sim	232	$1,4\pm0,1$	1,0
Não	1531	$1,23\pm0,04$	0,9

media das aprendizagens. Há décadas discute-se sobre a possibilidade de interpretar a escala Likert como variável numérica. Considerando que os valores de q_{antes} (idem para d_{epois}) guardam entre si uma relação de ordenação, a partir da qual e possível construir um espaço mético [Barata, cap. 27], mais as recomendações de [Harpe 2015], consideramos válido assumir que a (equação 1) e de fato numérica e que, por isso, $\bar{a}$, μ_f e $\mu_{\neg f}$ existem.

A Figura 7.3 representa a distribuição dos valores de aprendizagem a para os subconjuntos: *true* refere-se às aulas de ferramenta.

A Tabela 7.4 apresenta as médias, desvio-padrão e erro-padrão da média (nnível de significância de 5%) para cada um dos subconjuntos dos cursos de DA e DS. Vemos, por exemplo, que para DA a media amostral de aprendizagem nas aulas de ferramentas e de $1,9\pm0,1$, ou seja, ela reside no intervalo [1,8;2,0] com confiança de 95%, que não contém a media da aprendizagem nas demais aulas. Observação análoga pode ser feita para DS, `a direita na tabela. Esse resultado e um indício de que realmente há uma diferença entre as médias populacionais, mas para fazermos essa afirmação precisamos do teste t.

Antes de efetuarmos o teste da hipótese 1, vamos checar se ambos os subconjuntos apresentam distribuição normal. Para isso utilizamos o teste de Lilliefors com nnível de significância de 5% (valor-padrao na literatura). O resultado e um valor-p bastante inferior ao nnível de significância, o que significa que a distribuição *não* e normal. Apesar disso, segundo [Triola 2015, p. 259] e possível realizar o teste t a seguir mesmo que a amostra não provenha de uma distribuição normal, desde que a amostra tenha tamanho maior do que 30, que e o nosso caso: o menor dos subconjuntos tem 232 observações.

Executamos o teste t de duas amostras independentes com populações cuja variância e desconhecida. Para DA obtivemos valor-$p \approx 4\times 10^{-8}$ com a estatística t positiva: veja a Tabela 7.5. Isso significa que a probabilidade de observarmos uma distribuição amostral com a media $\bar{a} = 1,9 \pm 0,1$, sob a suposição de que essa amostra tem origem numa população cuja media $\mu_f = 1,55\pm 0,05$, e da ordem de 4×10^{-8}. Ou seja, e improvável. Na verdade, e mais improvável do que o que originalmente escolhemos aceitar, o nnível de significância de 5%. Logo, podemos rejeitar H_0 e afirmar que de fato $\mu_f > \mu_{\neg f}$.

Em palavras, os alunos aprendem mais nas aulas referentes a ferramentas do que nas demais aulas de DA.

Analise análoga pode ser feita para DS (Tabela 7.5): o valor p e inferior ao

Tabela 7.5. Resultado do teste da hipótese 1 nos cursos DA e DS.

Curso	Valor p	Estatística t
DA	$< 10^{-11}$	$7{,}24$
DS	$0{,}03$	$2{,}13$

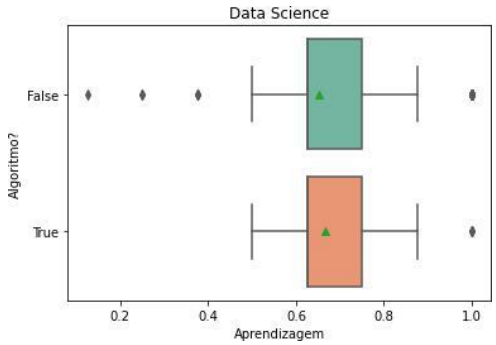

Figura 7.4. Distribui¸ao da aprendizagem nos cursos de DS para as aulas de algoritmos (true no eixo vertical) e as demais.

nnível de significância e a estatística t e positiva, significando que também para DS os alunos aprendem mais nas aulas referentes a ferramentas do que nas demais.

A luz da teoria do valor da expectativa, podemos argumentar que as aulas de ferramentas oferecem aos alunos uma relação explícita (instrumentalidade) com as demandas de vagas de postos de trabalho, de modo que a expectativa de obter um emprego (valência) promove no aluno motivação intrínseca para aprender. Essa relação não e tão evidente nas demais aulas, levando a uma instrumentalidade menor e, por conseguinte, menor motivação para aprender.

Essa aprendizagem aprimorada nas aulas de ferramentas pode ser, ainda, um efeito do chamado condicionamento operante [Petri], uma abordagem comportamentalista da motivação: a utilização das ferramentas leva a resultados concretos que, por sua vez, motivam aprendizagens subsequentes.

Independentemente do processo cognitivo, intangível `a experiência, o fato e que há diferença. Isso sugere que podemos utilizar essa motivação nas demais aulas. Duas possibilidades nos ocorrem: (1) desenvolver as demais aulas utilizando as ferramentas, se possível (aprendizagem significativa); e (2) tornando evidente, por discurso, a relação com os requisitos do mercado (teoria do valor da expectativa).

7.4. Hipótese 2

A verificação da hipótese 2 e similar a` primeira, exceto que dessa vez segmentamos os dados em aulas que envolvem explicitamente algoritmos ou não. Na Seção 7 nós mostramos como essas aulas foram identificadas (coluna *algorithm* do conjunto de dados), restando agora executar a segmentação e o teste t.

A Figura 7.4 e a Tabela 7.6 apresentam a distribuição da aprendizagem das aulas de ferramentas e das demais do curso de DS.

Tabela 7.6. Tamanho da amostra (#), media e desvio-padrão da aprendizagem em DS.

Algoritmo?	#	$\bar{a}$	s_a
Sim	110	$1,3 \pm 0,2$	0,88
Não	1653	$1,25 \pm 0,04$	0,88

Tabela 7.7. Estatísticas de teste da hipótese 2.

Curso	Valor p	Estatística t
DS	$0,41$	$0,83$

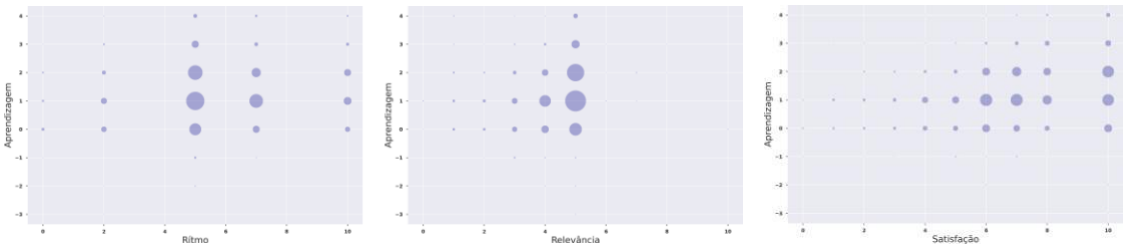

Figura 7.5. Bubble-plot da aprendizagem em função de cada um dos preditores.

O curso de DA não tem ênfase em algoritmos. Por isso não propusemos fazer teste análogo nele. De fato, podemos verificar no conjunto de dados que a quantidade de aulas com o atributo *algorithm* = *true* e zero.

Note, na Tabela 7.6, que a media amostral das aulas de algoritmo reside no intervalo $[1,1; 1,5]$, que envolve a media amostral das demais aulas. Além disso, os desvios-padrão são similares. Isso evidencia que as duas amostras têm origem na mesma população. Porém, apenas o teste t nos permitirá afirmar.

Ao efetuarmos o teste de hipótese, obtivemos valor-p de 41%. Ou seja, a probabilidade de observarmos a media $1,3 \pm 0,2$ numa amostra de uma população com media $1,25 \pm 0,04$ e de 41%, acima do nnível de significância de 5% assumido. Logo, podemos afirmar que as duas amostras advêm da mesma população. Consequentemente, *não* ha diferença de aprendizagem entre as aulas de algoritmos e as demais.

Utilizando novamente a teoria do valor da expectativa para interpretar esse resultado, isso significa que a instrumentalidade das componentes relacionadas a algoritmos e equivalente aquela associada as demais aulas. Ou seja, o aluno não vê distinção entre essas aulas e, com isso, apresenta o mesmo nnível de motivação para aprender.

7.5. Hipótese 3

Agora voltamos nossa atenção para a possível relação entre a aprendizagem e a satisfação, ritmo e relevância reportados pelos alunos.

O argumento aqui e baseado na aprendizagem individualizada: o ritmo, que pode ser completamente controlado pelo aluno nessa abordagem, oferece vantagem para a aprendizagem. E como temos informações extras sobre a satisfação e a relevância, vamos considerá-las também como critérios de comparação.

Tabela 7.8. Méticas de qualidade dos vários algoritmos de regressão aplicados a DS.

Algoritmo	Linear?	RMSE	MAE	R^2
XGBoost	Nao	0,769	0,581	**0,193**
Random Forest	Nao	0,772	0,594	0,186
Arvore de decisao	Nao	0,775	0,596	0,181
Adaboost	Nao	0,806	0,629	0,113
ElasticNet	Sim	0,819	0,640	0,083
SVR	Nao	0,835	0,612	0,048

Tabela 7.9. R_2 **e** $\bar{R}^2$ **para o algoritmo XGBoost aplicado a DA e DS.**

			Data Science		Data Analytics	
Relevância	Ritmo	Satisfação	R^2	$\bar{R}^2$	R^2	$\bar{R}^2$
C	C	C	0,193	0,191	0,213	0,211
	C	C	0,125	0,120	0,184	0,183
C		C	0,115	0,114	0,147	0,146
		C	0,075	0,074	0,130	0,130
C	C		0,071	0,070	0,096	0,095
C			0,052	0,051	0,019	0,018
	C		0,014	0,014	0,076	0,075

Como a aprendizagem a e uma variável numérica, temos em mãos um problema de regressão. Argumentamos que uma abordagem classificatória também e possível, desde que se tome o cuidado de garantir a integridade do espaço amostral de a, o intervalo $[-4, 4]$. Porém isso ficara para trabalhos futuros.

O modelo de regressão mais simples e o linear. Porém, uma análise visual dos gráficos de aprendizagem em função de cada um dos preditores não torna evidente qualquer possível relação linear (fig. 7.5). Então experimentamos outros algoritmos de regressão, conforme apresentado na Tabela 7.8.

O conjunto de dados usado para as regressões foi extraído do conjunto completo (Figura 7.3), tomando o cuidado de que um dado aluno não estivesse presente

Tabela 7.10. Métricas de desempenho dos vários algoritmos de regressão aplicados a DA.

Algoritmo	Linear?	RMSE	MAE	$R2$
XGBoost	Nao	0,958	0,760	**0,213**
Arvore de decisão	Nao	0,958	0,762	0,212
Random Forest	Nao	0,959	0,770	0,210
Adaboost	Nao	0,983	0,801	0,170
ElasticNet	Sim	0,985	0,793	0,167
SVR	Nao	0,994	0,797	0,152

mais do que uma vez. Fizemos esse tratamento com o intuito de evitar correla¸cao entre os exemplos.

Em seguida, aplicamos cada um dos algoritmos a 70% dos exemplos no conjunto de dados (conjunto de treinamento), explorando sistematicamente o espaço de hiper parâmetros a` procura de um mínimo global no erro quadrático medito da regressão.

O melhor índice de determinação, calculado sobre os 30% exemplos restantes (conjunto de teste), foi $R^2 \approx 0,193$ para o modelo XGBoost, composto por um conjunto de árvores de decisão simples [Friedman 2001]. Isso significa que nosso melhor modelo e capaz de explicar apenas 19% das variações na aprendizagem, a partir dos preditores propostos. Embora seja baixo, podemos argumentar que a aprendizagem sofre maior influência de outros parâmetros mais importantes, que não temos acesso aqui, como a metodologia de aprendizagem, a qualidade das atividades e conteúdo proposto *etc*.

Agora que sabemos que o modelo XGBoost obteve melhor desempenho, podemos experimentar variar os preditores: efetuamos a regressão considerando como preditores as combinações de um, dois e três preditores. Nesse caso, a mética mais relevante e o $\bar{R}^2$, que pondera o índice de determinação pela quantidade de preditores. Por exemplo, para dois modelos com o mesmo R^2, o primeiro com um e o segundo com dois preditores, o melhor modelo e aquele com maior $\bar{R}^2$.

Os resultados desse experimento estão na Tabela 7.9, ordenados por R^2 e $\bar{R}^2$. Concluímos que o melhor modelo e de fato o que utiliza todos os três preditores propostos: satisfação, relevância e ritmo.

A mesma análise pode ser feita para DA, o que nos leva inicialmente a` escolha do algoritmo, conforme ilustra a Tabela 7.10. Curiosamente, nesse caso o algoritmo de árvore de decisão obteve melhor desempenho que o *random forest* (o oposto para DS). Ainda assim, novamente o melhor (maior R^2) algoritmo foi o XGBoost; porém com R^2 similar ao caso de DS.

Em seguida variamos os preditores e obtemos os resultados apresentados na Tabela 7.9. O resultado e similar ao de DS: o modelo que utiliza todos os preditores e melhor.

Resumindo, para DA e DS o melhor modelo XGBoost utilizando os três preditores propostos consegue explicar apenas aproximadamente 20% das variações observadas (hipótese 3). Ainda assim, o fato de R^2 ser demasiadamente baixo torna essa conclusão duvidosa.

8. Conclusão

Neste trabalho nós avaliamos a influência da expectativa do aluno (de obter um posto de trabalho como cientista de dados) na sua aprendizagem e como isso pode ser incongruente com propostas de adotar currículos baseados em competências e habilidades em cursos livres de ciência de dados. Avaliamos ainda como a satisfação do aluno com o curso, a relevância percebida por ele sobre os tópicos abordados e o

ritmo de o curso também inferirem na aprendizagem.

Concluímos que nas aulas de ferramentas de ciências de dados os alunos apresentam maior aprendizagem, possivelmente devido `a relação explícita delas com os requisitos de vagas de trabalho. Esse resultado sugere duas estratégias para aprimorar a aprendizagem nas demais aulas: (1) desenvolvê-las utilizando as ferramentas (aprendizagem significativa) e (2) tornando mais evidente a relação dessas aulas com os requisitos do mercado (teoria do valor da expectativa). Concluímos ainda que efeito análogo *não* acontece para as aulas de algoritmos.

Finalmente, demonstramos o ritmo do curso, a relevância dos ttópicos abordados e a satisfação com o curso são capazes de explicar no máximo 20% da aprendizagem. Esse resultado e inconclusivo, mas sugere um prosseguimento: decompor a satisfação nas suas componentes, propondo novos experimentos com os alunos.

Referencias

[Anderson et al. 2014]ANDERSON, P. et al. An undergraduate degree in data sci-ence: Curriculum and a decade of implementation experience. In: *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. New York, NY, USA: Association for Computing Machinery, 2014. (SIGCSE '14), p. 145–150. ISBN 9781450326056. Disponıvel em: <https://doi.org/10.1145/2538862.2538936>.

[Baker, Isotani e Carvalho 2011]BAKER, R. S. J. d.; ISOTANI, S.; CARVALHO, A. M. J. B. d. Minera¸cao de dados educacionais: oportunidades para o brasil. *Revista Brasileira de Informatica na Educa¸cao*, v. 19(2), 2011.

[Barata]BARATA, J. C. A. *Notas para cursos de Fısica-Matematica*. [s.n.]. Dis-ponıvel em: <denebola.if.usp.br/jbarata/Notas_de_aula/notas_de_aula. html>.

[Base Nacional Comum Curricular]BASE Nacional Comum Curricular. Disponıvel em: <basenacionalcomum.mec.gov.br>.

[Bugnion 2016]BUGNION, P. *Scala for Data Science*. Birmingham, UK: Packt Pu-blishing, 2016.

[Butova 2015]BUTOVA, Y. the History of Development of Competency-Based Edu-cation. *European Scientific Journal*, v. 7881, n. June, p. 1857–7881, 2015. Dispo-nıvel em: <http://eujournal.org/index.php/esj/article/viewFile/5728/5535>.

[CF-DS-Release 2019]CF-DS-RELEASE, E. Data Science Competence Framework (CF-DS). 2019. Disponıvel em: <https://www.iabac.org/g-standards/IABAC-EDSF-CFDS-R2.pdf>.

[Cheng e VanDeGrift 2019]CHENG, H.; VANDEGRIFT. Course models for tea-ching data science. In: LU, B.; TUTTLE, S. (Ed.). *The journal of computing sciences in colleges*. [S.l.: s.n.], 2019. v. 35, n. 1.

[Demchenko, Belloum e Wiktorski 2017]DEMCHENKO, Y.; BELLOUM, A.; WIK-TORSKI, T. Data Science Body of Knowledge. n. July, 2017. Disponıvel em:

<http://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08 0.pdf>.

[Friedman 2001]FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001. Disponıvel em: <https://projecteuclid.org/euclid.aos/1013203451>.

[Harpe 2015]HARPE, S. E. How to analyze likert and other rating scale data. *Cur-rents in Pharmacy Teaching and Learning*, v. 7, p. 836–850, 2015. Disponıvel em: <https://www.sciencedirect.com/science/article/pii/S1877129715200196>.

[Hassan e Liu 2019]HASSAN, I. B.; LIU, J. Data science academic programs in the u.s. *J. Comput. Sci. Coll.*, Consortium for Computing Sciences in Colleges, Evans-ville, IN, USA, v. 34, n. 7, p. 56–63, abr. 2019. ISSN 1937-4771.

[Hayes 2019]HAYES, B. *Programming languages most used and recommended by Data Scientists*. 2019. Disponıvel em: <https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/>.

[He, Zhao e Chu 2020]HE, X.; ZHAO, K.; CHU, X. Automl: A survey of the state- of-the-art. *Knowledge-Based Systems*, p. 106622, 2020. ISSN 0950-7051. Disponıvel em: <http://www.sciencedirect.com/science/article/pii/S0950705120307516>.

[James et al. 2013]JAMES, G. et al. *An introduction to Statistical Learning with applications in R*. [S.l.: s.n.], 2013.

[Marques 2017]MARQUES, C. A. e Simone Cavalheiro e Adriana Bordini e M. O pensamento computacional por meio da robotica no ensino basico - uma revisao sistematica. *Brazilian Symposium on Computers in Education (Simposio Brasileiro de Informatica na Educa¸cao - SBIE)*, v. 28, n. 1, p. 82, 2017. ISSN 2316-6533. Disponıvel em: <https://www.br-ie.org/pub/index.php/sbie/article/view/7537>.

[McNicholas e Tait 2019]MCNICHOLAS, P. D.; TAIT, P. A. *Data Science with Ju-lia*. [S.l.]: CRC Press, 2019.

[Nagpal e Gabrani 2019]Nagpal, A.; Gabrani, G. Python for data analytics, scientific and technical applications. In: *2019 Amity International Conference on Artificial Intelligence (AICAI)*. [S.l.: s.n.], 2019. p. 140–145.

[NIST Big Data Public Working Group 2015]NIST Big Data Public Working Group. NIST Special Publication 1500-1 - NIST Big Data Interoperability Fra- mework: Volume 1, Definitions. *NIST Special Publication*, v. 1, p. 32, 2015. Dis-ponıvel em: <http://dx.doi.org/10.6028/NIST.SP.1500-1>.

[Pelizzari et al. 2002]PELIZZARI, A. et al. Teoria da aprendizagem significativa segundo ausubel. *Rev. PEC*, Curitiba, v. 1, p. 37–42, 2002. Disponıvel em: <http://portaldoprofessor.mec.gov.br/storage/materiais/0000012381.pdf>.

[Peng 2011]PENG, R. D. Reproducible research in computational science. *Science*, v. 334, p. 1226–1227, 2011.

[Petri]PETRI, H. L. *Behavioristic Approaches To Motivation*. Britannica Books. Disponıvel em: <https://www.britannica.com/topic/motivation/Behavioristic-approaches-to-motivation>.

[Raj et al. 2019]RAJ, R. K. et al. Data science education: Global perspectives and convergence. In: . New York, NY, USA: Association for Computing Ma-chinery, 2019. (ITiCSE '19), p. 265–266. ISBN 9781450368957. Disponıvel em: <https://doi.org/10.1145/3304221.3325533>.

[Triola 2015]TRIOLA, M. F. *Introdu¸cao `a estatística*. 9. ed. Rio de Janeiro: LTC, 2015.

[Wickham 2014]WICKHAM, H. Tidy data. *The Journal of Statistical Software*, v. 59, 2014. Disponıvel em: <http://www.jstatsoft.org/v59/i10/>.

[Zulauf 2006]ZULAUF, M. Ensino superior e desenvolvimento de habilidades para a empregabilidade: explorando a visao dos estudantes. *Sociologias*, 2006. ISSN 1517-4522.

A. *Reproducible research*

A análise de dados apresentada neste trabalho segue princípios de *reproducible research* [Peng 2011]. Isso significa que ela pode ser reproduzida utilizando os arquivos disponibilizados no seguinte repositorio Git:

http://github.com/irpagnossin/tcc-cae-icmc-usp