Um Farol para Criação e Avaliação de Cursos de Ciência de Dados: Os Referenciais Curriculares da SBC

Ângelo Brayner
brayner@dc.ufc.br
UFC

André P. L. de Carvalho
andre@icmc.usp.br
USP

Duncan Dubugras A. Ruiz
duncan.ruiz@pucrs.br
PUCRS - Escola Politécnica

Eduardo Ogasawara
eogasawara@ieee.org
CEFET/RJ

RESUMO

Este documento apresenta os referenciais de formação na área de Computação para os cursos de Bacharelado em Ciência de Dados (RF-CD-21). Estes Referenciais foram construídos em torno da noção de competência, em consonância com as competências definidas pela Força Tarefa em Ciência de Dados da *Association for Computing Machinery* (ACM) em 2021 (*ACM Data Science Task Force*). Assim como feito pela Socidade Brasileira de Computação (SBC) na preparação de um Currículo de Referência para outras áreas da Computação, as 17 (dezessete) competências apontadas como necessárias estão resumidas em oito eixos de formação, para facilitar a construção de currículos nas Instituições de Ensino Superior (IES) brasileiras. Cada eixo de formação relaciona os conteúdos considerados úteis no desenvolvimento das competências necessárias. Por fim, este referencial busca nortear a construção de um Projeto Pedagógico de Curso (PPC) para cursos de graduação em Ciência de Dados pelas IES, proporcionando flexibilidade para que cada uma delas defina seus PPC conforme sua vocação e seus objetivos.

CCS CONCEPTS

• **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Machine learning**; • **Applied computing → Education**; • **Information systems → Data management systems**.

PALAVRAS-CHAVE

Ciência de Dados, Referenciais Curriculares

1 INTRODUÇÃO

A Ciência de Dados é uma das áreas que tem recebido mais atenção e se tornado atrativa nas últimas décadas. A profissão de Cientistas de Dados foi inclusive rotulada como a mais sedutora deste século [11]. Exageros à parte, a procura do mercado por profissionais também tem sido maior que a oferta. Estes fatores proporcionaram a demanda por estabelecer referências de formação para graduações em Ciência de Dados [15] em diversos lugares do mundo. No Brasil, a demanda se concretiza em 2021, quando o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) atualizou a Classificação Internacional Normalizada da Educação (CINE) no Brasil com a inclusão de Ciência de Dados, na Área 6 (Computação e Tecnologia da Informação e Comunicação) e Área Detalhada 0617 (Desenvolvimento de Soluções Computacionais em Domínios Específicos) [21].

Em 2013 foram criados os primeiros cursos em Ciência de Dados, na Northern Kentucky University, na University of San Francisco e no College of Charleston, nos Estados Unidos. Na Europa, o primeiro curso foi criado 2014, na University of Warwick, Reino Unido. O rápido crescimento no número de cursos de graduação em Ciência de Dados, fez com que as duas maiores sociedades científicas do mundo, a *Association for Computing Machinery* (ACM) e o *Institute of Electrical and Electronics Engineers* (IEEE) se unissem em uma força tarefa para propor um currículo de referência para os cursos na área [15].

No início da década de 2020, observou-se um forte crescimento na criação de cursos com a denominação de Ciência de Dados. A partir deste fato, o INEP percebeu a necessidade de uma classificação do curso de Ciência de Dados no CINE Brasil (Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica do Brasil). A decisão daquele órgão foi a de incluí-lo na área de Computação e Tecnologia da Informação e Comunicação (área 6).

Nesse sentido, este trabalho apresenta o processo de construção dos referenciais curriculares de cursos de graduação em Ciência de Dados no Brasil. Tais referenciais servem como uma bússola na elaboração de um Projeto Pedagógico de Curso (PPC) para tais cursos de graduação. Assim, devem ser entendidos como diretrizes, que respeitam as vocações e objetivos das IES.

Além desta introdução, o artigo está organizado em mais seis seções. A Seção 2 apresenta uma discussão entre Ciência de Dados e Inteligência Artificial. A Seção 3 apresenta a demanda por Cientistas de Dados. As Seções 5 e 6 apresentam, respectivamente, o perfil de uma graduação em Ciência de Dados, bem como os eixos de formação. A Seção 7 apresenta uma expectativa do mercado quanto aos egressos deste curso. Finalmente, a Seção 8 apresenta as conclusões.

2 CIÊNCIA DE DADOS OU INTELIGÊNCIA ARTIFICIAL? EIS A QUESTÃO

Por terem vários aspectos em comum, não está claro para muitos no que a Ciência de Dados difere da Inteligência Artificial. A confusão

EduComp'24, Abril 22-27, 2024, São Paulo, São Paulo, Brasil (On-line)

entre as duas áreas tem como razão a sobreposição que existe entre elas, principalmente pelas duas trabalharem com aprendizado de máquina. Contudo, são áreas diferentes e é importante que essas diferenças sejam esclarecidas [45].

É difícil definir quem veio antes, pois as duas são resultados de evolução do conhecimento gerado e foram conhecidas por termos diferentes ao longo de suas histórias. Se nos ativermos ao nome hoje utilizado para denominá-las, a área de Inteligência Artificial, o termo surgiu em 1955, em uma proposta para um projeto de pesquisa de verão, que seria realizado no Dartmouth College, na cidade de Hannover, nos Estados Unidos. Dessa proposta, ou do evento em si, participaram pesquisadores de diferentes áreas, que entendiam os organizadores tinham relação com a definição do que seria a Inteligência Artificial: A ciência e engenharia de fazer máquinas inteligentes.

Atualmente, a Inteligência Artificial é formada por várias subáreas, pesquisadas isoladamente ou em conjunto, que inclui: aprendizado de máquina, busca, ética, lógica, mineração de dados, otimização, planejamento, processamento de linguagem natural, raciocínio, representação de conhecimento, robótica, sistemas multiagentes e visão computacional. A Inteligência Artificial é investigada por pesquisadores de diferentes áreas, com uma participação maior da área de Computação. Nas outras áreas, a Inteligência Artificial é utilizada principalmente em aplicações que buscam resolver problemas dessas áreas [4, 14, 24, 40].

Avanços recentes em redes neurais devem-se principalmente às arquiteturas de redes neurais profundas, aos algoritmos de Inteligência Artificial generativa, utilizada para gerar áudios, imagens, textos e vídeos [17, 20], aos modelos de atenção para reconhecimento de imagens e aos transformadores (*transformers*) [47] e grandes modelos de linguagem para processamento de linguagem natural, áudios e vídeos [38].

Paralelamente, em 1962, o estatístico e matemático John Tukey publicou o artigo *The Future of Data Analysis*, onde afirmava a importância da área de Computação para o futuro da análise de dados [46]. Em 1947, John Tukey havia proposto o termo *bit*, utilizado em 1948 por Claude Shannon quando formulou a teoria da informação.

O termo Ciência de Dados foi usado pela primeira vez pelo cientista da computação Peter Naur, ganhador do Turing Awards, prêmio da Computação que equivale a um prêmio Nobel, em 1965, como reconhecimento de suas contribuições na área de linguagens de programação. Naur propôs o termo em 1947, em seu livro *Concise Survey of Computer Methods* [33], onde definiu Ciência de Dados como a ciência de lidar com dados após definidos, enquanto a relação com o que eles representam é delegado a outros campos e ciências.

A Ciência de Dados é composta por diferentes subáreas, que são frequentemente trabalhadas em conjunto, como análise de riscos, arquiteturas de computadores de alto desempenho, aprendizado de máquina, banco de dados, estatística descritiva, engenharia de software, mineração de dados, otimização, pré-processamento de dados, processamento de linguagem natural, probabilidade, qualidade de dados, redes complexas, visão computacional e visualização [19].

Embora o termo Ciência de Dados já tivesse sido cunhado desde o milênio passado, as condições para se tornar em voga começaram na primeira década deste século. Elas são fruto da cada vez maior necessidade de desenvolver modelos preditivos [12] em um contexto de cada vez maior produção de dados [5]. Estes dados surgem com velocidade, variedade e volume. Em especial, estes três Vs caracterizaram o conceito amplamente difundido de *Big Data* [22].

Estas condições também levaram a maior demanda computacional, onde as primitivas de computação de alto desempenho (do inglês, *High Performance Computing*) passaram a apoiar as demandas de processamento de grandes volumes de dados, promovendo a chamada computação escalável intensiva de dados (do inglês, *Data Intensive Scalable Computing*). A realização de análise de dados nestas condições, seja por aprendizado estatístico ou aprendizado de máquina, tem sua origem na visão centrada em dados que exige processo e arcabouço computacional específico para o seu tratamento. Essa visão surge da percepção de que existe um ciclo de vida próprio para se realizar a Ciência de Dados [44], bem como a sua clarificação como ciência para os dados [27].

Enquanto a Inteligência Artificial cresceu principalmente com a contribuição de conhecimento da Computação, da Engenharia e da Física, a Ciência de Dados possui como tripé de formação as áreas de Computação, Matemática Aplicada e Estatística. É claro, frequentemente ambas fazem uso de métodos similares, como aprendizado de máquina. O aprendizado de máquina é justamente o que une mais fortemente a Ciência de Dados e a Inteligência Artificial. Ela tem apresentado um grande crescimento nos últimos anos; na maioria, por avanços nas redes neurais artificiais. A Ciência de Dados e Inteligência Artificial são campos intimamente relacionados, mas servem a propósitos diferentes.

A Ciência de Dados concentra-se principalmente nos dados, nos princípios de banco de dados, desde a coleta, preparação até chegar na extração de conhecimento. O foco está no apoio aos processos de tomada de decisão e nas resoluções de problemas complexos por meio da modelagem, análise e interpretação de dados. Há uma ênfase na preparação e gerência de dados [23]. Ao mesmo tempo, ela utiliza aprendizado de máquina, estatística e visualização para dar sentido a dados estruturados e não estruturados. É um campo versátil aplicável a diversas indústrias, desde finanças até saúde. Inclusive, é difícil pensar que a Inteligência Artificial consiga prosperar sem uma boa gerência de dados [18]. Ao mesmo tempo, fica difícil imaginar o futuro da Ciência de Dados sem caminhar lado a lado com a Inteligência Artificial. A visão centrada em dados, entretanto, é o princípio norteador que diferencia essas duas áreas [45].

3 PARA QUE PRECISAMOS DE UM CIENTISTA DE DADOS

3.1 O surgimento da ciência de dados

Na virada do século 20 para o século 21, a humanidade testemunhou um crescimento exponencial no volume de dados armazenados em mídias digitais. Esse fenômeno foi impulsionado pelo avanço de tecnologias de geração de dados como redes de sensores sem fio, Internet das coisas (IoT), entre outras. Naquele momento, a Ciência da Computação sentiu-se impelida a prover recursos tecnológicos, ao nível de hardware e software, que fossem capazes de garantir o gerenciamento daquele grande volume de dados de forma segura e eficiente, surgindo assim a tecnologia de *Big Data*.

Esse novo cenário fez com que se percebesse a necessidade de uma nova área dentro da computação, que abarcasse profissionais com conhecimentos sólidos em Computação e Estatística. Assim, surgiu em escala global a Ciência de Dados. Consequentemente, várias iniciativas surgiram em diferentes países, como Inglaterra, Alemanha, EUA, Canadá, entre outros, com o intuito de criar cursos de graduação para formar Cientistas de Dados. Por exemplo, já em 2014, a Universidade da Califórnia, em Berkeley, deu início a um curso de graduação em Ciência de Dados.

No Brasil, as primeiras iniciativas em formar cientistas de dados começam a surgir em 2020, a partir de duas abordagens distintas. Uma abordagem é a de adotar uma ênfase em Ciência de Dados dentro dos cursos de Estatística. Essa abordagem foi utilizada, por exemplo, em: (1) Instituto de Matemática da UFRJ, e no Instituto de Matemática, Estatística e Computação Científica da UNICAMP. A segunda abordagem é a de criar cursos com a denominação de Bacharelado em Ciência de Dados ou similares. Essa abordagem foi adotada: (1) pela Universidade Anhembi Morumbi e pela PUC-SP, na criação de bacharelado em Estatística e Ciência de Dados, (2) na Universidade Municipal de São Caetano do Sul (USCS);(3) pela Escola Politécnica da PUCRS, na criação, em 2020 de bacharelado em Ciência de Dados e Inteligência Artificial, já em seu terceiro ano de implantação, (4) pela UFC, com a criação do curso em Ciência de Dados pelos Departamentos de Computação e de Estatística, (5) pela USP, com a criação do curso em Ciência de Dados pelos Departamentos de Ciências de Computação, de Matemática, de Matemática Aplicada e Estatística e de Sistemas de Computação do ICMC-USP. Para ambos os casos, encontram-se Programas de Pós-graduação *stricto sensu* em Ciência de Dados. Na primeira categoria tem-se o curso precursor na área de Estatística: do IME-USP de 1970, enquanto na segunda tem-se o primeiro curso da área de Ciência da Computação da CAPES como foco em Ciência de Dados: do CEFET/RJ, desde 2016.

3.2 O papel do Cientista de Dados, Engenheiro de Dados e do Estatístico no processo de análise de dados

A profissão de Estatístico é regulamentada no Brasil. Ela tem suas diretrizes curriculares definidas na Resolução CNE/CES 8/2008 do MEC [28]. A sólida formação científica deve prover, ao egresso, proficiência para abordar os problemas de sua área de atuação, como coleta, organização e síntese de dados e ajuste de modelos, para investigar e implementar soluções para novos problemas e para assumir postura ética diante dos fatos.

O Cientista de Dados é um profissional com sólida formação em matemática, estatística e computação. Ele tem a competência de *pensar com dados*, em diferentes situações e para diversos campos de aplicação. Ele se vale da grande evolução da computação nas últimas décadas. Esta evolução ocorre tanto na infraestrutura física como processadores, dispositivos de armazenamento e de comunicação de dados, como nos algoritmos, métodos e técnicas para gestão dessa infraestrutura de forma eficiente e eficaz. Com desenvoltura para lidar com grandes e diversas quantidades de dados, armazenadas em unidades de armazenamento de alta capacidade ou em serviços na nuvem, o cientista de dados foca na análise de dados e no emprego de técnicas de aprendizado de máquina para munir seus usuários de padrões potencialmente interessantes que possam auferir vantagem competitiva aos mesmos.

O Engenheiro de Dados é um profissional que também possui sólida formação em computação, matemática e estatística aliada a formação em engenharia da computação no que diz respeito a equipamentos computacionais e de comunicação de dados, protocolos de comunicação, programação de dispositivos, de sistemas operacionais e de gestão de redes de computadores, e de ambientes computacionais em nuvem. O foco do engenheiro de dados é entender a produção, transmissão, armazenamento, preparação e disponibilização dos dados, de forma eficiente e reproduzível, para viabilizar o lidar com dados em grande volume, produzidos em grande velocidade e com bastante variedade de formatos.

Portanto, os papéis do Estatístico, do Cientista de Dados e do Engenheiro de Dados estão sinergicamente integrados e atuam de forma complementar.

4 PROCESSO DE CONSTRUÇÃO DO ARCABOUÇO DOS REFERENCIAIS CURRICULARES

O ponto de partida para a construção dos referenciais é o enquadramento de Ciência de Dados no CINE Brasil (Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica do Brasil) como um curso de Computação. Com isso, fez-se necessário obedecer às Diretrizes Curriculares Nacionais para os cursos de graduação na área da Computação ([31]) no *caput* de seu artigo 4º, onde são especificadas as competências e habilidades que todo o egresso deve ser dotado. Como um dos primeiros passos, foi proposto um detalhamento para este artigo 4º, levando em consideração a flexibilidade para atender domínios de aplicação diversos, além das vocações institucionais.

A construção dos referenciais também se valeu de documentos da Association of Computing Machinery - ACM, IEEE Computer Society e da National Academy of Sciences. São eles:

- Computer Science Curricula 2013 [36]
- Data Science for Undergraduates: Opportunities and Options 2018 [35]
- Computing Curricula 2020 - Paradigms for Global Computing Education [34]
- Computing Competencies for Undergraduate Data Science Curricula 2021 [15]

Foram também levados em consideração as iniciativas brasileiras na formação de cientistas de dados provenientes de diferentes IES brasileiras, como relatadas na seção 3.1. O estudo das convergências, divergências e especificidades desses diferentes documentos nortearam a equipe multidisciplinar formada, pela SBC, para a construção dos referenciais.

A metodologia adotada para desenvolvimento dos referenciais segue um modelo baseado em competências e habilidades, garantindo conformidade com os Referenciais de Cursos de Graduação em Computação ([48]). Assim, adotou-se como referência a Taxonomia de Bloom Revisada ([13]). Nesta taxonomia, uma competência pode expressar o conhecimento, as habilidades ou as atitudes esperadas do egresso do curso, sob a perspectiva de objetivos de aprendizagem. Além disso, as competências foram articuladas e estruturadas em eixos temáticos de formação ([10]), apresentados na seção 6.

Ao longo do processo, foram feitas reuniões conjuntas com a Associação Brasileira de Estatística (ABE),para buscar estabelecer referenciais com a maior sinergia possível entre as áreas de Ciência da Computação e Estatística.

Por fim, foi realizada consulta pública em nível nacional para colher críticas e sugestões aos referenciais propostos. Todas as manifestações foram consideradas, em maior ou menor grau, ou respondidas adequadamente. Ao final de 2023 foi publicada, pela SBC, a versão final dos referenciais curriculares para o curso de graduação em Ciência de Dados [42].

5 PERFIL, HABILIDADES E COMPETÊNCIAS

O volume e a complexidade de dados produzidos por pessoas e máquinas estão crescendo rapidamente. Esses dados contêm informações valiosas que podem resultar em avanços não só em suas áreas de origem, mas também na melhoria de serviços públicos, no desenvolvimento social, no crescimento econômico e em novas oportunidades de trabalho. Os empregos criados podem reduzir riscos não apenas para as pessoas, mas também para o planeta, contribuindo para a conservação do meio ambiente. Para alcançar isso, é essencial educar profissionais ao longo da cadeia de análise de dados, dotando-os de conhecimentos em Ciência da Computação, Estatística e Matemática. Essa educação de qualidade possibilita o desenvolvimento de ferramentas computacionais capazes de explorar eficaz e competentemente os dados gerados e de apoiar o processo de tomada de decisão em instituições acadêmicas, governamentais e empresariais [7, 8, 37].

O profissional formado em Ciência de Dados deve possuir a capacidade de compreender como funcionam os processos de coleta, administração e análise eficaz de grandes volumes de dados, em diferentes formatos e provenientes de diversas fontes. Em relação a esses dados, deve ser capaz de definir e implementar práticas de gestão e uso, criar estratégias para formular e testar hipóteses, interpretar e extrair conhecimentos valiosos, desenvolver e validar algoritmos para sua análise, colaborar com especialistas de outras áreas de conhecimento e realizar seu trabalho de forma ética e colaborativa [9, 11].

O profissional formado em Ciência de Dados deve possuir a capacidade de colaborar de maneira ética em várias áreas, trabalhando em conjunto com especialistas de diferentes campos, que compreendam a importância dos dados analisados e possam validar os resultados obtidos. Dessa forma, os graduados em Ciência de Dados devem estar aptos a analisar dados provenientes de uma ampla gama de disciplinas, incluindo administração, agronomia, ciências florestais, ciências sociais, economia, engenharia, geografia, história, medicina e veterinária, entre outras [1, 41].

A seguir, é apresentado o perfil geral dos graduados em cursos de Bacharelado em Ciência de Dados. O Bacharelado em Ciência de Dados visa a formação de profissionais com a capacidade de *pensar com dados*, possuindo competência teórica, técnica, metodológica e experiência prática para lidar com diversas situações e campos de aplicação. De forma resumida, o egresso deve ser apto a: (i) formular e aprimorar questões relevantes; (ii) adquirir, modelar e explorar dados associados; (iii) processar dados e realizar análises necessárias; (iv) comunicar conhecimento significativo; (v) apoiar o desenvolvimento e implementação de soluções com base nos resultados alcançados; e (vi) considerar os aspectos éticos e sociais relacionados à sua atuação. Para isso, deve demonstrar as seguintes habilidades e competências [15]:

(i) Ter uma base sólida em Computação, Matemática e Estatística, permitindo a aplicação de conceitos dessas áreas em tarefas de Ciência de Dados;
(ii) Usar eficazmente técnicas computacionais, matemáticas e estatísticas para avaliar de maneira analítica a viabilidade e, quando possível, extrair conhecimento dos dados disponíveis, a fim de realizar descobertas em diversos domínios de aplicação e apoiar a tomada de decisões;
(iii) Aplicar os princípios de Tecnologia da Informação e Comunicação (TIC) para pesquisar, projetar, implementar e avaliar novas abordagens e técnicas na construção de ferramentas de análise de dados;
(iv) Realizar experimentos utilizando várias infraestruturas para gerenciar e manipular eficazmente dados, tanto estruturados quanto não estruturados, ao longo do ciclo de vida dos dados;
(v) Definir e implementar estratégias de gerenciamento de dados, abrangendo a curadoria, coleta, integração, armazenamento, visualização, preservação e disponibilização de dados para processamento futuro;
(vi) Gerenciar projetos interdisciplinares que englobem as diversas fases do ciclo de vida dos dados;
(vii) Identificar novos desafios, necessidades e oportunidades de negócios, desenvolvendo soluções inovadoras;
(viii) Investigar, compreender e estruturar as características dos domínios de aplicação em vários contextos, levando em consideração aspectos ambientais, éticos, sociais, legais e econômicos;
(ix) Trabalhar de forma colaborativa, tanto com profissionais da mesma área quanto de áreas diferentes;
(x) Seguir os princípios de uma Ciência de Dados justa, transparente, imparcial e respeitosa da privacidade, cumprindo os requisitos das leis de proteção de dados em vigor;
(xi) Ter uma visão crítica e criativa na identificação e resolução de problemas, contribuindo para o avanço da área;
(xii) Atuar de forma empreendedora, abrangente e cooperativa, atendendo às demandas ambientais, sociais e econômicas em sua região, no Brasil e no mundo;
(xiii) Usar recursos de maneira racional e transdisciplinar;
(xiv) Ser capaz de atuar em um mundo de trabalho globalizado, buscando o domínio de idiomas estrangeiros, com ênfase no inglês.

Para adquirir essas habilidades e competências ao longo do curso, o graduado deve passar por várias experiências de aplicação do conhecimento adquirido em diversos contextos organizacionais e sociais. Com sua formação sólida, o profissional formado no curso estará preparado para trabalhar em empresas de diferentes setores, em órgãos públicos e do terceiro setor, além de instituições de ensino superior e institutos de pesquisa [3, 6].

6 EIXOS DE FORMAÇÃO

A estrutura do referencial de formação para Ciência de Dados (RF-CD-21) se baseia no modelo conceitual apresentado no primeiro capítulo dos referenciais de 2017 presentes para Ciência da Computação [2]. Ao mesmo tempo, levando-se em consideração o cenário

de Ciência da Ciência de Dados, o RF-CD-21 também leva em consideração as definições de Competências da ACM para Ciência de Dados [15]. As 17 competências e habilidades, tanto gerais quanto derivadas, definidas pelas Diretrizes Curriculares Nacionais (DCN) para graduados em cursos de Bacharelado em Ciência de Dados [30], foram organizadas em oito eixos de formação.

Cada eixo de formação do RF-CD-21 representa uma macro competência e abrange um conjunto de competências derivadas, que, por sua vez, estão relacionadas às competências e habilidades propostas mapeáveis a DCN vigente da Ciência da Computação. No conjunto, as competências associadas aos eixos de formação capacitam o formando em Ciência de Dados a abordar as diversas dimensões da computação. Os eixos de formação refletem a compreensão de que a formação em Ciência de Dados deve considerar: a capacidade de atuar em todas as etapas relacionadas à aplicação de Ciência de Dados em problemas variados, desde a concepção de sistemas computacionais até a implementação eficaz de soluções adequadas; a habilidade de se adaptar e adquirir novos conhecimentos; e a capacidade de buscar estudos avançados para promover o desenvolvimento da ciência e da tecnologia. Os eixos de formação compreendem:

(i) Fundamentos de Matemática, Estatística e Computação para Ciência de Dados;
(ii) Solução de Problemas;
(iii) Desenvolvimento de Sistemas;
(iv) Engenharia e Exploração de Dados;
(v) Dados em Grande Escala;
(vi) Mineração de Dados e Aprendizado de Máquina;
(vii) Aprendizado Contínuo e Autônomo;
(viii) Ciência, Tecnologia, Inovação e Empreendedorismo;

Cada eixo de formação segue a seguinte estrutura:

- Código: um número em algarismos arábicos que identifica o eixo de formação;
- Título: um rótulo que identifica o eixo de formação;
- Descrição: um breve texto que fornece contexto à competência associada ao eixo de formação;
- Competência de eixo: a descrição da competência ligada ao eixo de formação;
- Competências derivadas: uma lista de competências originadas das 48 competências e habilidades, tanto gerais quanto específicas, estabelecidas pelas DCN, necessárias para desenvolver a competência de eixo. As competências gerais das DCN são marcadas com o identificador CG, e as específicas do curso de Bacharelado em Ciência de Dados com o identificador CE. Cada competência derivada é composta por três elementos:
  - Código: formado pela combinação da letra C (inicial da palavra "competência", do código do eixo (de 1 a 8) e um número em algarismos arábicos que denota a ordem sequencial da competência derivada dentro do contexto do eixo de formação;
  - Classificação: um dos seis níveis do processo cognitivo da Taxonomia de Bloom Revisada [13];
  - Conteúdo: uma lista de conhecimentos que precisam ser abordados para desenvolver a competência derivada. Cada conteúdo é identificado por um título, proveniente da lista presente nas seções 3.1 e 3.2 do Parecer CNE/CSE 136/2012 [29]. A maioria dos detalhes sobre os conteúdos pode ser obtida nos Currículos de Referência da SBC de 1999 a 2005.

É importante destacar que conteúdo e disciplina não são termos intercambiáveis. A relação entre conteúdos e disciplinas é um dos principais desafios na elaboração da estrutura curricular de cada curso. Uma disciplina oferecida por uma instituição de ensino superior específica pode abranger diversos dos conteúdos listados nestes referenciais, combinando-os para abordar situações complexas. Da mesma forma, um determinado conteúdo pode ser abordado em mais de uma disciplina, demonstrando sua aplicabilidade em diferentes contextos, possivelmente com diferentes níveis de aprofundamento. A organização de conteúdos e disciplinas depende, fundamentalmente, da abordagem adotada por cada curso na formação das competências de seus alunos [25].

Uma competência das DCN pode estar relacionada a mais de um eixo, sendo que o conteúdo específico varia conforme a associação entre o eixo de formação e a competência das DCN. Isso significa que uma competência das DCN pode exigir diferentes conjuntos de conteúdos, dependendo do eixo em que está inserida. Da mesma forma, um conteúdo pode fazer parte de mais de um eixo. Além disso, um conteúdo pode ser vinculado a mais de uma competência das DCN em um determinado eixo.

Um curso pode optar por uma estratégia de estruturação de sua matriz curricular de modo que cada disciplina seja projetada para desenvolver uma ou mais competências das DCN, no contexto de um ou mais eixos de formação. Portanto, cada disciplina deve cobrir (total ou parcialmente) os conteúdos recomendados para as respectivas competências das DCN, conforme os eixos de formação aplicáveis.

A seguir, cada eixo de formação é explicado em relação às suas competências derivadas e aos conteúdos associados.

Fundamentos de Matemática, Estatística e Computação para Ciência De Dados. Neste eixo, os estudantes desenvolvem a compreensão das teorias e princípios fundamentais nas áreas de Computação, Estatística e Matemática. Essa compreensão é aplicada na resolução de desafios técnicos em Ciência de Dados, abrangendo sistemas de aplicação específica.

Resolução De Problemas. Este eixo tem foco na resolução de problemas por meio da computação, utilizando etapas claramente definidas. Os egressos devem ser capazes de criar soluções para problemas complexos que envolvem relações entre diferentes áreas de conhecimento. Isso inclui identificar problemas com soluções algorítmicas viáveis, escolher ou desenvolver algoritmos apropriados para situações específicas, implementar essas soluções com a programação adequada e compreender a complexidade das soluções encontradas.

Desenvolvimento de Sistemas. Este eixo abrange o desenvolvimento de sistemas computacionais, incluindo a criação e adaptação de sistemas, especialmente aqueles que envolvem análise de dados. Esse processo abrange a identificação, análise, especificação e validação de requisitos, bem como o projeto de soluções que estejam alinhadas com o ambiente de aplicação. Além disso, engloba a implementação, teste e manutenção de sistemas computacionais,

juntamente com a identificação e análise de possíveis vulnerabilidades.

Engenharia e Exploração de Dados. Neste eixo, os alunos adquirem conhecimento sobre como os dados são produzidos, armazenados e gerenciados. Isso envolve o estudo de Sistemas Gerenciadores de Banco de Dados, que fornecerá aos estudantes uma base sólida sobre como dados em abundância são manipulados. Os alunos também aprendem sobre governança de dados, que inclui aspectos de qualidade, privacidade e curadoria de dados. O desenvolvimento de projetos conceituais, lógicos e físicos de banco de dados, juntamente com a identificação de gargalos e soluções para melhorar o acesso a bancos de dados, são habilidades fundamentais nesse eixo.

Dados em Larga Escala. Neste eixo, os estudantes se concentram em soluções para lidar com dados em grande escala, que podem ser caracterizados por seu volume, variedade, velocidade, valor e veracidade. Os alunos aprendem a planejar e executar a implementação de sistemas computacionais baseados em programação e gerenciamento de dados em grande escala. Eles também se tornam aptos a identificar problemas que exigem soluções escaláveis e a selecionar ou criar algoritmos escaláveis para lidar com armazenamento de dados, computação de alto desempenho e teoria da complexidade. Garantir a conformidade com normas legais e éticas é uma parte essencial deste eixo.

Mineração de Dados e Aprendizado de Máquina. Neste eixo, os alunos exploram técnicas para extrair conhecimento de conjuntos de dados e criar modelos que capturem suas principais características. Isso envolve a análise, aplicação e desenvolvimento de técnicas e algoritmos para mineração de dados e aprendizado de máquina.

Aprendizado Contínuo e Autônomo. Este eixo se concentra no desenvolvimento de habilidades pessoais, em vez da simples aquisição de conteúdo tradicional. Os estudantes aprendem a aprender de forma contínua e autônoma, acompanhando a evolução da Ciência de Dados, avaliando novas ferramentas, tecnologias e métodos, e se adaptando rapidamente a mudanças tecnológicas e novos ambientes de trabalho.

Ciência, Tecnologia e Inovação. Este eixo prepara os estudantes para o desenvolvimento de estudos avançados e para enfrentar grandes desafios na área da Ciência de Dados. Os alunos devem compreender os fundamentos científicos e tecnológicos da computação, dominar ferramentas matemáticas e estatísticas, adaptar-se a novos domínios de aplicação e realizar ações inovadoras no desenvolvimento de soluções eficazes, incluindo novos produtos e processos. Eles também devem ser capazes de se adaptar rapidamente a mudanças tecnológicas e novos ambientes de trabalho.

7 O MERCADO DE TRABALHO

A área de Ciência de Dados está em alta no mercado de trabalho, tanto no Brasil como no cenário global, devido à crescente importância dos dados para empresas, que os utilizam para decisões estratégicas e melhoria de desempenho. No Brasil, o setor de Ciência de Dados promete um futuro promissor, com projeções indicando a criação de 200 mil novos empregos na área até 2025, segundo a consultoria IDC [26, 43].

O mercado de Ciência de Dados é comumente apresentado como *Big Data*. Normalmente o termo *Big Data* engloba as características associadas à indústria 4.0. Este termo é cunhado nas questões associadas a quanto a internet demandou e continua demandando mudanças nos processos produtivos e das empresas. Nota-se que na visão empresarial as questões de *Big Data* também são conhecidas como *Business Intelligence*, *Data Analytics* e Inteligência Artificial [16, 43].

A área de Ciência de Dados oferece diversas oportunidades de trabalho em setores como tecnologia, finanças, marketing, saúde e varejo, ocupando diferentes funções: analista de dados, engenheiro de dados e cientista de dados, propriamente dito. Como analista de dados, sua responsabilidade é preparar os dados para análise, realizar análises estatísticas e criar relatórios. Enquanto engenheiros de dados coletam, armazenam e processam dados. Já os cientistas de dados coletam, analisam e interpretam dados, desenvolvendo modelos e algoritmos para resolver problemas complexos. Esses profissionais desempenham um papel essencial na mineração e interpretação de dados complexos [16, 32].

No entanto, o mercado de trabalho em Ciência de Dados enfrenta desafios, como a escassez de profissionais qualificados. O Brasil ainda carece de cursos de graduação e pós-graduação na área. O mercado espera que os cientistas de dados possuam habilidades analíticas, conhecimento em estatística e matemática, capacidade para solucionar problemas, domínio da programação, habilidades de comunicação e trabalho em equipe [32, 39]. Observe-se que esta expectativa está em sintonia com os referenciais propostos neste documento.

8 CONSIDERAÇÕES FINAIS

Em um mundo em constante evolução, onde a Ciência de Dados desempenha um papel fundamental em uma variedade de setores, desde a pesquisa acadêmica até a indústria e a inovação, a formação em Ciência de Dados torna-se cada vez mais importante. O referencial de formação para Ciência de Dados (RF-CD-21) apresentado representa um marco na estruturação para graduação em Ciência de Dados no Brasil. O RF-CD-21 está organizado em oito eixos de formação foram elaborados e alinhados com as competências e habilidades estabelecidas pelas Diretrizes Curriculares Nacionais. Os eixos abordam desde os fundamentos de matemática, estatística e computação até a inovação e a capacidade de aprendizado contínuo. A estruturação flexível dos conteúdos e a possibilidade de relacioná-los com disciplinas específicas oferecem às instituições de ensino a liberdade de adaptar o currículo às suas necessidades e vocações individuais, garantindo, ao mesmo tempo, uma base sólida na formação dos Cientistas de Dados.

REFERÊNCIAS

[1] Paul Anderson, James McGuffee, and David Uminsky. 2014. Data science as an undergraduate degree. In *SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. Association of Computing Machinery - ACM, New York - NY - USA, 705 – 706.

[2] Renata Araujo, Alcides Calsavara, Alessandro Cerqueira, and Jair Leite. 2019. *Referenciais de Formação para os Cursos de Graduação em Computação no Brasil - Competências Atitudinais.* Sociedade Brasileira de Computação - SBC, Porto Alegre - RS.

[3] Monya Baker. 2015. Data science: Industry allure. *Nature* 520, 7546, 253 – 255.

[4] Y. Bengio, Y. Lecun, and G. Hinton. 2021. Deep learning for AI. *Commun. ACM* 64, 7, 58–65.

[5] Francine Berman. 2008. Got data?: A guide to data preservation in the information age. *Commun. ACM* 51, 12, 50 – 56.

[6] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, and Alexander S. Szalay. 2018. Realizing the potential of data science. *Commun. ACM* 61, 4, 67 – 72.

[7] Longbing Cao. 2017. Data science: A comprehensive overview. *Comput. Surveys* 50, 3, 1–42.

[8] Longbing Cao. 2017. Data science: Challenges and directions. *Commun. ACM* 60, 8, 59 – 68.

[9] Longbing Cao. 2019. Data Science: Profession and Education. *IEEE Intelligent Systems* 34, 5, 35 – 44.

[10] Lea das Graças Camargos Anastasiou. 2010. Desafios da Construção Curricular em Visão ntegrativa: Elementos para Discussão. In *Convergências e tensões no campo da formação e do trabalho – Textos do XV ENDIPE - Encontro Nacional de Didática e Prática de Ensino.* Autêntica, Campinas - SP.

[11] Thomas H. Davenport and D.J. Patil. 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review* 90, 10, 5.

[12] Vasant Dhar. 2013. Data science and prediction. *Commun. ACM* 56, 12, 64 – 73.

[13] Ana Paula do Carmo Marcheti Ferraz and Renato Vairo Belhot. 2010. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. *Gestão & Produção* 17, 421–431.

[14] S. Dong, P. Wang, and K. Abbas. 2021. A survey on deep learning and its applications. *Computer Science Review* 40.

[15] ACM Data Science Task Force. 2021. *Computing Competencies for Undergraduate Data Science Curricula.* Association of Computing Machinery - ACM, New York - NY - USA. 134 pages.

[16] Globo. 2023. Big data: análise de dados é aliada da indústria. https://oglobo.globo.com/patrocinado/dino/noticia/2023/09/27/big-data-analise-de-dados-e-aliada-da-industria.ghtml

[17] Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchán. 2023. A survey of Generative AI Applications.

[18] Christoph Gröger. 2021. There is no AI without data. *Commun. ACM* 64, 11, 98 – 108.

[19] David J. Hand. 2015. Statistics and computing: the genesis of data science. *Statistics and Computing* 25, 4, 705 – 711.

[20] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18, 7, 1527 – 1554.

[21] INEP. 2021. Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica (Cine Brasil). https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil/historico

[22] H.V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. *Commun. ACM* 57, 7, 86 – 94.

[23] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. 2023. The Principles of Data-Centric AI. *Commun. ACM* 66, 8, 84 – 92.

[24] A. Khan, A. Sohail, U. Zahoora, and A.S. Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* 53, 8, 5455–5516.

[25] José Carlos Libâneo. 2017. *Didática.* Cortez Editora, São Paulo.

[26] LinkedIn. 2020. Brasil: Empregos em alta. https://business.linkedin.com/pt-br/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-cont-fact-report

[27] Kate Matsudaira. 2015. The science of managing data science. *Commun. ACM* 58, 6, 44 – 47.

[28] MEC. 2008. Diretrizes Curriculares Nacionais do curso de Estatística. http://portal.mec.gov.br/cne/arquivos/pdf/2008/rces008_08.pdf

[29] MEC. 2012. Parecer CNE/CES nº 136/2012, aprovado em 9 de março de 2012. https://normativasconselhos.mec.gov.br/normativa/view/CNE_PAR_CNECESN1362012.pdf

[30] MEC. 2016. Diretrizes Curriculares - Cursos de Graduação. http://portal.mec.gov.br/component/content/article?id=12991

[31] MEC. 2016. Diretrizes Curriculares - Cursos de Graduação na área de Computação. http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=52101-rces005-16-pdf&category_slug=novembro-2016-pdf&Itemid=30192

[32] MIT. 2022. Empresas da América Latina investem em dados e IA para acelerar negócios - MIT Technology Review. https://mittechreview.com.br/empresas-da-america-latina-investem-em-dados-e-ia-para-acelerar-negocios/

[33] Peter Naur. 1974. *Concise Survey of Computer Methods* (1 ed.). Petrocelli Books, New York.

[34] Association of Computing Machinery ACM. 2020. *Computing Curricula 2020 - CC2020: Paradigms for Global Computing Education.* Association of Computing Machinery - ACM, New York - NY - USA. 205 pages.

[35] National Academies of Sciences Engineering and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options.* The National Academies Press, New York - NY - USA. 139 pages.

[36] The Joint Task Force on Computing Curricula. 2013. *Computer Science Curricula 2013.* Association of Computing Machinery - ACM and IEEE Computer Society, New York - NY - USA. 518 pages.

[37] C.L. Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, 314 – 347.

[38] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S.S. Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *Comput. Surveys* 51, 5.

[39] Redação. 2022. O grande valor da Ciência de Dados nas corporações. https://liga.ventures/insights/startups/o-grande-valor-da-ciencia-de-dados-nas-corporacoes/

[40] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B.B. Gupta, X. Chen, and X. Wang. 2022. A Survey of Deep Active Learning. *Comput. Surveys* 54, 9.

[41] Faryad Sahneh, Meghan A. Balk, Marina Kisley, Chi-Kwan Chan, Mercury Fox, Brian Nord, Eric Lyons, Tyson Swetnam, Daniela Huppenkothen, Will Sutherland, Ramona L. Walls, Daven P. Quinn, Tonantzin Tarin, David LeBauer, David Ribes, Dunbar P. Birnie, Carol Lushbough, Eric Carr, Grey Nearing, Jeremy Fischer, Kevin Tyle, Luis Carrasco, Meagan Lang, Peter W. Rose, Richard R. Rushforth, Samapriya Roy, Thomas Matheson, Tina Lee, C. Titus Brown, Tracy K. Teal, Monica Papes, Stephen Kobourov, and Nirav Merchant. 2021. Ten simple rules to cultivate transdisciplinary collaboration in data science. *PLoS Computational Biology* 17, 5, 1–12.

[42] SBC. 2023. SBC apresenta Referenciais de Formação para os Cursos de Bacharelado em Ciência de Dados. https://www.sbc.org.br/noticias/2505-sbc-apresenta-referenciais-de-formacao-para-os-cursos-de-bacharelado-em-ciencia-de-dados

[43] Vitor Soares. 2022. Por que o big data é uma das grandes apostas do setor de tecnologia. https://napratica.org.br/por-que-o-big-data-e-uma-das-grandes-apostas-do-setor-de-tecnologia/

[44] Victoria Stodden. 2020. The data science life cycle. *Commun. ACM* 63, 7, 58 – 66.

[45] M. Tamer Özsu. 2023. Data Science - -A Systematic Treatment. *Commun. ACM* 66, 7, 106 – 116.

[46] John W. Tukey. 1962. The Future of Data Analysis. *Annals of Mathematical Statistics* 33, 1–67.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 2017-December. 5999 – 6009.

[48] Avelino Zorzo, Daltro Nunes, Ecivaldo Matos, Igor Seinmacher, Jair Leite, Renata Araujo, Ronaldo Correia, and Simone Martins. 2017. *Referenciais de Formação para os Cursos de Graduação em Computação.* Sociedade Brasileira de Computação (SBC), Porto Alegre - RS. 153 pages. https://www.sbc.org.br/documentos-da-sbc/send/127-educacao/1155-referenciais-de-formacao-para-cursos-de-graduacao-em-computacao-outubro-2017