VII ENCONTRO INTERNACIONAL DE GESTÃO, DESENVOLVIMENTO E INOVAÇÃO
IMPACTOS DA INTELIGÊNCIA ARTIFICIAL NA SOCIEDADE E ORGANIZAÇÕES
ISSN: 2594-8083
VII EIGEDIN 4 a 7 de Junho de 2024 Totalmente Online e Gratuito

( ) Graduação ( X ) Pós-Graduação

# Cientistas de Dados: estudo das habilidades e ferramentas que utilizam

**Fabiano Castello**
**FEA-USP**
**fabianocastello@gmail.com**

**Cesar Alexandre de Souza**
**FEA-USP**
**calesou@usp.br**

**Rodrigo Baroni de Carvalho**
**PUC-MG**
**baroni@pucminas.br**

**Juliana N. do Nascimento Correa**
**FEA-USP**
**juliana.nelia.correa@usp.br**

## RESUMO

O cientista de dados é uma profissão em ascensão, mas que carece de uma definição consensual e de um entendimento claro das habilidades requeridas. O objetivo do estudo é investigar essa função sob a ótica da academia, empregadores brasileiros e profissionais, focando nas habilidades e ferramentas utilizadas. Utilizando uma abordagem múltipla que inclui revisão sistemática da literatura, análise de anúncios de emprego no LinkedIn, e levantamento de dados secundários de prática profissional, o estudo busca uma compreensão abrangente. A análise revelou alinhamento em grande parte entre as perspectivas, com dados de campo complementando dados da literatura com detalhes técnicos de ferramentas e atualização de técnicas. Um modelo conceitual de habilidades e ferramentas necessárias foi desenvolvido, incluindo hard skills, soft skills, linguagens de programação, frameworks, bancos de dados, técnicas de machine learning e visualização de dados. O estudo conclui que, embora haja complementaridade nas habilidades requeridas pelas três perspectivas, as necessidades são extensas e heterogêneas, indicando a existência de diferentes grupos de cientistas de dados.

**Palavras-chave:** Cientista de dados; habilidades profissionais; Revisão sistemática; Processamento de linguagem natural

ISSN: 2594-8083


## 1 INTRODUÇÃO

O termo "cientista de dados" na lista de profissões foi criado em 2008 por DJ Patil e Jeff Hammerbacher para descrever os papéis de suas equipes no LinkedIn e Facebook e se tornou popular quando foi descrito como "o trabalho mais sexy do século 21" (Davenport & Patil, 2012). Os cientistas de dados extraem conhecimento acionável diretamente dos dados por meio de descoberta ou formulação de teste de hipóteses. Entretanto, não há consenso sobre a definição do que seja um cientista de dados e sobre quais as habilidades que estes profissionais apresentam e que são requeridas.

Entender esta problemática é fundamental para que haja alinhamento entre: i) academia sob a forma de universidades que ofertam cursos superiores; ii) organizações recrutadoras e iii) os profissionais, que buscam qualificação na academia e empregos nas organizações.

Assim, o objetivo deste trabalho é estudar a função de um cientista de dados sob a ótica da academia, dos empregadores brasileiros e dos profissionais, considerando o conjunto de suas habilidades e principais ferramentas utilizadas.

Trabalhos anteriores, como os de Chatfield et al. (2014), De Mauro et al. (2017), e Costa & Santos (2017), analisaram anúncios online de vagas de emprego e propuseram classificações de habilidades, formação e conhecimento sobre ferramentas requeridas dos cientistas de dados. Este trabalho avança sobre o realizado pelos trabalhos anteriores ao triangular diferentes fontes de dados e perspectivas. Para tanto, o estudo desenvolveu-se em três etapas metodológicas: i) revisão sistemática da literatura para obter a perspectiva da academia; ii) Análise documental de anúncios de emprego no LinkedIn, para capturar a perspectiva das organizações; iii) Levantamento de dados secundários sobre as atividades desempenhadas por cientistas de dados, para capturar a perspectiva dos profissionais da área.

Na revisão sistemática, as palavras-chave "*data scientist*", "*analytics*", "*data literacy*", "*skills*", "*ability*", "*competency*" e "*knowledge*", incluindo suas variações na língua portuguesa e no plural, quando aplicável, foram pesquisadas nas principais bases de dados ACM, EBSCO, Elsevier Science Direct, Google Scholar, IEEE, Proquest, Scopus e Web of Science. Dos 2.245 documentos identificados restaram 30 artigos, após aplicados critérios de inclusão e exclusão conforme Kitchenham (2004)

Os documentos para análise documental foram obtidos por "*web scrap*" de anúncios de vaga da plataforma LinkedIn coletadas com base nos quatro passos propostos por Nolan & Lang (2015) e com uso de palavras de busca identificadas na revisão sistemática da literatura: cientista de dados (português e inglês); ciência de dados (português e inglês); especialista em dados; mineração de dados; modelagem; *big data*; aprendizado de máquina; gestor de dados; engenheiro de dados (português e inglês); analista de dados; arquiteto de dados; "*Analytics*". Foram realizadas 6021

ISSN: 2594-8083


pesquisas de forma automatizada, semanalmente, de agosto/2020 a novembro/2020 considerando as vagas no Brasil, totalizando 9427 vagas. Após eliminação de vagas duplicadas, fora do escopo de atividade e do recorte geográfico, 1.308 vagas foram analisadas.

Do conteúdo dos anúncios – descrição do cargo, responsabilidades e requisitos – foram removidas "*stop words*" – como preposições, artigos e outra categorias de palavras sem valor semântico. Então, foi realizado um estudo da frequência das palavras utilizando a técnica de PLN (Processamento de Linguagem Natural) denominada *"bag-of-words*" (BOW) adequada para mineração de texto (Matsubara *et al.*, 2003), com auxílio da plataforma NLTK gerando a distribuição de frequência em dois grupos de informação: termos únicos e bigramas. Em seguida, o cálculo por meio do método TF-IDF (*Term Frequency-Inverse Document Frequency*) (Hu *et al.*,2018) foi utilizado para a extração de características do conteúdo das vagas, com auxílio da biblioteca *Scikit-learn* da linguagem Python.

Os dados secundários utilizados para capturar a perspectiva dos profissionais de ciência de dados foram obtidos do levantamento realizado pelo grupo Data Hackers denominado "Pesquisa de Mercado de Data Science" disponibilizada de forma anonimizada e não agregada no site Kaggle (DATAHACKERS, 2020). A pesquisa foi realizada com 1.765 respondentes e compreendeu 35 questões de pesquisa, das quais 7 são de interesse para este estudo, pois se referem a métodos, linguagens de programação, fontes de dados, opções de nuvem listadas, bancos de dados/fontes de dados, ferramentas de *business intelligence*, ferramenta de ETL utilizadas pelos respondentes em seus trabalhos.

## 2 DISCUSSÃO E ANÁLISE DOS DADOS

A revisão da literatura, a busca por vagas e a pesquisa com os profissionais revelaram resultados convergentes, em sua maioria, apresentando pequenas diferenças. Há casos como ferramentas de nuvem e infraestrutura tecnológica em que os dados de campo adicionaram na maioria dos casos com os dados de campo trazendo detalhes técnicos a habilidades descritas de forma genérica na literatura, como no caso de ferramentas de nuvem – do campo emergiram ambientes específicos como Amazon Web Services ("AWS"), Microsoft Azure e Google Cloud Platform – de infraestrutura tecnológica – identificado Kubernetes em campo, em adição à Docker. Ferramentas de versionamento como o GitHub, emergiram dos dados de campo, mas não figuravam na literatura e a ferramenta de visualização de dados mais citada nos dados de campo - Power BI – não apareceu na revisão de literatura.

Um modelo conceitual preliminar foi submetido à validação de especialistas resultando em um modelo representado pelo quadro da Figura 1. O quadro apresenta as habilidades divididas em hard skills (habilidades técnicas), soft skills (competências comportamentais) e as

ferramentas de linguagem de programação: ecossistemas, frameworks e bancos de dados, técnicas de machine learning e visualização de dados. Para cada uma dessas categorias, apresentam-se as que são mandatórias, diferenciais e desejáveis.

**Figura 1** - Conjunto de habilidades e ferramentas do cientista de dados

| **Conjunto de Habilidades e Ferramentas de Cientistas de Dados**<br>Castello, F:"DSST.v2022 – Data Scientist Skills & Tools 2022" | | | | |
|---|---|---|---|---|
| **Soft Skills** | Pensamento Analítico, Curiosidade, Proatividade, Capacidade de comunicação, Comportamento Ético e Espírito Empreendedor | | | |
| **Hard Skills** | Habilidade em Programação e Banco de Dados, Conhecimento do Negócio, Metodologias ágeis e Frameworks de Projetos de Dados, Criação de Modelos Preditivos e Machine Learning, Business Intelligence , Data Visualization, Cloud Computing | | | |
| **Ferramentas** | *Linguagens de Programação* | *Frameworks* | *Técnicas de Machine Learning* | *Visualização de Dados* |
| *Mandatório* | Python (Pandas, Scikit-learn) ou [R]. | SQL e framworks de projeto de ciência de dados, p. ex. CRISP-DM | Regressão linear e logística, clustering, Random Forest e Gradient Boost | "Matplotlib", "Seaborn" e "Dash" , "ggplot2" e "Shiny" |
| *Desejável* | Python (Tensorflow/ PyTorch), Scala, SQL (linguagem). | Hadoop/MapReduce, Spark, NoSQL | Redes neurais, inferência bayesiana, SVM, NLP, Visão Computacional | Microsoft PowerBI , Tableau, Qlik, Google Data Studio ou Looker |
| *Diferencial* | Matlab, C/C++, Julia, Pig. | MongoDB, Cassandra, Hbase, Snowflake, Cloudera, | Ensemble (Combinação de Classificadores) | Microstrategy, Pentaho |

**Fonte:** desenvolvido pelos autores (2023)

## 3 CONSIDERAÇÕES FINAIS

Conclui-se que há, em geral, alinhamento e complementaridade nas habilidades e ferramentas requeridas dos cientistas de dados nas perspectivas da academia, do mercado e dos profissionais. Fica evidente que as habilidades e ferramentas necessárias são extensas e heterogêneas, é provável que existam de fato grupos de cientistas de dados. No entanto, como a profissão é recente, os limites de tais grupos ainda estão em definição.

É necessário ter conhecimento específico do setor em que o profissional está inserido para impactar o negócio por meio de dados. De fato, "ter conhecimento de domínio (de negócios)" é apresentado como uma das habilidades obrigatórias.

Embora o Python e o R sejam citados na literatura com a mesma frequência, o Python surge como o 5º termo mais frequente presente no total das vagas analisadas. Assim, Python deve ser considerada a primeira linguagem a ser dominada na carreira de cientista de dados.

## REFERÊNCIAS

CHATFIELD, A. T. et al. **Data scientists as game changers in big data environments**. 2014.

COSTA, C., SANTOS, M. Y. (2017). **The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. International Journal of Information Management**, 37(6), 726-734.

ISSN: 2594-8083


DATA HACKERS. **Data Hackers Survey 2019**. 2020. Access in: 05 nov. 2020.

DAVENPORT, T.; PATIL, D. **Data Scientist**: The Sexiest Job of the 21st Century. Harvard Business Review, 2012.

DE MAURO, A. et al. **Human resources for Big Data professions**: A systematic classification of job roles and required skill sets. Information Processing & Management, v. 54, n. 5, p. 807-817, 2017.

KITCHENHAM, B. A. **Procedures for Performing Systematic Reviews**. Keele University, 2004.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. **Pretext**: A tool for pre-processing texts using the bag-of-words approach. Technical Report, 209(4), 2003.

NOLAN, D.; LANG, D. T. **Exploring Data Science Jobs with Web Scraping and Text Mining. In: Data Science** in R: A Case Studies Approach to Computational Reasoning and Problem Solving, 2015. p. 457