A Mapping of Challenges and Methods for Data Science Application

Antonio Frauzo Santos Moura
frauso2012@hotmail.com
Universidade Federal de Sergipe
Itabaiana, Sergipe, Brasil

Jessica Santos Portilio
jessica.santos.portilio@gmail.com
Universidade Federal de Sergipe
Itabaiana, Sergipe, Brasil

Methanias Colaço Júnior
mjrse@hotmail.com
Universidade Federal de Sergipe
Itabaiana, Sergipe, Brasil

ABSTRACT

Context. Data Science is described as an interdisciplinary field combining statistics, mathematics, computer science, and economics to extract valuable knowledge and insights from data. The article discusses the differences between Data Science and conventional statistics, the importance of Big Data, and the techniques used to analyze large volumes of data. The emerging profession of Data Scientist is also addressed, highlighting the skills needed to work with complex, unstructured data. **Objective**. The objective of this article was to characterize the challenges and methodologies of Data Science applications. **Methodology**. The methodology of this study involved a systematic mapping of the literature, providing a comprehensive overview of Data Science and highlighting development methods with a focus on reproducibility, transparency, and strategic alignment. The differences between Data Science and conventional statistics were analyzed to better understand the evolution of the field and the competencies required of professionals. **Results**. Of the 29 articles retrieved from scientific databases, 4 met the inclusion and exclusion criteria. It was observed that 50% of the publications occurred in 2022, indicating a growing interest from researchers in the field. Conferences represented the primary publication format, accounting for 75% of the works, while journals accounted for 25%. **Conclusion.** The main challenges identified include the difficulty in ensuring the reproducibility of studies, the misalignment between projects and organizational goals, and the lack of standards and frameworks for the Data Science project lifecycle. The research suggests the need for investment in education and training, as well as more methodologies to evaluate the impact of projects on business, aiming for greater return on investment.

CCS Concepts

• Mathematics of computing→Probability and statistics • Information systems→Systems→Database administration

→Database applications→Information storage and retrieval →Content analysis and indexin • Computing methodologies→Pattern recognition→Applications • Computer Applications→Administrative data processing.

KEYWORDS

Data Science, Big Data, Data Analysis, Data Mining, Statistics, Data Scientist, Research Methodology, Data Engineering, Reproducibility, Machine Learning.

1. INTRODUÇÃO

Atualmente, há muitos dados ao nosso redor que são armazenados, como os dados capturados pelos celulares, incluindo a localização e outras informações dos usuários [2]. O grande volume de informações geradas exigiu o surgimento de uma nova área: a Ciência de Dados, que estuda os dados desde sua produção até o descarte, abrangendo todo o ciclo de vida [1]. Em suma, a Ciência de dados refere-se à disciplina moderna que se dedica à análise de dados, integrando técnicas estatísticas, aprendizado de máquina e tecnologias de gerenciamento e mineração de dados para enfrentar os desafios trazidos pelo Big Data [2]. O objetivo principal é apurar e coletar informações relevantes sobre os dados de uma fonte determinada [15].

A expansão da era digital transformou a geração, coleta e armazenamento de dados, exigindo o desenvolvimento de novas áreas. Conceitos como Big Data, mineração e análise de dados surgiram para enfrentar os desafios impostos pelo grande volume de informações complexas [20].

Conjuntos de dados em grande escala, geralmente variando de terabytes a exabytes, são descritos pelo termo “Big Data”. Esses dados se destacam por sua diversidade, velocidade e volume, exigindo tecnologias avançadas para armazenamento, gerenciamento, processamento e análise, a fim de obter “insights” importantes. De acordo com [3], o Big Data abrange a exploração de dados estruturados e não estruturados, permitindo a obtenção de “insights” valiosos para impulsionar a inovação e a tomada de decisões estratégicas.

A análise de dados é uma etapa fundamental no processo de transformação de dados brutos em informações acionáveis. Conforme enfatizado por [16], a análise de dados permite identificar padrões, compreender correlações e extrair “insights” relevantes por meio de técnicas estatísticas, modelagem preditiva e visualização de dados. Gerar conhecimento e resolver problemas complexos em vários setores é, portanto, considerado uma prática crucial.

Assim, a Ciência de Dados é um campo interdisciplinar que combina conhecimento de estatística, matemática, ciência da computação e economia. Segundo [5], o objetivo da Ciência de Dados é extrair conhecimento e “insights” dos dados, utilizando métodos avançados. Isso envolve a coleta, limpeza, análise e interpretação de dados para resolver problemas complexos e fundamentar decisões com base em evidências sólidas.

Para esse novo campo, surge também uma nova profissão, o cientista de dados, que desempenha um papel central na aplicação

dos princípios da Ciência de Dados. Segundo [12], o cientista de dados possui técnicas avançadas e conhecimento estatístico, o que lhe permite dominar ferramentas e métodos para coletar, analisar e interpretar grandes conjuntos de dados. Esses profissionais utilizam de técnicas de aprendizado de máquina e programação para obter informações valiosas e orientar decisões baseadas em dados.

Com base na abordagem proposta neste artigo, este mapeamento teve como finalidade identificar desafios e técnicas de implementação em Ciência de Dados que assegurem reprodutibilidade, transparência e alinhamento estratégico. À medida que grandes volumes de dados são gerados e armazenados diariamente no mundo moderno, é essencial investigar a evolução da Ciência de Dados para responder a essas demandas e oportunidades. Este estudo também explora as diferenças entre Ciência de Dados e estatística convencional, enfatizando a importância dessa distinção no contexto atual. Analisar cuidadosamente esses conceitos permitirá uma melhor compreensão do papel e das habilidades exigidas dos profissionais que atuam nesse setor em constante mudança.

O restante deste artigo está organizado da seguinte forma. A seção 2 descreve como o mapeamento sistemático foi planejado. A seção 3 apresenta como foi a condução do mapeamento sistemático. Na seção 4, foram destacados os resultados obtidos depois da leitura. A seção 5 descreve a discussão sobre os resultados. A seção 6 apresenta as ameaças à validade encontradas e, finalmente, na seção 7, a conclusão é apresentada.

2. PLANEJAMENTO DO MAPEAMENTO SISTEMÁTICO

2.1 Objetivo

Este mapeamento teve como objetivo apresentar desafios e métodos de aplicação de Ciência de Dados que visem reprodutibilidade, transparência e alinhamento estratégico.

2.2 Questão de Pesquisa

O objetivo das questões propostas neste estudo é oferecer uma visão abrangente da área, ressaltando os principais aspectos dos estudos primários [10]. A elaboração das mesmas usou como base o modelo PICO, com o intuito de apresentar os efeitos de uma intervenção em uma população específica e organizar a pesquisa em quatro elementos importantes: População, Intervenção, Controle e "Outcomes" (Resultados) [17]. De acordo com [17], esses elementos podem ser utilizados para construir questionamentos de pesquisa com diferentes naturezas. A Tabela 1 apresenta o modelo PICO utilizado neste trabalho.

Tabela 1. Modelo PICO na estruturação de questões de pesquisa

Categorias	Descrição
População	Publicações sobre Ciência de Dados.
Intervenção	A aplicação de métodos, processos ou metodologias para o desenvolvimento de aplicações em Ciência de Dados.
Comparação	Controle: Métodos tradicionais tais como o CRISP-DM (Cross-Industry Standard Process for Data Mining).
Outcome (Resultado)	Os resultados, impactos ou benefícios da aplicação de métodos, processos ou metodologias em Ciência de Dados, incluindo aspectos tais como alinhamento estratégico, reprodutibilidade, transparência, vantagem competitiva e validação estatística.

Conforme a definição do modelo PICO, as questões de pesquisa foram construídas tomando como base as diretrizes do protocolo de Mapeamento Sistemático da Literatura e são apresentadas [10]. São elas:

- QP1: Quais as características principais da área de Ciência de Dados?
- QP2: Qual a caracterização dos métodos encontrados?
- QP3: Quais os meios de publicações mais populares?
- QP4: Em quais anos foram publicados mais artigos nesta área?
- QP5: Quais países têm mais publicações nesta área?

2.3 Estratégia de Busca e de Seleção

Foram selecionadas as bases de dados IEEE Xplore, ACM Digital Library e SciElo devido à sua relevância na área de computação e Ciência de Dados. A IEEE Xplore e a ACM Digital Library são reconhecidas internacionalmente por indexarem publicações de alta qualidade em tecnologia e computação. A SciELO foi incluída por ser uma biblioteca digital que disponibiliza acesso gratuito a uma ampla coleção de periódicos científicos, especialmente na América Latina e por indexar publicações com foco na área de gestão. Embora outras bases, tais como Scopus e Web of Science, também sejam importantes, o foco principal na Tecnologia da Informação e na Computação teve o propósito de averiguar como a área técnica tem lidado com o alinhamento estratégico das aplicações de Ciência de Dados.

Para a realização da busca, foram utilizadas as ferramentas de filtragem disponíveis em cada base de dados: título, resumo e palavras-chave. Os termos de busca foram definidos com base nos elementos do modelo PICO apresentados na Tabela 1. Para concretizar a pesquisa nas bases de dados digitais, foi estabelecida a seguinte string de busca composta por termos em inglês e sinônimos:

STRING DE BUSCA: (Data Science) AND (Process OR Method OR Methodology) AND (Business Strategy OR Strategic Alignment OR Reproducibility OR Transparency OR Competitive Advantage OR Statistical Validation).

2.4 Critérios de Seleção de Fontes

Para a seleção de estudos mais relevantes para o mapeamento sistemático e, ao mesmo tempo, para responder às perguntas apresentadas na seção 2.2, foram definidos critérios de inclusão e exclusão. Esses critérios são os seguintes:

Os Critérios de Inclusão:

(1) Requer que o estudo contenha os termos de busca em seu título, ou resumo;

(2) A pesquisa deve investigar artigos, com fundamentos ligados à Ciência de Dados, a sua relação com áreas periféricas tais como: análise de dados; Big Data; mineração de dados e estatística;

(3) Exige que o estudo esteja disponível para pesquisa online;

Os Critérios de Exclusão:

(1) Estudos publicados antes de 2013;

(2) Estudos sem relação com Ciência de Dados, ou estatística, ou análise de dados;

(3) Estudos de bases não-científicas;

(4) Estudos duplicados;

2.5 Estratégia de Extração de Informações

Para analisar a qualidade do estudo e abordar as questões de pesquisa apresentadas na seção 2.2, foi elaborado um questionário que deveria ser preenchido para cada artigo revisado.

Tabela 2. Formulário de Extração.

1.	Qual o tipo de estudo utilizado?	[Estudo de Caso, Revisão da Literatura, Roadmapping, Estudo Experimental]
2.	O método apresenta o passo a passo de como desenvolver uma aplicação?	[Sim, Não]
3.	Existe alinhamento estratégico?	[Sim, Não]
4.	O processo abrange a fase de engenharia de dados?	[Sim, Não]
5.	Foram declaradas as ameaças à validade?	[Sim, Não]

2.6 Validação do Escopo de Pesquisa

Foi realizada uma pesquisa preliminar para avaliar a relevância da Ciência de Dados no contexto atual, com foco especial nos métodos de sua aplicação. Desta forma, foram selecionados quatro artigos de controle com o objetivo de fundamentar um conjunto de palavras-chave sobre o tema e fornecer uma contextualização histórica. Entre esses artigos, os trabalhos de [6], [14] e [18] foram considerados falso-positivos em relação à string de busca adotada; ainda assim, remetem à área de Ciência de Dados, contribuindo para a familiarização com o jargão técnico e oferecendo suporte adicional na definição do domínio de palavras-chave e na contextualização histórica das questões de pesquisa.

Os artigos analisados foram:

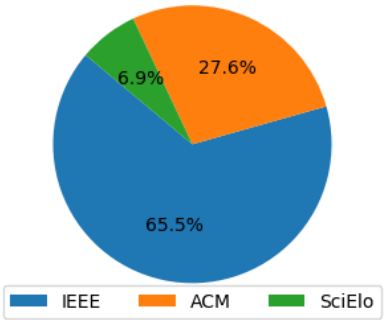
- [4] Colaço Júnior, Methanias; Cruz, Rodrigo Fontes; Araújo, Luciano Vieira de; Bliacheriene, Ana Carla; Nunes, Fátima de L. S. Evaluation of a process for the experimental development of data mining, AI and data science applications aligned with the strategic planning.
- [6] Fienberg, S. E. A brief history of statistics in three and one-half chapters: A review essay.
- [14] Martínez Plumed, Fernando; Contreras Ochando, Lidia; Ferri, Cèsar. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories.
- [18] Stigler, S. M. The history of statistics: The measurement of uncertainty before 1900.

3. CONDUÇÃO DO MAPEAMENTO SISTEMÁTICO

A sequência de pesquisa foi ajustada e implementada em cada um dos indexadores utilizados. Os mecanismos de busca foram configurados para identificar estudos a partir de 2013 que incluíssem, em seus títulos, resumos ou palavras-chave, ao menos um dos termos de pesquisa. No total, foram localizados 29 artigos: 19 (65,5%) do IEEE, 8 (27,6%) da ACM e 2 (6,9%) da SciElo,
conforme ilustrado na Figura 1.

Figura 1: Realização da Consulta de String de Busca em Repositórios Digitais

Após obter os artigos nas bases de dados, iniciou-se o processo de triagem, utilizando os critérios de seleção estabelecidos na seção



2.4. Cada artigo foi classificado como “Aceito” ou “Rejeitado”. Do total de 29 artigos analisados, 23 (79,3%) não atenderam aos critérios de inclusão e foram classificados como “Rejeitados”. Os artigos restantes foram considerados como “Aceitos” para uma análise mais detalhada, conforme ilustrado na Figura 2.

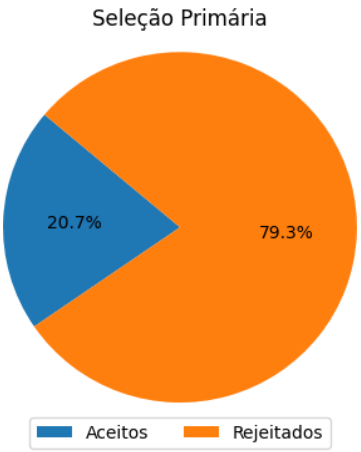


Figura 2: Etapa Inicial de Seleção dos Estudos

Com o objetivo de verificar se os artigos restantes utilizam abordagens para avaliar a relevância da Ciência de Dados no contexto atual e examinar as questões de investigação, o método e o processo seguiram a leitura completa dos textos. Dos 6 artigos selecionados para avaliação, após a leitura completa, 4 foram escolhidos. A escolha dos quatro artigos finais foi conduzida a partir da aplicação de cinco questões de qualidade, utilizadas para avaliar a consistência metodológica e a relevância das publicações em relação aos objetivos do estudo. As questões consideraram a clareza dos objetivos, a descrição do método empregado, a adequação das técnicas de análise, a coerência entre resultados e conclusões e a aplicabilidade das contribuições no campo da Ciência de Dados. Cada artigo foi avaliado de forma binária (sim/não) em cada critério, sendo estabelecido como ponto de corte a obtenção de pelo menos quatro respostas positivas. Dessa forma, os estudos selecionados apresentaram alinhamento metodológico e teórico satisfatório, garantindo a validade dos resultados e a representatividade do conjunto analisado.
A Figura 3 apresenta o resultado da seleção secundária.

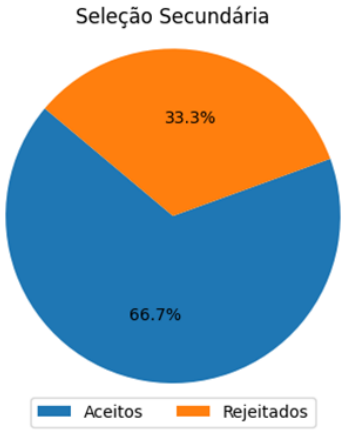


Figura 3: Etapa Secundária da Seleção de Estudos

4. SÍNTESE DOS DADOS E APRESENTAÇÃO DOS RESULTADOS

Esta seção apresenta os resultados do mapeamento sistemático. A Figura 4 ilustra o fluxo que descreve o processo de extração dos artigos em cada fase, seguido das respostas às questões de pesquisa com base nos dados obtidos.

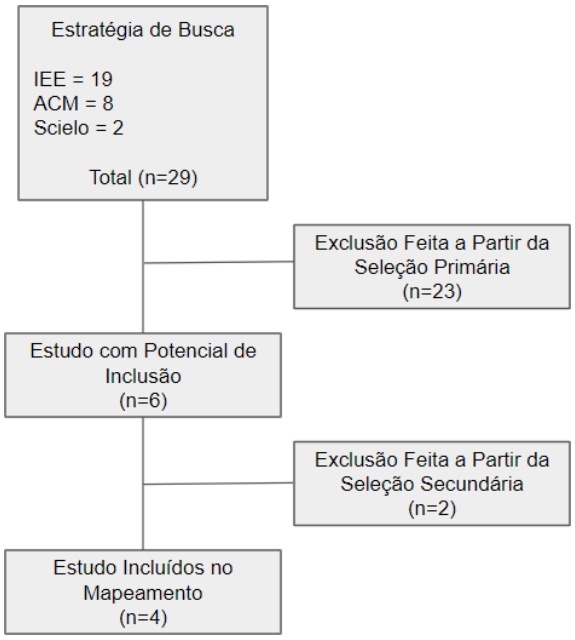


Figura 4: Diagrama Prima mostrando a extração dos Dados

4.1 Quais as características principais da área de Ciência de Dados?

A Ciência de Dados é fundamental para entender a análise estrutural do grande volume de dados presentes no ambiente empresarial. Com a evolução das organizações e o aumento de dados não estruturados, torna-se evidente que boa parte desses dados não é planejada de forma organizada. Dessa maneira, a quantidade, a variedade e a velocidade impedem uma análise manual. A Ciência de Dados lida com dados de diferentes fontes e formatos, o que não era o foco principal da estatística [9].

O termo 'Ciência de Dados' surgiu para abranger um campo mais amplo do que o tradicionalmente coberto pela estatística. Embora a estatística seja uma parte fundamental da Ciência de Dados, o termo reflete a necessidade de lidar com o grande volume, a variedade e a velocidade dos dados modernos, aspectos que a estatística tradicional, focada principalmente em amostras e inferência, não abrange completamente. Com o crescimento exponencial de dados digitais, novas ferramentas e técnicas, como aprendizado de máquina e Big Data, tornaram-se essenciais. Assim, a Ciência de Dados surgiu como um campo interdisciplinar que vai além da análise estatística, integrando computação avançada, gestão de dados e modelagem preditiva para resolver problemas reais em diversas áreas [9].

A análise científica de dados origina-se do fenômeno denominado Big Data. Em outras palavras, refere-se à coleta, processamento e à avaliação de grandes volumes de dados, permitindo extrair "insights" e informações importantes. Segundo [4], o Big Data é composto de 9 Vs: volume, velocidade, variedade, viscosidade, virilidade, visualização, veracidade, validade e valor. Essas características fazem do Big Data uma ferramenta extremamente eficaz em abordagens de Business Intelligence, permitindo a

obtenção de “insights” valiosos e auxiliando na tomada de decisões estratégicas.

Portanto, numerosos desafios enfrentados no mundo real apresentam uma complexidade que transcende a capacidade de compreensão por meio de abordagens estatísticas tradicionais. A disciplina da Ciência de Dados utiliza métodos avançados de aprendizado de máquina e inteligência artificial para extrair discernimentos valiosos de conjuntos de dados complexos e não lineares [9].

Desta forma, enquanto a estatística se concentra na análise de dados estruturados por meio de técnicas matemáticas para inferência e teste de hipóteses, a Ciência de Dados amplia esse campo ao integrar métodos de computação avançada, aprendizado de máquina e gestão dos dados modernos, que frequentemente são massivos, não estruturados e provenientes de diversas fontes.

Neste contexto, o pensamento estatístico atravessa muitos outros campos científicos atuais, e a sua história é um aspecto importante para o desenvolvimento científico mais amplo. A estatística possui uma trajetória extensa que se inicia nos tempos antigos, com os primeiros indícios de métodos estatísticos observados em civilizações como a Babilônia, Egito e China. Contudo, o estabelecimento formal da estatística como uma disciplina científica ocorreu mais recentemente [6].

Já [18] aponta que, embora a estatística tenha sido usada em áreas como a demografia, a economia e as ciências sociais, o desenvolvimento teórico robusto que permitiu sua aplicação em diversas disciplinas só ocorreu no século XIX, quando os matemáticos começaram a formalizar as ideias sobre variabilidade e incerteza. Ele destaca a gradual construção da estatística como um campo científico interdisciplinar, consolidado como ciência independente no final do século XIX e início do século XX.

Na perspectiva de [6], a estatística é a disciplina científica dedicada à coleta, organização e interpretação de dados, além de se dispor a relatar dados que representam observações derivadas do mundo real.

Segundo [18], os estatísticos tendem a ver a história da ciência como algo que gira em torno da medição e do raciocínio estatístico. A estatística moderna oferece uma tecnologia quantitativa para a ciência empírica, além da lógica e metodologia para medir incertezas e avaliar suas consequências durante o planejamento, experimentação e observação [18].

A origem moderna da estatística pode ser traçada até o século XVII, com contribuições significativas de pensadores como John Graunt, William Petty e Blaise Pascal. Graunt, um comerciante e demógrafo inglês, é frequentemente considerado o pai da estatística demográfica por seu trabalho pioneiro na análise de dados populacionais [6]. A estatística moderna, no seu estágio atual, complementa o processo de extração de valor a partir de dados, os quais, pela natureza heterogênea e volumosa, precisam de um verdadeiro processo de engenharia para que se tornem acessíveis e precisos.

Neste sentido, a engenharia de dados enfrenta desafios significativos, como a credibilidade e a incerteza dos dados não estruturados. Segundo [8], quando se trata dos modelos de Ciência de Dados, é fundamental manter a documentação constante do projeto e de um repositório de conhecimento para garantir a reprodutibilidade e rastreabilidade de toda a análise realizada.

Em outras palavras, um aspecto relevante da engenharia de dados é a proveniência dos dados, que se refere à capacidade de localizar a origem e o desenvolvimento dos dados com o decorrer do tempo. É essencial garantir a reprodutibilidade confiável dos modelos de aprendizado de máquina, realizando o controle de fatores que podem afetar a replicação dos resultados [19].

4.2 Qual a caracterização dos métodos encontrados?

Foi utilizado um formulário de extração (Tabela 2) para filtrar artigos deste trabalho, pois, em uma pesquisa, essa é uma estratégia eficaz para garantir a relevância e a qualidade dos dados coletados. Esse método permite selecionar artigos com base em critérios específicos, como palavras-chave, metodologia ou datas de publicação, agilizando o processo de revisão e análise. Além disso, ao estruturar o formulário com campos bem definidos, facilita-se a comparação entre os estudos e a identificação de padrões ou lacunas na literatura, tornando o processo de revisão mais rigoroso e organizado.

A seção em questão foi reformulada para alinhar-se ao título e esclarecer a caracterização dos métodos adotados. O estudo utilizou um mapeamento sistemático da literatura (SLR) como principal abordagem metodológica, seguindo critérios definidos de busca, seleção e exclusão de artigos com base na relevância, reprodutibilidade e aderência ao tema. Na Tabela 3, os resultados evidenciam os principais desafios e métodos identificados, permitindo observar tendências metodológicas e lacunas de pesquisa. Os dois artigos excluídos foram retirados por não apresentarem metodologia claramente descrita nem relação direta com os objetivos centrais do estudo, o que comprometeria a consistência e a validade da análise final.

Tabela 3. Caracterização dos Artigos.

Artigo	Qual o tipo de estudo utilizado?	O método apresenta o passo a passo de como desenvolver?	Existe alinhamento estratégico?	O processo abrange a fase de engenharia de dados?	Foram declaradas as ameaças à validade?
Wonsil, Seltzer (2023)	Estudo Experimental	Sim	Sim	Não	Não
Haertel et al. (2022)	Revisão da Literatura	Sim	Sim	Sim	Não
Júnior et al. (2022)	Estudo de Caso	Sim	Sim	Sim	Não
Kayabay et al.(2020)	Roadmapping	Sim	Sim	Sim	Não

4.3 Quais os meios de publicações mais populares?

A Figura 5 exibe os trabalhos selecionados de acordo com o tipo de publicação.

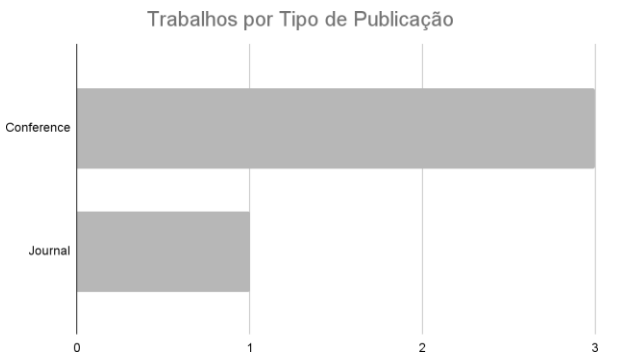


Figura 5: Classificação dos Trabalhos Selecionados por Categoria de Publicação

4.4 Em quais anos foram publicados mais artigos nesta área?

A Figura 6 apresenta os trabalhos selecionados por ano de publicação. Nota-se que a maior parte dos artigos foi publicada no ano de 2022.

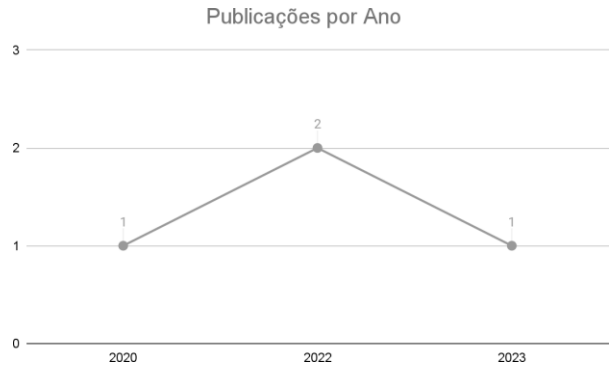


Figura 6: Distribuição dos Artigos Selecionados por Ano de Publicação.

4.5 Quais países têm mais publicações nesta área?

A Figura 7 exibe os países que publicaram trabalhos relacionados ao tema explorado neste mapeamento sistemático. Nenhum país se destacou, uma vez que cada um contribuiu com apenas uma publicação.

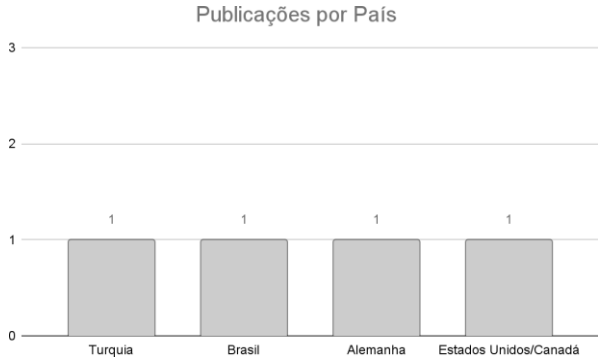


Figura 7: Distribuição de Publicações por País.

5. SÍNTESE NARRATIVA

Esta seção apresenta os aspectos fundamentais e aprendizados identificados nos artigos analisados. O maior número de publicações de artigos sobre o tema foi registrado em 2022, sugerindo que essa área de pesquisa é recente. Portanto, os resultados indicam que as publicações sobre o tema em questão estão distribuídas por diversos países, evidenciando que a Ciência de Dados é um assunto de interesse global entre pesquisadores.

Na pesquisa de [9], é descrita a técnica de "roadmapping", ou elaboração de roteiros estratégicos, para enfrentar os desafios da Ciência de Dados. O trabalho explora o desenvolvimento do "Data Science Roadmapping" (DSR), baseado na técnica de "roadmapping" tecnológico, com o objetivo de auxiliar as organizações a iniciar um "roadmapping" para projetos de Ciência de Dados. A aplicação de um DSR em uma empresa contribui diretamente para a produção de dados que impactam positivamente os negócios. Nesse contexto, os autores discutem os obstáculos que as organizações enfrentam ao implementar Big Data e Ciência de Dados, destacando barreiras tecnológicas, culturais e organizacionais. As empresas lidam com grandes volumes de dados e enfrentam desafios relacionados à diversidade e velocidade das informações, à falta de ferramentas adequadas, à ausência de infraestrutura escalável e à escassez de profissionais qualificados. Além disso, existe uma resistência cultural ao uso de dados nas tomadas de decisões e dificuldades na integração do Big Data aos processos empresariais. O "roadmapping" foi customizado para cada organização, ajustando escopo e metas em camadas, alinhadas às tendências do setor e aos recursos necessários.

Em [4] é ressaltada a importância da experimentação na análise dos algoritmos empregados na mineração de dados, peça-chave da Ciência de Dados, uma vez que envolve a identificação de padrões e "insights" a partir de vastos conjuntos de dados. A Ciência de Dados, por sua vez, é apresentada como uma disciplina voltada à extração de conhecimento e valor a partir dos dados, utilizando métodos e técnicas estruturadas. O destaque deste método, além da experimentação, que é uma peça-chave nos Planejamentos Estratégicos flexíveis e modernos, é a

preocupação com o alinhamento dos projetos com os objetivos organizacionais.

[19] discutem a reprodutibilidade em modelos de aprendizado de máquina, aspecto fundamental da Ciência de Dados. A reprodutibilidade permite que os resultados obtidos em experimentos de Ciência de Dados sejam verificados e validados por outros pesquisadores, aumentando a confiança nas conclusões tiradas a partir dos dados analisados. Nesse contexto, o trabalho mostra como o sistema MERIT, integrado à biblioteca Tribuo, facilita a coleta de procedência e o controle de variáveis em ambientes de aprendizado de máquina. Isso é essencial para a Ciência de Dados, pois garante a reprodutibilidade dos experimentos, permitindo a validação de modelos e métodos. Além disso, promove a transparência ao documentar os processos de modelagem e facilita a colaboração entre pesquisadores, aspecto indispensável em equipes multidisciplinares que enfrentam desafios complexos.

6. AMEAÇAS À VALIDADE

Validade de Construção: A validade de construção refere-se à adequação das medidas utilizadas para representar os conceitos estudados. No nosso caso, a string de busca e as questões de pesquisa podem não ter capturado todos os aspectos relevantes da Ciência de Dados. Para mitigar essa ameaça, realizamos uma revisão preliminar da literatura para identificar os termos mais utilizados na área. Além disso, testamos a string de busca em cada base de dados e a ajustamos conforme necessário para melhorar sua abrangência. Reconhecemos, contudo, que termos emergentes ou menos comuns podem ter sido omitidos.

Validade Interna: A validade interna está relacionada à capacidade de estabelecer relações causais confiáveis. Como este é um mapeamento sistemático, uma ameaça é a possível subjetividade na seleção e análise dos estudos. Para minimizar esse risco, seguimos rigorosamente a metodologia proposta por
[10] e [9], incluindo a aplicação independente dos critérios de inclusão e exclusão por dois pesquisadores, com discussão e consenso em casos de divergência.

Validade Externa: A validade externa refere-se à generalização dos resultados para outros contextos. Nossa pesquisa pode estar limitada pela escolha das bases de dados e pelo período de publicação. Para melhorar a validade externa, selecionamos bases reconhecidas internacionalmente e definimos critérios de inclusão que consideram a relevância atual dos estudos. No entanto, admitimos que estudos relevantes publicados em outras bases ou em idiomas diferentes do inglês e português podem não ter sido incluídos, limitando a generalização dos achados.

7. CONCLUSÃO

Esta pesquisa teve como objetivo destacar os desafios e métodos de aplicação de Ciência de Dados que promovam a reprodutibilidade, transparência e alinhamento estratégico. Destacamos sua importância na era digital, em que grandes volumes de dados são gerados constantemente. Dos 29 resultados obtidos após a busca nas bases científicas, utilizando a string exposta na seção 2.3, 4 foram aceitos pelos critérios de inclusão e exclusão, dos quais 50% foram publicados em 2022. Esse aspecto demonstra a tendência crescente no campo, com o aumento do interesse de pesquisadores em relação ao tema. Entre os principais meios de publicação, destacaram-se as “conferences”, com 3 (75%) dos trabalhos, enquanto os “journals” foram representados por apenas 1 artigo (25%).

O trabalho em Ciência de Dados enfrenta várias dificuldades que impactam a implementação e o desenvolvimento de projetos. A

reprodutibilidade é uma preocupação importante, visto que muitos estudos não conseguem replicar seus resultados devido a variáveis não controladas e à falta de documentação adequada. A falta de alinhamento entre os projetos de Ciência de Dados e os objetivos das organizações leva ao desperdício de recursos e à falta de autenticidade, resultando em uma falta de colaboração limitada entre equipes de Ciência de Dados e líderes empresariais. A ausência de estruturas e padrões para o ciclo de vida da Ciência de Dados causa discrepâncias em projetos, dificultando a colaboração e a transferência de conhecimento. Além disso, a complexidade dos dados, principalmente, quando desorganizados, exige habilidades avançadas de pré-processamento e manipulação, aumentando os desafios no campo.

Portanto, este estudo oferece contribuições valiosas para a academia, servindo como referência para futuras pesquisas na área de Ciência de Dados. Para enfrentar as dificuldades observadas, sugere-se que estudos futuros abordem a educação e o treinamento em Ciência de Dados, avaliando a eficácia de programas de formação para garantir que os profissionais estejam adequadamente preparados para os desafios do campo. Além disso, é essencial comparar metodologias existentes para assegurar a reprodutibilidade dos estudos, identificando pontos fortes e fracos de cada abordagem. Outra sugestão é desenvolver métricas para avaliar o impacto dos projetos de Ciência de Dados nos negócios, garantindo o retorno sobre o investimento.

Essas recomendações buscam aprimorar a prática da Ciência de Dados, assegurando um maior alinhamento com os objetivos organizacionais e promovendo uma colaboração mais eficaz entre equipes de Ciência de Dados e líderes empresariais.

8. REFERÊNCIAS

[1] Amaral, Fernando. Introdução à ciência de dados: mineração de dados e big data. Alta Books Editora, 2016.

[2] C, Luís. Big data e data science. Boletim da APDIO, p. 11-14, 2014.

[3] Chen, H.; Chiang, R. H. L.; Storey, V. C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, v. 36, n. 4, p. 1165-1188, 2012.

[4] Colaço Júnior, Methanias; Cruz, Rodrigo Fontes; Araújo, Luciano Vieira de; Bliacheriene, Ana Carla; Nunes, Fátima de L. S. Evaluation of a process for the experimental development of data mining, AI and data science applications aligned with the strategic planning. Journal of

Information Systems and Technology Management – Jistem USP, v. 19, 2022, e202219018. DOI= https://doi.org/10.4301/S1807-1775202219018.

[5] Dhar, V. Data science and prediction. Communications of the ACM, v. 56, n. 12, p. 64-73, 2013.

[6] Fienberg, S. E. A brief history of statistics in three and one-half chapters: A review essay. Statistical Science, v. 7, n. 2, p. 208-255, 1992.

[7] Gomes, W.; Colaço Júnior, M. Applications of Artificial Intelligence for Auditing and Classification of Incongruent Descriptions in Public Procurement. In: Proceedings of the XVIII Brazilian Symposium on Information Systems (SBSI), 2022. DOI= https://doi.org/10.1145/3535511.3535551.

[8] Haertel, C.; Pohl, M.; Turowski, K.; Staegemann, D. Project Artifacts for the Data Science Lifecycle: A Comprehensive Overview. 2022.

[9] Kayabay, K.; Gökalp, M. O.; Gökalp, E.; Eren, P. E.; Koçyigit, A. Data Science Roadmapping: Towards an Architectural Framework. 2020.

[10] Kitchenham, B. Procedures for performing systematic reviews. Keele, UK: Keele University, 2004.

[11] Kitchenham, B. et al. The impact of limited search procedures for systematic literature reviews — A participant-observer case study. In: 3rd International Symposium on Empirical Software Engineering and Measurement, Lake Buena Vista, FL, 2009.

[12] Loukides, M. What is data science?. O'Reilly Media, 2011.

[13] Neves, D. F.; Santos, L. C.; Santos, M. A.; Júnior, M. C.; Júnior, M. C. Governança Colaborativa em Instituições de Ensino: Uma Revisão Quasi-Sistemática da Literatura. RENOTE, 2019.

[14] Martínez Plumed, Fernando; Contreras Ochando, Lidia; Ferri, Cèsar. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, v. PP, n. 99, p. 1-1, dez. 2019. DOI: 10.1109/TKDE.2019.2962680.

[15] Pimentel, João Felipe et al. Ciência de dados com reprodutibilidade usando jupyter. Sociedade Brasileira de Computação, 2021.

[16] Provost, F.; Fawcett, T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, 2013.

[17] Santos, C. M. da C.; Pimenta, C. A. de M.; Nobre, M. R. C. A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências. Revista Latino-Americana de Enfermagem, v. 15, n. 3, p. 508-511, 2007. DOI= https://doi.org/10.1590/S0104-11692007000300023.

[18] Stigler, S. M. The history of statistics: The measurement of uncertainty before 1900. Belknap Press of Harvard University Press, 1986.

[19] Wonsil, J.; Seltzer, M. Integrated Reproducibility with Self-describing Machine. 2023.

[20] Zikopoulos, P.; Eaton, C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2011.