24º ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
Perspectivas Contemporâneas na Ciência da Informação
• Vitória - ES • Ancib • PPGCI/UFES

XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL APLICADAS À CURADORIA AUTOMATIZADA DE DADOS

ARTIFICIAL INTELLIGENCE TOOLS APPLIED TO AUTOMATED DATA CURATION

Denise Fukumi Tsunoda – Universidade Federal do Paraná (UFPR)
André José Ribeiro Guimarães – Universidade Federal do Paraná (UFPR)

Modalidade: Trabalho Completo

Resumo: O crescimento exponencial da disponibilidade dos dados digitais tem gerado uma demanda crescente por soluções que permitam a curadoria eficiente e automatizada de grandes volumes de dados. Para mensurar o interesse da comunidade científica em ferramentas de IA para curadoria automatizada de dados, foram analisados 277 documentos publicados entre 2003 e 2023. O primeiro documento, de 2003, da área de bioinformática, e desde então, o número de publicações cresceu 19,75% ao ano, atingindo 65 em 2023. Este artigo revisa e compara ferramentas de Inteligência Artificial (IA) que podem ser utilizadas no suporte à curadoria automatizada de dados, incluindo GPT-4, IBM Watson, Google Cloud AutoML, Amazon Sagemaker, DataRobot, Alteryx, Microsoft Azure Machine Learning e SAS Viya. As ferramentas são avaliadas e comparadas com base em critérios tais como inovação tecnológica, versatilidade, facilidade de uso, integração com outros sistemas e eficácia. Cada uma oferece um conjunto específico de funcionalidades que contempla desde a limpeza e organização de dados até a análise preditiva e a extração de informações relevantes. Conclui-se que a escolha da ferramenta depende das necessidades específicas de cada aplicação e, muitas vezes, é necessário o uso de mais de uma ferramenta para que seja realizado o processo de curadoria de dados.

Palavras-chave: curadoria de dados; ferramentas de IA; automação de dados.

Abstract: The exponential growth in the availability of digital data has generated a growing demand for solutions that enable the efficient and automated curation of large volumes of data. To measure the scientific community's interest in AI tools for automated data curation, we analyzed 277 documents published between 2003 and 2023. The first document, from 2003, came from the field of bioinformatics, and since then, the number of publications has grown by 19.75% per year, reaching 65 in 2023. This article reviews and compares Artificial Intelligence (AI) tools that can support automated data curation, including GPT-4, IBM Watson, Google Cloud AutoML, Amazon Sagemaker, DataRobot, Alteryx, Microsoft Azure Machine Learning, and SAS Viya. We evaluated and compared these products based on criteria such as technological innovation, versatility, ease of use, integration with other systems, and effectiveness. Each offers a unique set of functionalities, ranging from data cleaning and organization to predictive analysis and the extraction of relevant information. In conclusion, the choice

of the right tool depends on the specific needs of each application, and it is often necessary to use more than one tool to conduct the data curation process.

Keywords: data curation; AI tools; data automation.

1 INTRODUÇÃO

Nos últimos anos, o aumento exponencial de dados gerados por empresas e usuários, a velocidade de compartilhamento das bases de dados, o incentivo à padronização e disponibilização dos dados para acesso a quem interessar possa têm revolucionado a forma como o dado é coletado, selecionado, armazenado, processado e acessado (utilizado). Este fenômeno, conhecido como "Big Data", apresenta tanto desafios quanto oportunidades para diversos setores, uma vez que transformar grandes volumes de dados brutos em informações úteis e acionáveis é interesse da maioria das organizações, sejam públicas, privadas ou do terceiro setor.

A Curadoria de Dados (CD) envolve a identificação de fontes de dados relevantes, extração ou coleta, limpeza, organização, integração, enriquecimento, validação e manutenção dos dados, visando garantir sua qualidade e utilidade (Beheshti *et al.*, 2017). Por meio da CD, é possível adotar um suporte metodológico e tecnológico para a gestão de dados, o que permite abordar problemas relacionados à qualidade, maximizando sua usabilidade (Freitas; Curry, 2016). De maneira geral, a CD é o processo que os transforma os dados brutos em dados prontos para análise, fornecendo uma camada de abstração que isenta os usuários de tarefas demoradas, tediosas e propensas a erros (Beheshti *et al.*, 2018).

A curadoria de dados e a ciência da informação estão interligadas por seu foco comum no gerenciamento e uso eficiente de dados e informações. Segundo Tenopir, Birch e Allard (2012, p. 7), durante as fases do ciclo de vida dos dados, como captura, processamento, preservação e compartilhamento, a curadoria desempenha um papel crucial na aplicação de boas práticas de gestão, garantindo a reutilização e disseminação dos dados de forma ética e eficaz. Os autores ainda pontuam que a ciência da informação abrange o estudo de como a informação é coletada, organizada, recuperada e utilizada, desenvolvendo metodologias para melhorar seu acesso e uso. Ambas as áreas atuam de maneira interdisciplinar, sendo aplicáveis a contextos como bibliotecas digitais, big data e repositórios acadêmicos, com ênfase na preservação e sustentabilidade da informação. Em resumo, a curadoria de dados é uma aplicação prática dos princípios da ciência da informação, especialmente em contextos

de grandes volumes de dados. Além disso, conforme Tenopir, Birch e Allard (2012), as bibliotecas acadêmicas já desempenham um papel relevante na curadoria de dados, refletindo como a ciência da informação contribui para a preservação e uso estratégico dos dados em ambientes acadêmicos.

No entanto, quando realizadas manualmente, as tarefas relacionadas à CD são, além de demoradas, suscetíveis a erros humanos, inconsistências e vieses. Com os avanços na Inteligência Artificial (IA), têm surgido ferramentas capazes de automatizar esse processo, tornando a curadoria de dados mais rápida, precisa e escalável. As soluções de IA são projetadas para lidar com diversas tarefas alinhadas à curadoria, como a limpeza (ou higienização) de dados, extração de informações relevantes, categorização de conteúdos e geração de metadados. Nesse contexto, levanta-se as questões: como as principais ferramentas de inteligência artificial podem ser utilizadas para curadoria automatizada de dados? Qual o atual panorama das pesquisas científicas que relacionam ferramentas de IA à curadoria de dados?

Para responder a essas questões, este artigo tem como objetivo identificar como as ferramentas de inteligência artificial podem ser aplicadas à curadoria automatizada de dados. Além de identificar as ferramentas, é realizada uma comparação entre as soluções encontradas, com base em seu conjunto de funcionalidades, especialmente aquelas relacionadas aos diferentes aspectos da curadoria de dados. A seleção das ferramentas foi baseada em critérios como inovação tecnológica, versatilidade, facilidade de uso, integração com outros sistemas e eficácia comprovada, aspectos que serão discutidos detalhadamente. Nas seções seguintes, cada ferramenta será apresentada, com uma avaliação de suas capacidades e aplicações práticas. Ademais, o Quadro 1 foi elaborado para auxiliar a compreensão das potencialidades, limitações e possíveis aplicações empíricas de cada ferramenta em diferentes contextos.

Para oferecer um panorama abrangente das pesquisas científicas sobre o tema, a investigação foi complementada com uma busca na base de trabalhos acadêmicos da Lens. Ao final do artigo, espera-se que a combinação das estratégias metodológicas (pesquisa em base de periódicos científicos e análise e comparação de ferramentas) forneça uma visão crítica e detalhada do estado atual da curadoria automatizada de dados, destacando as tendências emergentes e os desafios futuros.

2 ENCAMINHAMENTOS METODOLÓGICOS

A primeira etapa da pesquisa envolveu a realização de uma busca no indexador Lens (2024), uma ferramenta que oferece acesso a um *corpus* global de metadados da literatura acadêmica, com indexação de citações. A escolha pela base de dados Lens se deve à amplitude de seu catálogo, que, com mais de 200 milhões de registros, destaca-se como um dos maiores indexadores de trabalhos acadêmicos disponíveis (Penfold, 2020). Além disso, a Lens é uma iniciativa gratuita, voltada para a democratização do acesso à informação, tanto acadêmica quanto de patentes.

Para a busca, realizada em 26 de junho de 2024, foi utilizada a expressão *((("automating" OR "automated") AND ("data curation" OR "digital curation")) AND ("artificial intelligence" OR "machine learning")) AND (tool OR tools)*, restrita a documentos de cunho acadêmico, sem a aplicação de filtros adicionais, como recorte temporal ou tipo de documento científico. A pesquisa recuperou 277 documentos, os quais foram submetidos a uma análise bibliométrica para identificar padrões na produção, publicação e comunicação de informações (Diodato, 2013). A finalidade dessa etapa não foi analisar o conteúdo dos documentos em si, mas sim identificar a evolução do tema em número de publicações, os países que mais pesquisam o assunto, as referências mais citadas, os autores e periódicos mais relevantes, além de outros indicadores observáveis.

A principal ferramenta utilizada na extração dos indicadores bibliométricos foi o pacote Bibliometrix da linguagem R (Aria; Cuccurullo, 2017), por meio da sua interface gráfica Biblioshiny (Aria; Cuccurullo, 2022), considerada uma ferramenta bastante completa e abrangente para esse tipo de análise (Moreira; Guimarães; Tsunoda, 2020). No entanto, como os dados originais exportados pela Lens não continham informações sobre os países dos documentos, utilizou-se o pacote openalexR (Aria *et al.*, 2024) para acessar a API do OpenAlex, um serviço aberto e gratuito que fornece metadados sobre milhões de trabalhos (artigos de periódicos, livros etc.) (Priem; Piwowar; Orr, 2022). O acesso foi realizado por meio do identificador DOI (*Digital Object Identifier*), presente em 275 dos documentos identificados inicialmente, permitindo, assim, o acesso a estatísticas relacionadas aos países.

Em seguida, foi necessário definir os critérios de seleção das ferramentas de IA para curadoria automatizada de dados a serem analisadas. O filtro inicial foi baseado no Quadrante Mágico do Gartner de 2024 para plataformas de Ciência de Dados e Machine Learning (CDML),

ferramenta voltada a ajudar profissionais de dados a selecionar as plataformas mais adequadas para seus trabalhos (Jaffri *et al.*, 2024). Do Quadrante Mágico do Gartner de 2024, foram selecionadas as plataformas identificadas como "líderes", apresentadas na Figura 1, que indicam as soluções com maior destaque no mercado. Além dessas, foi incluído o GPT-4 devido à sua popularidade de uso e recente ascensão.

Figura 1 – Quadrante Mágico do Gartner para Ciência de Dados e *Machine Learning*

Fonte: Extraído de Gartner (Jaffri *et al.*, 2024).

As ferramentas incluídas na pesquisa:

- são reconhecidas por suas capacidades tecnológicas e por já terem apresentado inovações no campo da IA e curadoria de dados;
- são aplicáveis em uma ampla gama de setores, desde mídia e comércio eletrônico até finanças e governo;
- contam com vasta documentação disponível, suporte técnico e comunidades ativas para auxiliar na obtenção de ajuda;
- Apresentam histórico comprovado de implementações reais.

Ainda que não tenha sido um critério formalmente definido, algumas das ferramentas permitem acesso e uso gratuito por pesquisadores interessados e serão destacadas nos resultados.

Para a comparação das ferramentas, foram selecionados critérios com potencial para fornecer visão abrangente e comparativa das ferramentas, levando em consideração não

apenas suas capacidades técnicas, mas também a acessibilidade e a aplicabilidade prática em diferentes contextos. Os critérios utilizados foram:

- **descrição:** visão geral e seu principal objetivo ou funcionalidade;
- **funcionalidades:** lista de alguns recursos oferecidos pela ferramenta;
- **usabilidade:** facilidade de uso e a curva de aprendizado;
- **aplicações:** exemplos de uso da ferramenta em situações reais;
- **distribuição:** como a ferramenta é disponibilizada, como SaaS (*software as a ervisse*), software desktop, API etc.;
- **linguagem de programação:** linguagens de programação suportadas pela ferramenta para personalização e uso avançado (quando for o caso);
- **acesso:** *link* (URL) de acesso para mais informações sobre a ferramenta;
- **educação**: indica se há acesso gratuito ou planos especiais para pesquisadores, docentes e discentes;
- **vantagens**: principais vantagens percebidas, a exemplo de interface amigável, fluxos bem definidos e outros;
- **desvantagens**: principais desvantagens percebidas, a exemplo do custo, limitações em personalizações das análises e outros.

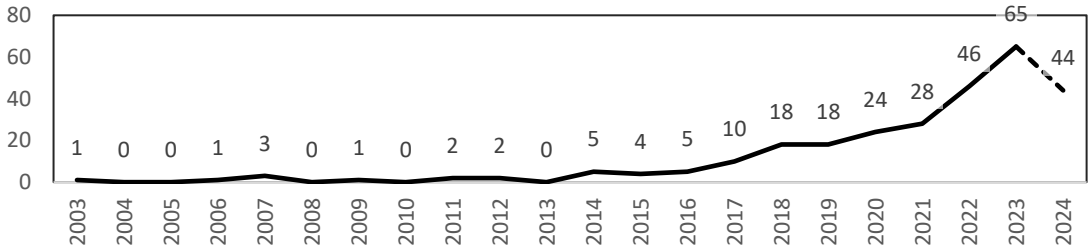
Com esses elementos de comparação, é apresentada a distinção entre as ferramentas selecionadas, destacando as principais características e finalidades de cada uma, conforme apresentado a seguir.

3 RESULTADOS E DISCUSSÃO

3.1 Análise da produção científica sobre ferramentas de IA para curadoria de dados

Inicialmente, para mensurar o grau de interesse da comunidade científica acerca das ferramentas de IA para curadoria automatizada de dados, foi verificada a distribuição dos 277 documentos recuperados ao longo dos anos. Essa etapa, cujo resultado é apresentado na Figura 2, permitiu identificar que o primeiro documento encontrado foi publicado em 2003, um artigo da área de bioinformática que enfatizava a modelagem, aquisição, recuperação e integração de dados sobre biologia molecular. Desde então, o número anual de documentos tem aumentado a uma taxa de 19,75%, atingindo seu auge em 2023, com 65 publicações.

Figura 2 – Número de publicações ao longo dos anos



Fonte: Elaborado pelos autores.

É importante ressaltar que, embora o intervalo temporal encontrado abranja mais de 20 anos, as publicações sobre o tema começaram a aumentar significativamente a partir de 2018, com 87,73% dos documentos recuperados sendo publicados a partir desse ano. Ainda mais expressiva é a proporção de publicações no período de 2021 a 2024, que representa 66,06% do total, mesmo considerando que o ano 2024 ainda não esteja finalizado.

Em relação aos tipos de documentos analisados, 251 são artigos de periódicos (90,61% do total), e o restante se divide em trabalhos de eventos, livros, capítulos de livros, conjuntos de dados (*datapapers*), editoriais, *preprints*, entre outros. Outras informações relevantes sobre o conjunto de dados incluem a disponibilização aberta de 198 documentos (71,48%), a identificação de 160 fontes diferentes (periódicos, eventos etc.), 2.000 autores distintos, 18 documentos com apenas um autor, uma média de 7,57 coautores por documento, uma "idade" média de 3,2 anos por documento, 41,25 citações médias por documento e 13.381 referências citadas.

Dentre a produtividade dos autores, verifica-se o padrão de poucos autores produzirem muito, enquanto muitos autores produzem pouco. Dentre os 2.000 autores identificados, há dois indivíduos, H. Wang e Y. Zhang, que colaboraram em cinco documentos, sendo que são coautores em quatro deles. Na sequência, nove autores possuem três documentos, enquanto 73 autores produziram dois documentos e, por fim, 1.916 autores, equivalente a 95,8% do total, publicaram apenas um documento do *corpus*.

Para analisar o impacto das publicações no campo, foram considerados dois indicadores principais de citação. Primeiro, verificou-se o número de citações dos documentos recuperados de forma global, ou seja, incluindo citações de outros documentos não pertencentes à base de dados. Entre os documentos da base de dados, o mais citado foi publicado em 2018 pelo periódico Nature Review Cancer, intitulado "Artificial intelligence in radiology" (Hosny *et al.*, 2018), que recebeu, segundo informações da base Lens, 2.122 citações, resultando em uma média de 303,14 citações por ano. Em segundo lugar, está um

artigo da Nature Biomedical Engineering (Yu; Beam; Kohane, 2018), também de 2018, com 1.579 citações e uma média de 225,57 citações por ano, seguido por um artigo da International Journal of Precision Engineering and Manufacturing-Green Technology (Kang *et al.*, 2016), publicado em 2016, que possui um total de 981 citações, com uma média de 109,00 citações anuais. Os 10 documentos mais citados da base possuem uma média de 672,10 citações, com desvio padrão de 679,43, demonstrando uma diferença considerável entre o primeiro e o 10º mais citado.

Ainda em relação ao número de citações, foram verificadas as referências mais encontradas e compartilhadas entre os documentos recuperados. Essa análise, denominada cocitação de referências, revelou que o artigo intitulado “Deep learning” (Lecun; Bengio; Hinton, 2015), publicado em 2015 pela revista Nature, foi citado por oito dos 277 documentos analisados, tornando-se a referência mais citada. Em seguida, foram identificados dois artigos com cinco citações cada: o primeiro, publicado em 2016 pelo periódico JAMA (The Journal of the American Medical Association) (Gulshan *et al.*, 2016), e o segundo, publicado em 2022 pelo IEEE Transactions on Network Science and Engineering (Ni *et al.*, 2022). Todas as demais referências foram citadas por quatro ou menos documentos.

A análise de cocitação, aplicada aos periódicos, destacou as principais revistas científicas citadas pelos documentos recuperados. Essa lista é liderada pela revista Nature, com 234 cocitações, seguida pela revista Science, com 175 cocitações, e pela revista Nucleic Acids Research, com 141 cocitações, sendo as únicas identificadas com mais de 100 cocitações. Por outro lado, considerando citações não compartilhadas, a Nature Reviews Cancer é o periódico mais citado, com 2.009 citações, seguida pela Nature Biomedical Engineering, com 1.493 citações, e pelo International Journal of Precision Engineering and Manufacturing-Green Technology, com 904 citações. Ainda em relação aos periódicos, foram analisadas as fontes que mais publicaram documentos entre os 277 analisados. Nesse quesito, os principais destaques foram o Journal of Intelligent & Robotic Systems, com 13 artigos publicados, Scientific Data, com 12 artigos, e o Journal of Grid Computing, com 11 artigos. Nenhum dos demais periódicos identificados apresentou mais de 10 publicações. Esses resultados indicam quais periódicos possuem maior relevância e influência na disseminação de pesquisas sobre curadoria automatizada de dados.

Por fim, os últimos indicadores bibliométricos analisados para este mapeamento do cenário de pesquisa sobre ferramentas automatizadas para curadoria de dados foram a produção e a colaboração entre países. Em primeiro lugar, com participação em 35,50% do total dos documentos, estão os Estados Unidos, que apresentam autores em 35 documentos,

sendo 55 deles produzidos apenas por pesquisadores estadunidenses. Ou seja, da produção desse país, 20 documentos, equivalentes a 26,70%, foram desenvolvidos em colaboração com pesquisadores de outros países. A China aparece em segundo lugar, com um total de 20 documentos, sendo sete produzidos em colaboração internacional (35,0%), seguida pela Alemanha, com 17 documentos, sendo 10 assinados com pesquisadores de outros países (58,8%). O Brasil aparece na nona posição desta lista, com participação em cinco documentos, sendo quatro com colaboração internacional (80,0%). Embora esses dados atestem a predominância dos Estados Unidos na produção científica sobre curadoria automatizada de dados, eles também ressaltam a presença expressiva da colaboração internacional, destacando a importância dessas iniciativas no avanço do conhecimento nessa área.

Assim, com a apresentação desse panorama acerca das pesquisas científicas relacionadas ao tema, a próxima seção é dedicada à exposição das principais características das ferramentas analisadas.

3.2 Características das ferramentas

O Altair é uma biblioteca de visualização de dados em Python que utiliza a API Vega-Lite. É conhecida por sua simplicidade e expressividade, permitindo a criação de gráficos interativos de alta qualidade. Na Curadoria de Dados (CD), o Altair pode contribuir com a visualização interativa, permitindo a criação de gráficos dinâmicos para explorar dados de forma responsiva. Essa funcionalidade pode auxiliar na identificação de padrões, anomalias e tendências nos dados por meio de visualizações; limpeza de dados por meio de análise de gráficos de dispersão para identificar *outliers* e dados inconsistentes; criação de gráficos de barras para analisar a distribuição de categorias em grandes conjuntos de dados; produção de *dashboards* que permitem aos usuários finais acessarem e explorarem os dados sob diferentes perspectivas, por exemplo, em segurança ou saúde pública.

A Amazon SageMaker é a plataforma de *machine learning* (ML) da AWS que oferece ferramentas para construir, treinar e implantar modelos de ML. Como contribuições para a CD, pode atuar no pré-processamento de dados com as ferramentas integradas para limpeza e transformação de dados; treinamento de modelos de ML com grandes volumes de dados; implantação de modelos treinados em produção para fazer previsões em tempo real; análise preditiva para prever tendências de mercado com base em dados históricos, como a previsão de vendas de determinados produtos em diferentes datas comemorativas.

A Databricks é uma plataforma unificada de análise de dados baseada em Apache Spark, otimizada para engenharia de dados e ML. Na CD, pode auxiliar no processamento de grandes volumes de dados de forma distribuída; criar *pipelines* de dados que automatizam a ingestão, limpeza e transformação de dados; auxiliar no desenvolvimento de *pipelines* ETL (*Extract*, *Transform*, *Load*) para coletar dados de múltiplas fontes, limpá-los e prepará-los para análise; utilizando Apache Spark para processar e analisar dados em tempo real, como *logs* de servidores e dados de sensores IoT; e viabilizar os algoritmos de ML para análise preditiva.

O Dataiku é uma plataforma colaborativa de ciência de dados que facilita a construção, implantação e gestão de fluxos de trabalho de ML. Para a CD, pode contribuir com as ferramentas para colaboração entre equipes de ciência de dados, TI e negócios; criação de *workflows* automatizados para coleta, limpeza e análise de dados de múltiplas fontes; integração com bibliotecas de ML para desenvolver e treinar modelos preditivos; permitir que equipes multidisciplinares colaborem na exploração e análise de dados para extrair *insights* acionáveis e implementação de modelos preditivos, como previsão de comportamentos de clientes com base em dados históricos de transações.

O DataRobot oferece uma plataforma de ML automatizada e permite a criação de modelos preditivos. Na CD, pode ser empregado para criar modelos que preveem tendências futuras com base em dados históricos; curadoria de dados temporais, como vendas e dados financeiros, para identificar padrões sazonais e anomalias; agrupamento automático de clientes em segmentos com base em comportamentos. Uma empresa de marketing, por exemplo, pode utilizar o DataRobot para segmentar sua base de clientes e prever comportamentos de compra, otimizando campanhas de marketing direcionadas.

O Google Cloud AutoML permite que desenvolvedores com pouco conhecimento em ML criem modelos de aprendizado de máquina personalizados de alta qualidade. Na CD, pode ser usado para criar modelos que classificam imagens ou textos em categorias específicas; identificação e anotação de objetos em imagens, útil para curadoria de dados visuais; criação de modelos de tradução automática personalizados para diversos idiomas, facilitando a curadoria de dados multilinguísticos. Por exemplo, uma empresa de e-commerce pode utilizar a ferramenta para classificar automaticamente produtos em imagens enviadas por vendedores, melhorando a precisão e a eficiência do processo de catalogação.

O Microsoft Azure Machine Learning é uma plataforma abrangente para construir, treinar e implantar modelos de ML. Na CD, apresenta vantagens como a criação e treinamento de modelos personalizados de ML; integração facilitada com outros serviços da Microsoft, o que viabilizar a CD em ambientes corporativos; e recursos de AutoML para simplificar a criação

de modelos por usuários não técnicos. Por exemplo, uma instituição financeira poderia utilizar o Azure Machine Learning para desenvolver modelos preditivos que analisam tendências de mercado e auxiliam na gestão de riscos.

O OpenAI GPT-4 é uma das ferramentas de processamento de linguagem natural (NLP) mais difundidas e utilizadas atualmente. Desenvolvido pela OpenAI, oferece potencial para gerar textos de alta qualidade, responder a perguntas, resumir documentos, entre outras funções. Na CD, poderia auxiliar em diversas atividades, como a limpeza e normalização de dados, identificando e corrigindo inconsistências em grandes conjuntos de dados textuais; extração de informações relevantes de textos não estruturados, como e-mails, relatórios e artigos científicos; geração de metadados úteis para catalogação e indexação de documentos; e produção de resumos automáticos de textos para facilitar a análise de grandes volumes de dados. Uma empresa de pesquisa científica poderia utilizar o GPT-4, por exemplo, para analisar milhares de artigos de pesquisa, extrair dados relevantes e gerar resumos automáticos para facilitar a revisão de literatura.

O SAS Viya é uma plataforma de análise de dados que fornece capacidades de ML, *deep learning* e NLP, embora seja a menos popular no Brasil quando comparada a outras ferramentas mencionadas. Na CD, pode ser utilizada para análise de dados, pois oferece recursos avançados para análise e curadoria de grandes volumes de dados, modelagem preditiva, uma vez que facilita a criação de modelos preditivos com alta precisão e apresenta facilidades de integração com diferentes fontes de dados, possibilitando a promoção de uma curadoria mais eficiente. Uma empresa de saúde, por exemplo, pode usar SAS Viya para analisar dados de pacientes e prever surtos de doenças, melhorando a resposta e o planejamento estratégico.

O Quadro 1 compara de forma sumarizada os estudos realizados, considerando os critérios explicados na seção de encaminhamentos metodológicos.

Pesquisadores podem usar o Altair gratuitamente para criar gráficos interativos e personalizáveis, explorar dados visualmente, identificar padrões, tendências e anomalias, e comunicar resultados de forma clara e compreensível. A Amazon SageMaker oferece 12 meses de uso gratuito com recursos limitados, acessível por meio de uma conta AWS. O Databricks possui uma versão comunitária com recursos básicos para testes e desenvolvimento.

Quadro 1 - Comparação das ferramentas

Ferramenta	Descrição	Funcionalidades	Usabilidade	Aplicações	Distribuição	Linguagem	Acesso	Educação	Vantagens	Desvantagens
Altair	Ferramenta de visualização e análise estatística; intuitiva e poderosa para análises interativas e exploratórias.	Visualização de dados, criação de gráficos interativos, integração com Python.	Interface amigável e fácil de usar, adequada para análises interativas.	Visualização de dados, análise exploratória de dados, criação de dashboards interativos.	Software desktop, SaaS	Python	https://www.databricks.com/	Gratuito pelo programa acadêmico	Ferramenta intuitiva e poderosa para visualização de dados e análise estatística interativa.	Funções limitadas para análises avançadas, custo elevado para uso corporativo.
Amazon SageMaker	Plataforma completa para construção, treinamento e implantação de modelos de ML; forte integração com AWS.	Treinamento de modelos, implantação de modelos em produção, notebooks integrados, AutoML.	Interface rica e documentação detalhada; integração completa com o ecossistema AWS.	Desenvolvimento de modelos de ML, implantação de modelos, análise preditiva.	SaaS	Python, Jupyter, R	https://aws.amazon.com/pt/sagemaker/	Oferece créditos gratuitos para estudantes e professores através do AWS	Plataforma completa para construção, treinamento e implantação de modelos de ML; forte integração com AWS.	Pode ser complexo para iniciantes; custos podem aumentar com uso intensivo de recursos.
Databricks	Plataforma unificada de análise de dados baseada em Apache Spark, otimizada para ML e engenharia de dados.	Análise de dados em larga escala, ML, integração com Apache Spark, Delta Lake para armazenamento de dados.	Interface baseada em notebooks, documentação abrangente; integração com serviços de nuvem.	Engenharia de dados, análise de dados em tempo real, treinamento de modelos de ML.	SaaS	Python, Scala, SQL	https://www.databricks.com/	Oferece acesso gratuito para estudantes através do Databricks University Alliance	Plataforma unificada de análise de dados baseada em Apache Spark, otimizada para ML e engenharia de dados.	Curva de aprendizado íngreme, custo elevado para uso empresarial.
Dataiku	Plataforma colaborativa para ciência de dados; permite a construção e gestão de fluxos de trabalho de ML.	Preparação de dados, visualização, ML, automação de fluxo de trabalho, colaboração entre equipes.	Interface web intuitiva, para colaboração em equipe, documentação robusta.	Desenvolvimento de modelos de ML, análise preditiva, automação de fluxos de trabalho.	SaaS, software desktop	Python, R, SQL	https://www.dataiku.com/	Acesso gratuito e descontos para estudantes e professores pelo programa Dataiku Academic Program	Plataforma colaborativa para ciência de dados, permite a construção, implantação e gestão de fluxos de trabalho de ML.	Custo elevado, complexidade inicial para novos usuários.
DataRobot	Plataforma de AutoML focada em modelos preditivos para empresas.	Previsão de tendências, análise de séries temporais, segmentação de clientes.	Interface intuitiva, ideal para empresas sem expertise em IA.	Previsão de comportamento de mercado, segmentação de clientes.	SaaS, acesso mediante assinatura	Python, R	https://www.datarobot.com/	Não disponibiliza planos educacionais	Excelente para modelagem preditiva; interface intuitiva para usuários de negócios.	Custo elevado; algumas limitações em personalizações avançadas.
Google Cloud AutoML	Plataforma de AutoML para usuários com pouco conhecimento técnico.	Classificação de imagens e textos, detecção de objetos, tradução automática.	Muito amigável para não-especialistas, boa documentação.	Classificação de produtos, curadoria de conteúdo visual e textual.	SaaS, acesso mediante assinatura	Python, suporte via API REST	https://cloud.google.com/	Sim, Google Cloud for Education	Amigável a usuários sem conhecimento técnico; forte em classificar imagens e textos.	Limitações nas personalizações; custo pode aumentar com uso intensivo.
GPT-4 (OpenAI)	Modelo avançado de NLP para diversas aplicações textuais.	Geração e resumo de texto, extração de informações, limpeza de dados.	Requer algum conhecimento técnico para customização.	Análise de textos, geração de metadados, resumos automáticos.	API em nuvem, acesso mediante assinatura	Python, com wrappers para outras linguagens	https://openai.com/	Não possui planos educacionais, mas alguns créditos diários	Avançada capacidade de processamento de linguagem natural; flexível para diversos usos textuais.	Requer conhecimentos técnicos para customização; custo elevado de assinatura.
Microsoft Azure Machine Learning	Plataforma de ML integrada ao ecossistema Azure.	Criação e treinamento de modelos, AutoML, integração com serviços Azure.	Boa para empresas que já utilizam Azure, documentação extensa.	Desenvolvimento de modelos preditivos, análise de dados em larga escala.	SaaS, acesso mediante assinatura	Python, R, suporte via SDK	https://azure.microsoft.com/en-us/products/machine-learning/	Sim, Azure for Students	Forte integração com Azure; excelente para desenvolvimento de modelos preditivos em larga escala.	Requer alguma familiaridade com Azure; custos podem aumentar com uso intensivo de recursos.
SAS Viya	Plataforma de análise moderna com capacidades avançadas de ML.	Análise avançada de dados, modelagem preditiva, NLP.	Interface robusta, usado por empresas de diversos setores.	Análise de dados de saúde, finanças, marketing.	SaaS, acesso mediante assinatura	SAS Language, Python, R	https://www.sas.com/pt_br/	Sim, SAS OnDemand for Academics	Robusta para análise de dados avançada; grande histórico de uso em diversos setores industriais.	Custo elevado; interface pode ser complexa para novos usuários.

Fonte: Elaborado pelos autores.

Uma plataforma de ciência de dados que se destaca pela facilidade em promover a colaboração é a Dataiku, que possibilita a construção, implantação e gestão de fluxos de trabalho de ML, além de *workflows* automatizados para coleta e análise de dados. A Dataiku oferece uma versão gratuita com funcionalidades limitadas. A DataRobot é uma plataforma de automação de ML que automatiza tarefas de curadoria de dados, seleção de modelos de ML e análises preditivas, e conta com uma versão de avaliação gratuita disponível no site oficial.

Alternativamente, o Google Cloud AutoML permite treinar modelos de ML personalizados para classificação de imagens, processamento de linguagem natural e previsão de séries temporais, oferecendo um nível gratuito com créditos mensais limitados. Em contrapartida, o GPT-4 é uma ferramenta de processamento de linguagem natural voltada à geração de texto, tradução automática, análise de sentimentos e sumarização de textos, oferecendo créditos gratuitos diários limitados.

O Microsoft Azure Machine Learning é uma plataforma abrangente para construir, treinar e implantar modelos de ML em um ambiente colaborativo e oferece um nível gratuito com recursos básicos. Por fim, também reconhecida como uma plataforma flexível, o SAS Viya é uma plataforma de análise de dados e ML que suporta toda a jornada de análise, desde a preparação até a modelagem e implementação, oferecendo versões de avaliação acessíveis mediante solicitação no site oficial da ferramenta.

3.3 Discussão dos resultados

A crescente necessidade de pesquisas voltadas à automação de tarefas relacionadas à curadoria de dados (Ehrlinger; Wöss, 2022) é refletida no aumento encontrado da produção científica nessa área. Apesar da importância crítica da curadoria no ciclo de vida dos dados, a falta de automação faz com que muitas atividades dependam de um esforço humano intensivo, criando um gargalo para analistas, que acabam dedicando muito mais tempo à preparação dos dados do que à análise propriamente dita (Talburt; Ehrlinger; Magruder, 2023). Sem a implementação de soluções automatizadas, essa situação tende a se agravar com o aumento de dados gerados automaticamente, como aqueles provenientes de sensores, por exemplo.

Neste contexto, diante da combinação de diferentes fontes de dados, modelos, códigos e infraestrutura, a demanda por plataformas de Ciência de Dados e Machine Learning

(CDML) como um ativo estratégico empresarial nunca foi tão elevada, especialmente por soluções de IA, incluindo aquelas do tipo generativo (Jaffri *et al.*, 2024). Esse cenário também se reflete no aumento da oferta e na diversificação de soluções CDML, que apresentam diferentes funcionalidades, recursos e finalidades, conforme discutido neste trabalho.

Além do crescimento expressivo dos últimos anos, as publicações científicas e as plataformas CDML têm em comum o fato de estarem majoritariamente concentradas nos Estados Unidos. Aproximadamente um terço das publicações analisadas conta com a participação de pesquisadores estadunidenses; dos 277 documentos encontrados, 97 incluem ao menos um autor nascido nesse país. Da mesma forma, das nove ferramentas analisadas, todas possuem sede nos Estados Unidos, sendo que oito delas se identificam como companhias americanas. Além disso, outro ponto de convergência entre as plataformas analisadas e o meio acadêmico é a ampla disponibilidade de versões destinadas a fins educacionais e de pesquisa, evidenciando o interesse das empresas em colaborar com pesquisadores e estudantes.

4 CONSIDERAÇÕES FINAIS

A exploração metodológica e comparativa das ferramentas de IA para curadoria automatizada de dados (GPT-4, IBM Watson, Google Cloud AutoML, DataRobot, Alteryx, Microsoft Azure Machine Learning e SAS Viya) permitiu traçar um panorama abrangente das soluções disponíveis para diversos perfis interessados no tema. As ferramentas avaliadas, reconhecidas como líderes no Quadrante Mágico do Gartner, demonstraram um grande potencial para enfrentar os desafios atuais na conversão de grandes volumes de dados brutos em informações úteis.

A breve análise bibliométrica empregada permitiu constatar um aumento no número de publicações relacionadas ao tema, especialmente a partir de 2021. Considerando os documentos do *corpus*, predominantemente composto por artigos publicados em periódicos, foi possível verificar a importância para a área da revista Nature, que apresentou 234 cocitações, e da Nature Reviews Cancer, com 2.009 citações. Além disso, os 13 artigos publicados pelo periódico Journal of Intelligent & Robotic Systems posicionaram essa revista como a principal fonte entre os documentos analisados. Por fim, a análise da produção nos países destacou a colaboração internacional, mesmo diante do evidente predomínio das pesquisas realizadas nos Estados Unidos.

Dentre as limitações desta pesquisa, destaca-se o escopo da análise bibliométrica, que não examinou o conteúdo dos documentos, mas buscou, de forma objetiva, fornecer um panorama atual das publicações na área. Assim, para pesquisas futuras, recomenda-se a investigação das leis bibliométricas, a construção de redes bibliométricas e o uso de outros indicadores que não foram abordados neste estudo. Além disso, o método de seleção das ferramentas analisadas também pode ser considerado metodologicamente restritivo. Um próximo passo poderia ser a identificação de ferramentas mencionadas nas publicações para a realização de uma nova comparação. Ademais, sugere-se para futuros estudos a comparação entre as ferramentas por meio de testes de *benchmark* para problemas de dados similares, assim como avaliações de usabilidade com a adoção de critérios bem definidos. Outros aspectos de comparação também poderiam ser explorados, ampliando a compreensão sobre a aplicabilidade das soluções.

REFERÊNCIAS

ARIA, M.; CUCCURULLO, C. bibliometrix : An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959–975, nov. 2017. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1751157717300500.

ARIA, M.; CUCCURULLO, C. **Biblioshiny**: bibliometrix for no coder. Disponível em: https:// bibliometrix.org/biblioshiny/biblioshiny1.html. Acesso em: 8 dez. 2022.

ARIA, M.; LE, T.; CUCCURULLO, C.; BELFIORE, A.; CHOE, J. openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex. **The R Journal**, v. 15, n. 4, p. 167–180, 11 abr. 2024. Disponível em: https://journal.r-project.org/articles/RJ-2023-089.

BEHESHTI, A.; BENATALLAH, B.; TABEBORDBAR, A.; MOTAHARI-NEZHAD, H. R.; BARUKH, M. C.; NOURI, R. DataSynapse: A Social Data Curation Foundry. **Distributed and Parallel Databases**, v. 37, n. 3, p. 351–384, 2018. Disponível em: https://doi.org/10.1007/s10619-018-7245-1.

BEHESHTI, S. M. R.; TABEBORDBAR, A.; BENATALLAH, B.; NOURI, R. On automating basic data curation tasks. In: 26th International World Wide Web Conference 2017, WWW 2017 Companion, 2017, [...]. 2017. p. 165–169.

DIODATO, V. P. **Dictionary of bibliometrics**. New York: Routledge, 2013.

EHRLINGER, L.; WÖSS, W. A Survey of Data Quality Measurement and Monitoring Tools. **Frontiers in Big Data**, v. 5, n. March, 2022.

FREITAS, A.; CURRY, E. Big Data Curation. *In*: CAVANILLAS, J.; CURRY, E.; WAHLSTER, W. (ed.) **New horizons for a Data-Driven Economy**: a roadmap for usage and exploitation of Big Data in Europe. Madrid, Spain: Springer Open, 2016. p. 87–118.

GULSHAN, V.; PENG, L.; CORAM, M.; STUMPE, M. C.; WU, D.; NARAYANASWAMY, A.; VENUGOPALAN, S.; WIDNER, K.; MADAMS, T.; CUADROS, J.; KIM, R.; RAMAN, R.; NELSON, P. C.; MEGA, J. L.; WEBSTER, D. R. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. **JAMA**, v. 316, n. 22, p. 2402, 13 dez. 2016. Disponível em: http://jama.jamanetwork.com/article.aspx?doi= 10.1001/jama.2016.17216.

HOSNY, A.; PARMAR, C.; QUACKENBUSH, J.; SCHWARTZ, L. H.; AERTS, H. J. W. L. Artificial intelligence in radiology. **Nature Reviews Cancer**, v. 18, n. 8, p. 500–510, 17 ago. 2018. Disponível em: https://www.nature.com/articles/s41568-018-0016-5.

JAFFRI, A.; POPA, A.; KRENSKY, P.; HARE, J.; BHATI, R.; HASSANLOU, M.; ZHANG, T. **Magic Quadrant for Data Science and Machine Learning Platforms**. Disponível em: https:// www.gartner.com/doc/reprints?id=1-2HNJ79HD&ct=240523. Acesso em: 26 jun. 2024.

KANG, H. S.; LEE, J. Y.; CHOI, S.; KIM, H.; PARK, J. H.; SON, J. Y.; KIM, B. H.; NOH, S. Do. Smart manufacturing: Past research, present findings, and future directions. **International Journal of Precision Engineering and Manufacturing-Green Technology**, v. 3, n. 1, p. 111–128, 23 jan. 2016. Disponível em: http://link.springer.com/10.1007/s40684-016-0015-5.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LENS.ORG. **About The Lens**. Disponível em: https://about.lens.org/. Acesso em: 26 jun. 2024.

MOREIRA, P. S. da C.; GUIMARÃES, A. J. R.; TSUNODA, D. F. QUAL FERRAMENTA BIBLIOMÉTRICA ESCOLHER? um estudo comparativo entre softwares. **P2P E INOVAÇÃO**, v. 6, p. 140–158, 31 mar. 2020. Disponível em: http://revista.ibict.br/p2p/article/view/5098.

NI, Q.; GUO, J.; WU, W.; WANG, H.; WU, J. Continuous Influence-Based Community Partition for Social Networks. **IEEE Transactions on Network Science and Engineering**, v. 9, n. 3, p. 1187–1197, 1 maio 2022. Disponível em: https://ieeexplore.ieee.org/document/9658135/.

PENFOLD, R. Using the Lens database for staff publications. **Journal of the Medical Library Association**, v. 108, n. 2, p. 341–344, 2020.

PRIEM, J.; PIWOWAR, H.; ORR, R. **OpenAlex**: A fully-open index of scholarly works, authors, venues, institutions, and conceptsarXiv, 2022. Disponível em: https://arxiv.org/abs/2205. 01833.

TALBURT, J. R.; EHRLINGER, L.; MAGRUDER, J. Editorial: Automated data curation and data governance automation. **Frontiers in Big Data**, v. 6, 2023.

TENOPIR, C.; BIRCH, B.; ALLARD, S. **Academic Libraries and Research Data Services: Current Practices and Plans for the Future**. [s.l: s.n.]. Disponível em: https://alair.ala.org/items/ 60ed0e28-bdda-4616-a66d-de4b343af101.

YU, K.-H.; BEAM, A. L.; KOHANE, I. S. Artificial intelligence in healthcare. **Nature Biomedical Engineering**, v. 2, n. 10, p. 719–731, 10 out. 2018. Disponível em: https://www.nature.com/ articles/s41551-018-0305-z.