DOI: 10.5433/1981-8920.2022v27n3p751

APRENDIZAGEM DE MÁQUINA E CIÊNCIA DA INFORMAÇÃO: CONTRIBUIÇÕES PARA UMA AGENDA DE PESQUISA E ENSINO NA CI BRASILEIRA

MACHINE LEARNING AND INFORMATION SCIENCE: CONTRIBUTIONS TO A RESEARCH AND TEACHING AGENDA IN THE BRAZILIAN INFORMATION SCIENCE

Dalton Lopes Martins[a]
Rafaella Carine Monterei[b]

RESUMO
Introdução: O artigo discute e apresenta os conceitos de inteligência artificial, aprendizagem de máquina e discute potenciais aplicações e estudos técnica de referência na área da Ciência da Informação com maior ênfase no campo das bibliotecas e biblioteconomia. **Objetivo:** tem por intenção apoiar na construção de uma agenda de pesquisa e ensino do tema da aprendizagem de máquina por meio do conhecimento de suas principais aplicações e algoritmos mais usados. **Metodologia:** artigo estuda as referências de algoritmos de aprendizagem de máquina em documentos indexados na área da Ciência da Informação na base *Web of Science*. **Resultados e Conclusões:** avalia 3111 documentos identificados e conclui que a abordagem supervisionada por meio de técnicas de classificação é a mais utilizada no campo, com evidência para os algoritmos de Support Vector Machine, Decision Tree, Random Forest.

Descritores: Aprendizagem de Máquina. Ciência da Informação. Algoritmos. Web of Science. Ensino. Pesquisa.

1 INTRODUÇÃO

O presente artigo é parte de um esforço maior de pesquisa envolvendo vários estudos em andamento visando compreender melhor como a área da

[a] Doutor em Ciências da Informação pela Universidade de São Paulo (USP). Docente do Programa de Pós-graduação em Ciência da Informação da Faculdade de Ciência da Informação na Universidade de Brasília (UnB), Brasília, Brasil. E-mail: daltonmartins@unb.br

[b] Mestranda em Ciência da Informação pela Universidade de Brasília (UnB). Analista Judiciário apoio especializado em Biblioteconomia do Superior Tribunal de Justiça. Brasília, Brasil. E-mail: rafaellamonterei@gmail.com

Ciência da Informação (CI) tem aplicado as técnicas de Aprendizado de Máquina (AM) em seus problemas característicos, que questões têm sido priorizadas, que resultados têm sido obtidos, os cuidados a serem tomadas, as ferramentas e técnicas utilizadas. O objetivo de tal esforço de pesquisa é tanto compreender como essas questões impactam os procedimentos de ensino e pesquisa na área da CI visando futuras alterações em currículos de disciplinas de graduação e pós-graduação, bem como de identificar oportunidades de sistematização e construção de procedimentos metodológicos que permitam avançar nas contribuições que a área da CI pode aportar ao contexto contemporâneo da ciência e tecnologia.

No presente trabalho, o objetivo da pesquisa consiste em analisar quais são os algoritmos de AM mais mencionados nos artigos científicos em nível internacional da área da CI. Entende-se que conhecer os algoritmos mais mencionados, como estão sendo usados e suas aplicações é uma etapa importante para compreender o estado da arte do tema na CI.

De forma a posicionar conceitualmente a questão e contextualizar o debate, na seção 2 são apresentadas definições conceituais e históricas a respeito da inteligência artificial, considerando um campo mais amplo que contém a abordagem da aprendizagem de máquina em seu contexto. Na seção 3, a aprendizagem de máquina é tratada conceitualmente e são discutidas suas diferentes abordagens e desdobramentos metodológicos. Na seção 4, discute-se a luz de pesquisas e estudos técnicos recentes como a aprendizagem de máquina e diferentes visões da inteligência artificial têm sido aplicadas em problemas e instituições características da Ciência da Informação, com maior ênfase no campo das bibliotecas. Na seção 5, são apresentados os procedimentos metodológicos da pesquisa, sendo os resultados apresentados na seção 6 e a conclusão na sessão 7.

2 LINHAS CONCEITUAIS E HISTÓRICAS ACERCA DA INTELIGÊNCIA ARTIFICIAL

A busca por soluções capazes de simular o raciocínio humano não é algo recente e remontam aos estudos filosóficos que buscavam compreender as leis

que governam a parte racional da mente. No entanto, elas começaram a se tornar realidade (com algumas restrições) a partir de 1950. Desta forma, o desenvolvimento do campo de estudos da Inteligência Artificial (IA) em paralelo com o avanço da infraestrutura tecnológica, converteram-se em elementos com uma capacidade transformadora para o desenvolvimento da sociedade contemporânea sob a égide de diversos fatores, sejam eles sociais, culturais, econômicos, políticos e científicos.

Um importante elemento conceitual do campo foi proposto por John Searle (1980) por meio de uma dicotomia composta pelas expressões “IA forte” e “IA fraca”. A IA forte simula o comportamento inteligente humano, o qual as máquinas se tornam autoconscientes. Este tipo de inteligência é considerada por muitos especialistas uma realidade um pouco distante, no entanto, de acordo com Taulli (2020, p. 19), existem algumas empresas que já se concentram nesta categoria, como é o caso do Google por meio da *DeepMind*. Já em relação à IA fraca é aquela designada para realizar a correspondência entre padrões por meio de tarefas especificas, porém sem grande autonomia. Um exemplo de aplicações nesta categoria são os *chatbots*, que dependem de insumos (dados) fornecidos pelo ser humano.

No que se refere à sua origem, houve vários trabalhos publicados sobre Inteligência Artificial antes de 1950, durante o período da Segunda Guerra Mundial, que poderiam ser classificados sob este domínio, todavia, a pesquisa publicada por Alan Turing foi considerada por muitos teóricos a mais influente e, portanto o marco zero para o desenvolvimento deste campo. Em 1950, Alan Turing publicou um artigo denominado *Computing machinery and intelligence*, o qual propôs um teste com o objetivo de verificar se a máquina conseguia representar o papel de um humano no jogo da imitação (conhecido também pela expressão ‘teste de Turing’), “enganando” o interrogador, de maneira que ele não consiga fazer distinção entre o humano e a máquina.

Seis anos mais tarde, em 1956, John McCarthy organizou um evento de dois meses em Dartmouth (RUSSEL; NORVIG, 2021), no estado de New Hamshire, nos Estados Unidos. Neste evento havia 10 participantes, dentre os quais se destacavam os pesquisadores: Claude Shannon, Nathaniel Rochester,

Marvin Minsky, Allen Newell, O. G. Selfridge, Raymond Solomonoff e Arthur Samuel. A proposta de estudo deste evento, de acordo com McCarthy *et al.* (1955, tradução nossa, p. 1) era:

> [...] prosseguir com base na conjectura de que cada aspecto do aprendizado ou qualquer outra característica da inteligência pudesse, em princípio, ser descrita tão precisamente a ponto de ser construída uma máquina para simulá-la.

Logo, McCarthy denominou este estudo como "[...] um estudo da inteligência artificial", e, desta forma, foi a primeira vez que a expressão Inteligência Artificial foi utilizada (TAULLI, 2020, p. 22). Neste contexto, McCarthy (2007, p. 2) definiu a IA como "[...] a ciência e engenharia de máquinas inteligentes, especialmente programas inteligentes de computador". Em suma, ela se baseou na capacidade das máquinas simularem ações de maneira análoga ao raciocínio humano, cujo objetivo é resolver problemas, simular situações ou tomar decisões de maneira inteligente.

Entre os anos de 1956 e 1970, que sucederam ao evento em Dartmouth, foram denominados por muitos pesquisadores como a Era de Ouro da IA. Durante este período, houve um substancial desenvolvimento tecnológico devido à produção dos primeiros computadores e ferramentas de programação. Além disso, dado ao contexto da Guerra Fria, o governo dos Estados Unidos, por meio da *Advanced Research Projects Agency* (ARPA) – a mesma instituição que desenvolveu a internet – foi a principal fonte de financiamento da IA (TAULLI, 2020, p. 24). Já em relação ao financiamento por parte do setor privado, com exceção da IBM, houve pouco envolvimento (TAULLI, 2020, p. 22). No entanto, grande parte da inovação em IA aconteceu no âmbito acadêmico a partir da utilização de sistema informáticos ainda primitivos, por meio da publicação de estudos que incluíam tópicos acerca dos métodos bayesianos, *machine learning* e redes neurais. Um importante destaque deste período foi a criação do primeiro *chatbot* da história denominado Elisa, que de acordo com Barbosa e Bezerra (2020, p. 6), foi um sistema baseado em palavras-chaves e estrutura sintática que conversava de forma automática imitando o comportamento de uma psicanalista.

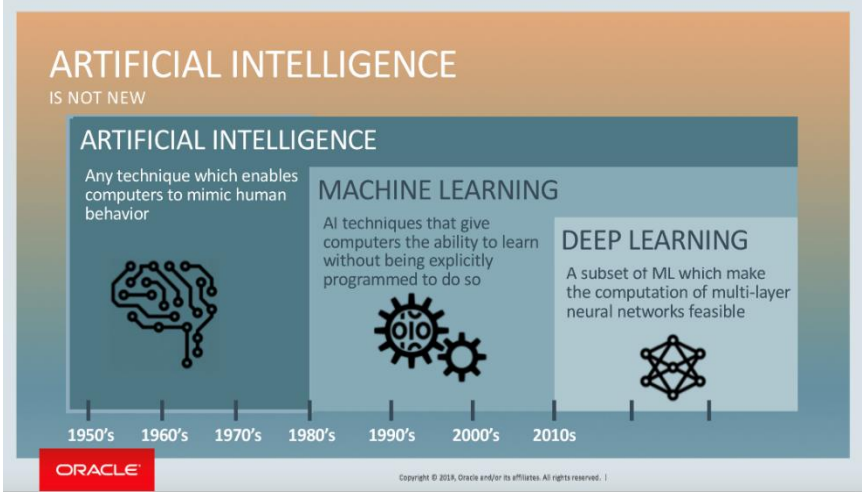
Após os anos de 1970 até 1980 perdurou um período batizado de "Inverno da IA", o qual foi caracterizado por uma época de agravamento da crise

econômica. O marco histórico do campo da IA ao longo deste período se deve a publicação do Relatório de Lighthill em 1973 pelo professor britânico Sir James Lighthill. Neste relatório financiado pelo parlamento do Reino Unido, foi divulgado uma nota de repúdio total aos “objetivos grandiosos” da IA forte (TAULLI, 2020, p. 29). Lighthill (1972) concluiu que em nenhuma área do campo, as descobertas feitas até o momento produziram o grande impacto então prometido.

No entanto, a partir de 1980 a IA recomeçou a florescer de maneira acentuada e, desta forma, houve o surgimento dos primeiros sistemas especialistas, a criação de redes neurais, todos eles apoiados no rápido desenvolvimento da indústria da tecnologia. Em 1996 foi criado pela IBM o Deep Blue, este supercomputador e software conquistou as manchetes dos principais jornais em 1997, devido a disputa de xadrez com o campeão Garry Kasparov, o qual a máquina saiu vitoriosa. Outra iniciativa de destaque no contexto dos jogos de tabuleiro, foi o AlphaGo, que em 2016 derrotou o então campeão de Go Lee Sedol, por meio de conhecimentos adquiridos para análise de lances através do aprendizado de máquina e das redes neurais. No entanto, apesar destas aplicações de destaque, a IA não se restringe apenas ao âmbito dos jogos de tabuleiro, tornando-se um campo de estudos aplicável a uma variedade de setores, de modo a ajudar a identificar padrões, prever problemas, corrigir erros e tomar decisões.

Neste cenário efêmero de grandes transformações ao longo de um curto intervalo de tempo, surgiram diversas aplicações, dentre as quais se destacam o *Machine Learning* e o *Deep Learning*, conforme a representação da linha do tempo da IA produzida pela Oracle (2018).

Figura 1 - Linha do tempo da Inteligência Artificial



Fonte: Oracle (2018)

Ao explorar a linha do tempo da Oracle (2018), a Inteligência Artificial apresenta-se como um conjunto de técnicas, que no decorrer de 60 anos, ou seja, menos de um século desenvolvimento e com o apoio de uma infraestrutura em expansão, permite que os computadores simulem o comportamento humano. A IA engloba tanto técnicas de *Machine Learning* (serão apresentadas em seção específica neste trabalho) quanto de *Deep Learning*, que confere aos computadores a capacidade de aprender.

3 APRENDIZAGEM DE MÁQUINA: UMA ABORDAGEM CONCEITUAL

A Inteligência Artificial, como já analisado em seção anterior, produziu um vasto conjunto de técnicas, dentre as quais se destaca o Aprendizado de Máquina (AM), conhecida também por sua expressão em inglês *Machine Learning* (ML). No que diz respeito a origem desta expressão, o seu termo foi cunhado pela primeira vez pelo cientista americano Arthur Samuel em 1959. Samuel (1959) definiu o AM como o campo de estudo que permite que computadores aprendam sem que sejam programados explicitamente.

Samuel (1959) descreveu ainda, um programa denominado *Game of Checkers,* que simulava um jogo de damas entre um computador e um ser humano. Este trabalho revolucionário para a época, demostrou que um computador poderia aprender por meio do processamento de dados sem ter sido explicitamente programado para realizar tal tarefa (TAULLI, 2020, p. 64).

Já em 1997, o cientista e professor americano Tom M. Mitchell apresentou uma definição da área que se tornou mais popular, o qual define AM como “a área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio da experiência” (SILVA; FERRARI, 2016, p. 14). Esta área de pesquisa proveniente da Ciência da Computação, também recebe contribuições da Estatística, da Neurociência, da Teoria da Informação, das Ciências Cognitivas dentre outros campos científicos.

Em essência, o objetivo do Aprendizado de máquina segundo Faceli *et al.* (2019, p. 341) é a construção de modelos computacionais que descrevem sistemas complexos a partir da observação do comportamento do sistema. Um exemplo de sua aplicação é a utilização de algoritmos em plataformas de *streaming* como a Netflix e o Prime Vídeo da Amazon para realizar sugestões de mídias baseadas no comportamento do usuário. Desta forma, os algoritmos programados com AM aprendem com base na experiência passada, por meio de um princípio de inferência denominado indução, no qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos (FACELI *et al.*, 2019, p. 3).

No que se refere aos algoritmos, o seu desenvolvimento tem favorecido cada vez mais a expansão do AM, aumentando a sua produtividade e eficiência, e, desta forma apresenta-se como um dos elementos-chave do campo. Os algoritmos aprendem por meio de um conjunto de dados de treinamento, cujo objetivo de acordo com Faceli *et al.* (2019, p. 5) é procurar uma hipótese capaz de descrever as relações entre os objetos e que melhor se ajuste aos dados de treinamento.

Tradicionalmente o *Machine Learning* é subdivido em duas grandes categorias, a saber: aprendizagem supervisionada e aprendizagem não supervisionada, conforme a figura abaixo.

Figura 3 - Hierarquia das categorias do aprendizado de máquina.

Fonte: Faceli *et al.* (2019, p. 6)

No que se refere ao aprendizado supervisionado ele é baseado, segundo Silva e Ferrari (2016, p. 16), em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado pelo sistema. Além disso, inclui a figura do supervisor que auxilia na classificação dos dados a *priori*, por meio de algoritmos de classificação ou de regressão. Um exemplo de aplicação, é a utilização de algoritmos de aprendizado para a filtragem de *spams* em *e-mails*. Os sistemas de comunicação aprendem através diversos dados fornecidos pelos próprios e-mails tais como os metadados provenientes do endereço utilizado pelo remetente, as palavras utilizadas no corpo da mensagem e no assunto do e-mail, elementos que podem indicar ataque virtuais se antecipando as investidas de golpistas.

Já o aprendizado não supervisionado é baseado, segundo Silva e Ferrari (2016, p. 16), apenas nos objetos da base, cujos rótulos são desconhecidos. Basicamente, o algoritmo deve aprender a “categorizar” ou rotular os dados brutos, sem dispor da figura do supervisor. O aprendizado não supervisionado é empregado de maneira a encontrar padrões em conjunto de dados que normalmente se encontram desorganizados, e a abordagem mais comum, segundo Taulli (2020, p. 77), é agrupamento (*clustering*), que manipula dados não rotulados e usa algoritmos para colar itens semelhantes em grupos.

Neste cenário, aprendizado não supervisionado pode ser aplicado na mineração de dados de mídia social, no registro de empréstimos de livros ao traçar o perfil dos usuários de uma biblioteca e na verificação de transações bancárias. Por fim, é importante mencionar um fato curioso, de acordo com Taulli

(2020, p. 78), as aprendizagens humana e animal são, em grande parte, não supervisionadas, pois o mundo, sob o ponto de vista destes dois sujeitos, é descoberto por meio de observações.

Além das abordagens citadas, existem algumas tarefas de aprendizado que não se enquadram nas suas subdivisões anteriormente abordadas, que segundo Faceli *et al.* (2019, p. 7) são: o aprendizado semissupervisionado, o aprendizado ativo e o aprendizado por reforço.

O aprendizado semissupervisionado de acordo com Taulli (2020, p. 79) é uma mistura de aprendizado supervisionada e não supervisionada que surge quando se tem uma pequena quantidade de dados não rotulados. Este tipo de aprendizado utiliza um conjunto de dados de treinamento rotulados e não rotulados com o objetivo de induzir um modelo preditivo.

Já o aprendizado ativo, segundo Faceli *et al.* (2019, p. 7) apresenta a estratégia de selecionar interativamente os exemplos a serem rotulados e os rótulos a ser atribuído a cada um deles. Esta abordagem é adequada quando não há muitos dados disponíveis ou os dados são muito caros para serem adquiridos.

Por fim, o aprendizado por reforço que de acordo com Taulli (2020, p. 79) refere-se ao processo de tentativa e erro, em que o aprendizado será aperfeiçoado por meio de reforços positivos e negativos em um ambiente de *feedback* entre o sistema de aprendizado e as suas experiências. Este tipo de abordagem é muito utilizado em jogos e na robótica.

Neste cenário, o aprendizado de máquina dispõe de um variado conjunto de aplicações bem-sucedidas baseados em problemas reais, o que torna possível a sua aplicação nos mais diversos cenários. Desta maneira, o próximo tópico apresentará as possíveis relações entre a Inteligência Artificial e a Ciência da Informação com maior ênfase no campo das bibliotecas e da biblioteconomia.

4 A INTELIGÊNCIA ARTIFICIAL APLICADA AO CAMPO DE ÁREAS AFINS À CIÊNCIA DA INFORMAÇÃO

O cenário informacional vem sofrendo profundas transformações ao longo das últimas décadas. E isso se deve, em grande medida, à criação e ao

desenvolvimento de novas tecnologias da informação e da comunicação (TICs), o qual resultou em um ambiente propicio para o desenvolvimento exponencial da informação e, provocou uma mudança na maneira em como acessamos e consumimos a informação.

À vista disso, este novo panorama informacional se apresentou de maneira desafiadora, em especial, para as bibliotecas, ao transformar as suas tradicionais funções e, desta forma, demandar novas habilidades e competências de seus profissionais frente a este novo cenário.

Diante disso, acrescenta-se ainda a este cenário *sui generis*, a Inteligência Artificial (IA), um elemento já presente em nossas vidas, e que pode desempenhar, um papel fundamental em diferentes setores da sociedade, dentre eles nas bibliotecas.

Assim, a IA/ML pode auxiliar na transformação da natureza das bibliotecas, de instituições relegadas exclusivamente ao depósito de livros para entidades disseminadoras do conhecimento, assumindo assim um papel mais crítico na sociedade. Ademais, ao ressignificar o seu papel, as bibliotecas podem ofertar produtos e serviços personalizados, de modo a proporcionar experiências interativas e intuitivas, para melhor atender uma geração de usuários cada vez mais autossuficiente e, onde os recursos informacionais encontram-se cada vez mais dispersos.

Cordell (2020, p. 6) afirma que a literatura sobre aprendizado de máquina em bibliotecas é mais longa do que podemos imaginar devido à intensa atenção que o campo tem recebido nos últimos anos. À vista disso, Smith em 1976 escreveu acerca da transição entre sistemas de recuperação baseados em fita adesiva para sistemas de recuperação *online*. Cordell (2020, p. 6-7) destaca que Smith (1976) descreveu uma série de intervenções do ML e da IA em processos de recuperação da informação. E ainda afirma (2020, p. 6-7) que essas tecnologias ajudariam os pesquisadores por meio de processos de reconhecimento de padrões, classificação, recuperação, representação de informações, resolução de problemas e mais centralmente na descoberta de conhecimento.

Já no século XXI, e em convergência os prognósticos de Smith (1976), Bohyun Kim (2020) afirma que aplicar a IA em bibliotecas, pode melhorar a

descoberta e recuperação de informações e extrair informações a partir de um grande número de documentos. Além disso, a AI tem um enorme potencial para automatizar os processos de representação descritiva e temática de documentos (catalogação, classificação, indexação etc.), tendo em vista que esses processos consomem muito tempo e esforços (KIM, 2020).

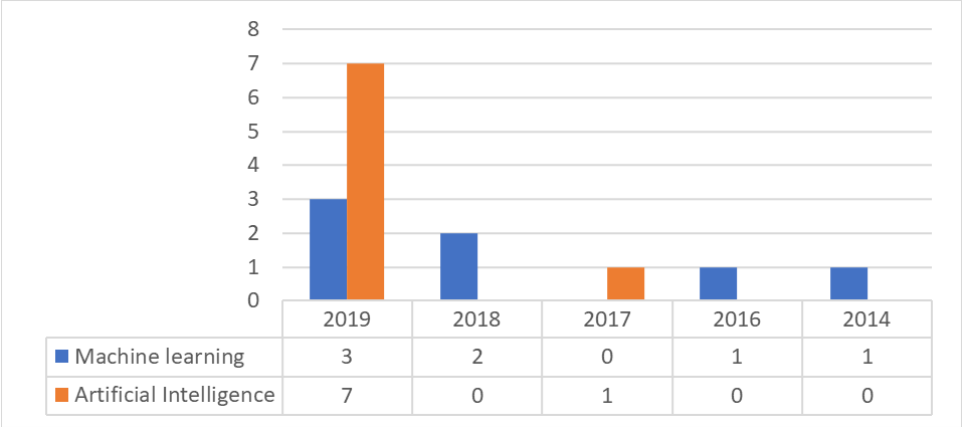
Já Vijayakumar e Sheshadri (2019, p. 136) afirmam que o uso de sistemas especialistas, redes neurais, processamento de imagem, processamento de linguagem natural, reconhecimento de voz, robótica, aprendizado de máquina (AM) etc., podem enriquecer os serviços de uma biblioteca.

Outras aplicações da IA já descritas na literatura são o desenvolvimento de *chatbots* ou assistentes virtuais para apoiar o serviço de referência, o uso de robôs na gestão de acervos, a análise de grandes coleções textuais e imagéticas para a atribuição de metadados, a criação de laboratórios de gamificação, a coleta automática de metadados para coleções digitais, o reconhecimento óptico de caracteres (OCR) de textos em documentos digitalizados dentre outros.

Em pesquisa realizada pela Europeana (2021, p. 5) com 56 representantes de instituições culturais, dentre as quais galerias, bibliotecas, arquivos, museus (GLAMs), instituições de pesquisa e do setor cultural, provenientes de 20 países dentre os quais o Brasil. Realizou uma pesquisa acerca dos impactos da IA e do ML no patrimônio cultural e, revelou que 91,8% os entrevistados estão interessados em pelo menos um tópico de AI e, 54% das instituições, ou seja, mais da metade, desenvolvem projetos em AI. E, conclui que o uso da IA desempenhará um papel cada vez mais importante, especialmente no que diz respeito ao fornecimento de acesso, metadados, extração e enriquecimento de dados (EUROPEANA, 2021, p. 22).

Nesta seara ao realizar uma pesquisa com as expressões “*artificial intelligence*” e “*machine learning*”, entre os anos de 2013 a 2019, em trabalhos publicados em congressos realizados pela Federação Internacional de Associações de Bibliotecas e Instituições (IFLA) e depositados no repositório institucional da IFLA, foram recuperados 8 registros com a expressão “*Artificial intelligence*” e 7 registros com a expressão “*machine learning*”, conforme apresentado no gráfico abaixo.

Gráfico 1 - Documentos recuperados no repositório da IFLA



	2019	2018	2017	2016	2014
Machine learning	3	2	0	1	1
Artificial Intelligence	7	0	1	0	0

Fonte: Adaptado de IFLA (2021)

A representação gráfica dos resultados recuperados no repositório, demonstram uma tendência de crescimento de discussão dos temas nos últimos anos, com destaque para a grande concentração de trabalhos sobre aplicações de IA e AM publicados no ano de 2019.

Tendo em vista os resultados das pesquisas da EUROPEANA (2021) e da IFLA (2021), infere-se, portanto, que as técnicas de IA e ML já são uma realidade em algumas instituições, deste modo é possível apresentar algumas iniciativas de destaque desenvolvida instituições de renome internacional.

A *Library of Congress* (LC) em parceria com a Universidade de Nebraska-Lincoln, desenvolveram o projeto *Digital libraries, intelligence data analytics, and augmented description* (AIDA). O projeto visa a extração e destaque de conteúdo visual do *Chronicling America*[7], para isso faz uso de uma série de métodos de processamento de imagem e AM em coleções de manuscritos minimamente processados apresentados no projeto *By the People* (LORANG *et al.*, 2020, p. 2). Ademais, o projeto da LC e da Universidade de Nebraska explora os desafios sociais e técnicos em relação ao desenvolvimento do AM no setor de patrimônio cultural.

Outra iniciativa em IA foi desenvolvida pela Biblioteca Nacional da Noruega, que realizou um experimento para uso da AM em uma coleção de artigos, cujo objetivo é aplicar a classificação automática em documentos com base nas categorias propostas por Dewey em sua Classificação Decimal (CDD) (BRYGFJELD; WETJEN; WALSØE, 2018).

Já a Biblioteca Central Oodi de Helsinque na Finlândia, segundo Hammais, Ketamo e Koivisto (2019), elaborou um aplicativo de celular com tecnologia IA, projetado para usuários da biblioteca, com uma interface semelhante a um bate-papo, cuja finalidade é realizar sugestões personalizadas de leitura.

No âmbito do Reino Unido, a British Library é uma das principais colaboradas do projeto "*Living with machine*", que investiga os impactos da tecnologia na vida de pessoas durante a Revolução Industrial. Desta forma, o projeto visa analisar documentos históricos do século XIX, por meio da aplicação de ferramentas de ML e Ciência de dados (EUROPEANA, 2021, p. 13).

Já na França, a Biblioteca Nacional da França (BnF) produziu um plano de ação com o roteiro de aplicações da IA para o período 2021-2026. Ademais, os projetos de IA estão concentrados na instituição em cinco áreas principais: apoio as atividades de catalogação; gerenciamento de coleções; pesquisa, análise e acesso à informação; engajamento do usuário; e, por fim, tomada de decisão e governança (BIBLIOTECA NACIONAL (França), 2021). Além disso, a BnF por meio do projeto Gallicapix aplicou o *TensorFlow* no desenvolvimento de seus próprios modelos para treinamento de conjunto de dados imagéticos, cuja finalidade é classificar imagens e detectar objetos.

No Brasil, a Faculdade de Odontologia da Universidade de São Paulo (USP), em um projeto denominado Centro de Recursos de Aprendizagem e Investigação (CRAI), desenvolveu um *software* denominado "Minerador de Inovação" (QUINTO, 2020), que processa teses de doutorado de modo a cruzar informações e gerar dados e relatórios sobre toda a produção científica da biblioteca de Odontologia da USP.

Essas iniciativas inovadoras permitem que bibliotecários identifiquem padrões, reduzam custos e identifiquem tarefas repetitivas e passiveis de serem automatizada por sistemas inteligentes, de maneira a liberar a equipe para se concentrar em tarefas complexas.

Além disso, estes novos desafios disruptivos às tradicionais atividades desempenhadas pelas bibliotecas, forçam os gestores a repensarem os seus modelos e práticas gerenciais de modo a identificarem novas oportunidades e

otimizarem fluxos de trabalho.

À vista disso, sistemas de IA também podem auxiliar os bibliotecários na coleta e análise de dados acerca do comportamento do usuário, por meio de algoritmos de predição e inferência. A partir da geração destes dados é possível compreender hábitos e interações realizados nos sistemas, além de identificar necessidades informacionais se antecipar a demandas ainda não expressas pelos usuários.

Entretanto, à medida em que bibliotecas e fornecedores detêm grandes volumes de informações acerca dos hábitos de seus usuários, é necessário estar atento aos mecanismos de coleta e armazenamento de dados sensíveis, no que se refere as regras de privacidade dos dados pessoais.

Princípios fundamentais e éticos em sociedades de direito, devem ser respeitados, de modo a preservar a privacidade, o consentimento e o uso adequado dos dados de usuários de bibliotecas. E, tendo em vista essa preocupação, a IFLA (2020) publicou uma declaração, cujo objetivo é delinear as principais considerações sobre o uso de tecnologias de IA e AM em bibliotecas e sugerir os papéis que elas devem se esforçar para assumir em uma sociedade em crescente integração com a IA. Desta forma, a IFLA produziu recomendações para os principais atores da sociedade ligados à causa: governos, bibliotecas e associações de bibliotecas. Neste documento a IFLA destacou preocupações acerca dos padrões éticos, da privacidade ou equidade, da liberdade intelectual, da liberdade de expressão, do direito autoral dentre outros princípios que devem nortear os usos da IA em bibliotecas.

Além da IFLA, uma outra instituição que demostrou preocupação acerca da IA e da liberdade intelectual é a Federação Canadense de Bibliotecas (CFLA-FCAB). A FCAB produziu um documento, resultado de um fórum sobre IA, o qual alertou em um dos painéis (FCAB, 2018, p. 4) acerca dos efeitos negativos da IA e dos riscos potenciais, incluindo o viés humano em programação e no desenvolvimento de sistemas, bem como os potenciais vieses que podem ser reforçados quando os sistemas de IA são treinados utilizando conjuntos de dados de fontes questionáveis, ou que apresentem dados incompletos, incorretos ou tendenciosos.

À vista disso, tanto a IFLA (2020) como a FCAB (2018), defende que é necessário que usuários compreendam como os algoritmos e outros processos digitais impactam na maneira como eles acessam e recebem informação. Desta forma, a IFLA produziu um anexo em sua declaração e em destacou a importância da alfabetização digital e algorítmica (Anexo II da Declaração).

No entanto, assim como os usuários, os profissionais da informação devem buscar caminhos para lidar com a complexidade deste novo cenário. E, desta forma, é necessário que estes adquiram novas habilidades por meio de cursos de capacitação e atualização, de modo a aplicar técnicas de IA em bibliotecas, bem como replicar o conhecimento básicos sobre tecnologia para seus usuários, de modo a auxiliá-los na obtenção de competências tecnológicas para fins educacionais, laborais e pessoais. Considerando esse contexto, conhecer os algoritmos usados para resolver problemas específicos, as aplicações e os usos implementados por pesquisas na área torna-se um elemento estratégico e tático para avanço das pesquisas na área da CI

Deste modo, na próxima seção serão abordados os princípios metodológicos utilizados para se avaliar o uso e a presença dos algoritmos de aprendizagem de máquina na área da Ciência da Informação.

5 METODOLOGIA

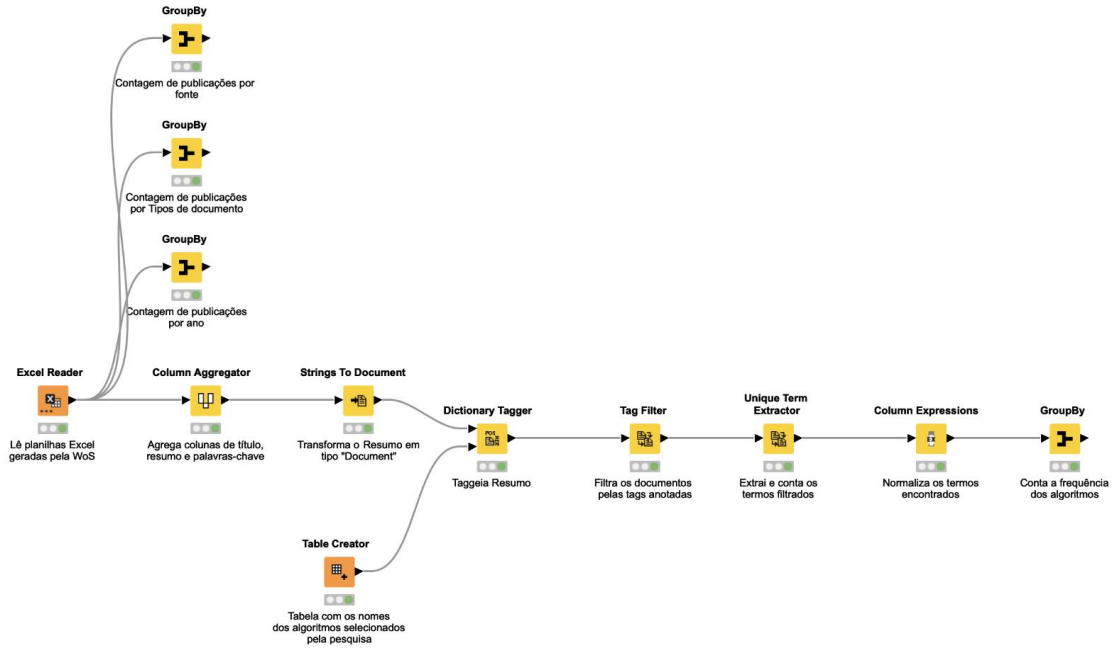
Utilizou-se para a realização da pesquisa a base de dados Web of Science (WoS) com a expressão de busca "*machine learning*" OR "*data mining*" em todos os campos e sem corte temporal nos resultados. A expressão foi construída incluindo o termo "mineração de dados" pois ele é citado como sendo um sinônimo de AM (SILVA; FERRARI, 2016, p. 14). Os resultados foram filtrados para separar apenas aqueles indexados na área "*Library and Information Science*".

A WoS permite exportar os arquivos com os resultados de busca em diferentes formatos. Os dados foram baixados em formato Excel e foram tratados utilizando o aplicativo KNIME2. O KNIME é um aplicativo para a construção visual de fluxos de processos de ciência de dados, disponibilizando recursos para as etapas de coleta, pré-processamento e modelagem de dados utilizando técnicas

de AM. O fluxo criado para a presente pesquisa é apresentado na figura 4.

O fluxo é composto da etapa de coleta (nó Excel Reader) onde todos os arquivos obtidos da WoS são lidos e agregados em uma única tabela. Os 3 nós acima da imagem (nós GroupBy) são utilizados para contagem do número de documentos por fontes de publicação, número de documentos por tipo de documento (na pesquisa foram coletados artigos, capítulos de livros, comunicações científicas, entre todos os outros tipos fornecidos pela WoS) e o número de documentos por ano. Seguindo após o nó de coleta de dados, as colunas de "Título", "Resumo" e "palavras-chave" foram agregadas para se buscar os termos dos algoritmos nesses 3 campos (nó Column Aggregator). Em seguida, a coluna resultado da agregação é transformada em um tipo de dados denominado "documento" para análise textual pelo KNIME (nó String to Document). Os documentos são analisados por um dicionário (nó Dictionary Tagger) que pesquisa texto a texto para identificar todos os termos cadastrados em uma tabela de apoio (nó Table Creator) ao dicionário. Por fim, os resultados são filtrados (nó Tag Filter e Unique Term Extractor) para ficar apenas os termos anotados pelo dicionário, normalizados (nó Column Expressions) para corrigir diferentes formas de escrita e agrupados para a contagem dos termos (nó GroupBy).

Figura 4 - Fluxo de processos de tratamento de dados implementados no KNIME



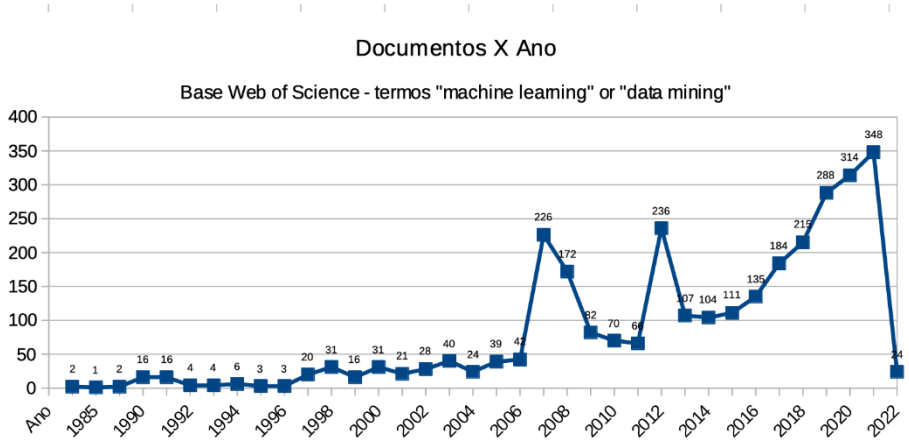
Fonte: dos autores.

Uma vez configurado e validado o fluxo de processos é possível tratar automaticamente grandes quantidades de dados. Tal recurso automatiza parte importante da identificação dos termos de interesse da pesquisa e reduz erros em potencial pelo processamento humano dos termos. Cabe ressaltar que a denominação dos algoritmos de AM utilizada para anotação nos dados coletados bem como sua tipologia foi extraída de Sarker (2021).

6 RESULTADOS

Foram identificados 3111 documentos a partir dos critérios da pesquisa mencionados na seção Metodologia. A distribuição do número de documentos por ano pode ser vista na figura 5. Pode-se notar que as primeiras menções aparecem no ano de 1984, permanecendo em número de pequeno volume até o ano de 1997. A partir de 1998 ocorre um primeiro salto na quantidade de documentos que se mantém constante até aproximadamente 2006. A partir desse período, observa-se dois picos importantes nos anos de 2007 e 2013 e um crescimento linear e contínuo do ano de 2014 em diante. Cabe ressaltar que o ano de 2022 apresenta baixa quantidade de documentos pelo fato da pesquisa ter sido realizada no primeiro mês do ano. Optou-se por incluir esses documentos para não perder as tendências mais recentes de menção dos algoritmos de AM.

Figura 5 - Distribuição dos documentos X Ano.



Fonte: dos autores.

As 10 fontes de documentos com maior frequência são apresentadas na tabela 1. Essas fontes são responsáveis por 45,2% dos documentos identificados. É notável que a interface entre a área médica e a CI é uma das mais importantes áreas de menção dos algoritmos de AM na presente pesquisa. Entende-se que essas fontes podem ser estratégicas de monitorar para se acompanhar as tendências da área. Chama a atenção que a revista Scientometrics é uma das que mais apresenta menções, indicando potencialmente que o campo da cientometria pode ser uma das áreas da CI que mais tem se beneficiado dos algoritmos de AM.

Tabela 1 - As 10 principais fontes de documentos

Fonte do documento	Documentos
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	372
INFORMATION PROCESSING & MANAGEMENT	192
2012 6TH INTERNATIONAL CONFERENCE ON NEW TRENDS IN INFORMATION SCIENCE, SERVICE SCIENCE AND DATA MINING (ISSDM2012)	158
FIRST INTERNATIONAL WORKSHOP ON KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS	145
SCIENTOMETRICS	129
INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE	121
DATA ANALYSIS, MACHINE LEARNING AND APPLICATIONS	83
JOURNAL OF INFORMATION SCIENCE	78
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY	70
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY	57

Fonte: dos autores.

O resultado objetivo da presente pesquisa é apresentado na tabela 2. A tabela inicia com a coluna "Técnicas de aprendizagem de máquina" classificando os algoritmos por uma tipologia de técnicas apresentada por Sarker (2021). Segue-se a coluna "Algoritmos" que apresenta o nome específico do algoritmo, a coluna "Frequência do Termo" que apresenta a quantidade de vezes que aquele termo apareceu na base de dados da pesquisa, a coluna "Frequência de documentos" que apresenta a quantidade de diferentes documentos que o termo apareceu e a última coluna "%FD" mostra a quantidade relativa de documentos em que um algoritmo foi mencionado.

Tabela 2 - Frequência de termos e documentos que referenciam os algoritmos de AM

Técnicas de aprendizagem de máquina	Algoritmo	Frequência do Termo	Frequência de documentos	% (FD)
Análise de Classificação	SUPPORT VECTOR MACHINE	168	143	4,6%
Análise de Classificação	DECISION TREE	164	112	3,6%
Análise de Classificação	RANDOM FOREST	156	123	4,0%
Análise de Classificação	RULE-BASED	122	82	2,6%
Análise de Classificação	NAIVE BAYES	112	85	2,7%
Análise de Classificação	LOGISTIC REGRESSION	102	81	2,6%
Análise de Agrupamento	K-MEANS	67	39	1,3%
Redes neurais e deep learning	CONVOLUTIONAL NEURAL NETWORK	48	34	1,1%
Análise de Redução de dimensionalidade	PRINCIPAL COMPONENT ANALYSIS	30	23	0,7%
Análise de Regras de associação	APRIORI	28	18	0,6%
Análise de Regressão	LINEAR REGRESSION	25	19	0,6%
Aprendizado por reforço	MONTE CARLO	12	7	0,2%
Análise de Classificação	ADAPTIVE BOOSTING	11	5	0,2%
Análise de Agrupamento	DBSCAN	9	5	0,2%
Análise de Classificação	EXTREME GRADIENT BOOSTING	8	6	0,2%
Análise de Classificação	K-NEAREST NEIGHBORS	8	8	0,3%
Redes neurais e deep learning	MULTILAYER PERCEPTRON	7	6	0,2%
Análise de Redução de dimensionalidade	ANOVA	5	4	0,1%
Análise de Regras de associação	FP-GROWTH	5	3	0,1%
Análise de Classificação	LINEAR DISCRIMINANT ANALYSIS	5	4	0,1%
Análise de Agrupamento	AGGLOMERATIVE HIERARCHICAL	4	3	0,1%
Análise de Agrupamento	GAUSSIAN MIXTURE MODELS	4	2	0,1%
Análise de Redução de dimensionalidade	PEARSON CORRELATION	4	4	0,1%
Análise de Redução de dimensionalidade	RECURSIVE FEATURE ELIMINATION	2	2	0,1%
Análise de Regras de associação	AIS	1	1	0,0%
Análise de Classificação	STOCHASTIC GRADIENT DESCENT	1	1	0,0%

Fonte: dos autores.

Destaca-se logo de início a importância dos problemas denominados "Análise de Classificação" que são responsáveis por 79,3% da frequência somada de documentos na tabela 01. Com base nesses dados, pode-se inferir que o uso das técnicas de classificação, uma abordagem supervisionada da AM, é uma das áreas que têm demonstrado maior interesse no desenvolvimento de pesquisas e produção científica da CI em termos internacionais. É essa temática que concentra os algoritmos mais referenciados, sendo eles o *Support Vector Machine* (4,6% dos documentos), *Decision Tree* (3,6%), *Random Forest* (4,0%), *Rule-based* (2,6%), *Naive-Bayes* (2,7%) e *Logistic Regression* (2,6%).

Na sequência, aparecem técnicas de agrupamento com a menção do algoritmo *K-Means* (1,3%) e logo em seguida as redes neurais e *deep learning* (1,1%) com a menção do algoritmo *Convolutional Neural Network*. Todos os outros resultados apresentam valores de menos de 1% dos documentos encontrados, demonstrando problemas de menor volume de produção científica, apontando temas de menor interesse ou temas ainda emergentes e que estão no início de seu interesse científico.

7 CONCLUSÃO

A presente pesquisa apresentou um método de automação da identificação de menções de algoritmos de AM nos títulos, resumos e palavras-chave de documentos científicos obtidos pela WoS. A metodologia se mostrou adequada para o volume de dados tratados e apresenta resultados que podem facilitar novos pesquisadores e grupos de pesquisa no Brasil acompanharem as técnicas mais utilizadas, os algoritmos mais usados e fontes de publicação mais usadas. Tal objetivo visa apoiar o domínio do tema e estimular a pesquisa na área facilitando a construção de currículos de disciplinas e o desenvolvimento de escopo de projetos aplicados que favoreçam o desenvolvimento de conhecimento na área da CI no país. Futuras pesquisas pretendem ampliar a investigação usando outras formas de denominar os algoritmos, incluindo siglas e outras expressões disponíveis na literatura.

O tema da aprendizagem de máquina e seu potencial de aplicações tem se tornado cada vez mais promissor e entende-se que dominar seus princípios técnicos, científicos tem tanto aplicações para a pesquisa quanto para a dimensão do ensino de graduação e pós-graduação. Conhecer seus fundamentos e entender seus desdobramentos torna-se fundamental para o próprio desenvolvimento da área da CI no Brasil. Logo, conhecer os algoritmos aqui expostos como os mais usados, compreender suas aplicações, os problemas que resolvem e desenvolver formas de ensino e aprendizagem de seus princípios torna-se uma pauta que pode contribuir de forma sistemática ao avanço dos esforços na área no Brasil.

REFERÊNCIAS

BARBOSA, X. de C.; BEZERRA, R. F. Breve introdução à história da inteligência artificial. **Jamaxi**, Rio Branco, v. 4, n. 1, 2020, p. 90-97. Disponível em: https://periodicos.ufac.br/index.php/jamaxi/article/view/4730. Acesso em: 3 set. 2021.

BIBLIOTECA NACIONAL (França). **BnF and Artificial intelligence**. Paris: BnF, 2021 Disponível em: https://www.bnf.fr/en/feuille-de-route-ia. Acesso em: 1 set. 2022.

BRYGFIELD, S. A.; WETJEN, F.; WALSØE, A. Machine learning for production of Dewey Decimal. **IFLA WLIC**, 2018, Kuala Lumpur. Disponível em: http://library.ifla.org/id/eprint/2216/1/115-brygfjeld-en.pdf. Acesso em: 23 set. 2021.

CORDELL, Ryan. **Machine Learning + Libraries: A Report on the State of the Field.** Washington, Estados Unidos da América: Biblioteca do Congresso Americano, 2020. 97p. Disponível em: http://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf. Acesso em 30/01/2022.

EUROPEANA. **AI in relation to GLAMS task force: report and recommendations.** [Amsterdam]: EUROPEANA, 2021. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/AI%20in%20relation%20to%20GLAMs%20Task%20Force%20Report.pdf. Acesso em: 27 ago. 2022.

FACELI, K. *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina.** Rio de Janeiro: LTC, 2019.

FEDERAÇÃO CANADENSE DE ASSOCIAÇÕES DE BIBLIOTECAS (FCAB). Artificial intelligence and intellectual freedom, key policy concerns for Canadian libraries. **CFLA-FCAB Forum**, [S. l.], 2 may 2018. Disponível em: http://cfla-fcab.ca/wp-content/uploads/2018/07/CFLA-FCAB-2018-National-Forum-Paper-final.pdf. Acesso em: 23 set. 2021.

HAMMAIS, E.; KETAMO, H.; KOIVISTO, A. Virtual information assistants on mobile app to serve visitors at Helsinki Central Library Oodi. **IFLA WLIC**, 2019, Atenas. Disponível em: http://library.ifla.org/id/eprint/2536/1/114-hammais-en.pdf. Acesso em: 23 set. 2021.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **The IFLA Library is IFLA's institutional repository.** IFLA: Haia, 2021 Disponível em: http://library.ifla.org/cgi/search/advanced. Acesso em: 23 set. 2021.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **IFLA Statement on Libraries and Artificial Intelligence**. The Hague, Holanda: International Federation of Library Associations and Institutions, 2020. 14p. Disponível em: https://www.ifla.org/publications/node/93397. Acesso em 30/01/2022.

KIM, B. A new tech revolution: AI, big data, and other disruptive technology. **American Libraries Magazine**, [S. l], 1 may 2020. Disponível em: https://americanlibrariesmagazine.org/2020/05/01/new-tech-revolution/. Acesso em: 23 set. 2021.

LIGHTHILL, J. **Artificial intelligence: a general survey.** Lighthill report, 1972, [S. l.: s. n.]. Disponível em: http://www.chilton-

computing.org.uk/inf/literature/reports/lighthill_report/p001.htm. Acesso em: 3 set. 2021.

LORANG, Elizabeth, LEEN-KIAT, Soh, YI, Liu, e CHULWOO, Pack, Digital Libraries, **Intelligent Data Analytics, and Augmented Description: A Demonstration Project**, Nevada, Estados Unidos da América: University of Nevada, 2020. 47p. Disponível em: https://digitalcommons.unl.edu/libraryscience/396/. Acesso em 30 jan. 2022.

MCCARTHY, J. *et al.* **A proposal for the Dartmouth summer research project on artificial intelligence.** [S. l.: s. n], 1955. Disponível em: http://www-formal.stanford.edu/jmc/history/dartmouth.pdf. Acesso: 3 set. 2021.

MCCARTHY, J. **What is artificial intelligence?** Stanford: Universidade de Stanford, 2007. Disponível em: http://35.238.111.86:8080/jspui/bitstream/123456789/274/1/McCarthy_John_What%20is%20artificial%20intelligence.pdf. Acesso em: 3 set. 2021.

ORACLE. **What's the difference between AI, Machine Learning, and Deep Learning?** [S. l.:] Oracle, 11 jul. 2018. Disponível em: https://blogs.oracle.com/bigdata/post/whatx27s-the-difference-between-ai-machine-learning-and-deep-learning. Acesso em: 3 set. 2021.

QUINTO, A. C. Inteligência artificial agiliza busca pela inovação em biblioteca. **Jornal da USP**, São Paulo, 11 mar. 2020. Disponível em: https://jornal.usp.br/ciencias/ciencias-exatas-e-da-terra/inteligencia-artificial-agiliza-busca-pela-inovacao-em-biblioteca/. Acesso em: 23 set. 2021.

RUSSELL, S.; NORVIG, P. **Inteligência artificial**. 3. ed. Rio de Janeiro: LTC, 2021. *E-book.*

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal,** [S. l.], v. 3, n. 3, jul. 1959, p. 535-554. Disponível em: https://www.cs.virginia.edu/~evans/greatworks/samuel1959.pdf. Acesso em 3 set. 2021.

SARKER, Igbal H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**. N. 2, v. 160, 2021. 21p. Disponível em: https://doi.org/10.1007/s42979-021-00592- . Acesso em 30/01/2022.

SEARLE, J. R. Minds, brains, and programs. **Behavioral and Brain Science**, [S. l.] v. 3, n. 3, 1980, p. 417-457. Disponível em: http://cogprints.org/7150/1/10.1.1.83.5248.pdf. Acesso: 3 set. 2021.

SILVA, Leandro Nunes de Castro; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados. Conceitos Básicos, Algoritmos e Aplicações**. 2. ed. São Paulo: Saraiva, 2016. 376p.

SMITH, L. Artificial intelligence in information retrieval systems. **Information**

processing & management, [S. l], v. 12, n. 3, 1976, p. 189-222. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/0306457376900054. Acesso em: 1 set. 2022.

TAULLI, T. **Introdução à inteligência artificial:** uma abordagem não técnica. 1. ed. São Paulo: Novatec, 2020.

VIJAYAKUMAR, S.; SHESHADRI. Applications of artificial intelligence in academic libraries. **International Journal of Computer Science and Engineering**, [S. l], v. 7., n. esp. 16, may 2019. Disponível em: https://ijcseonline.org/full_spl_paper_view.php?paper_id=1294. Acesso em: 23 set. 2021.

MACHINE LEARNING AND INFORMATION SCIENCE: CONTRIBUTIONS TO A RESEARCH AND TEACHING AGENDA IN THE BRAZILIAN INFORMATION SCIENCE

ABSTRACT
Introduction: The article discusses and presents the concepts of artificial intelligence, machine learning and discusses potential applications and reference technique studies in the area of Information Science with greater emphasis in the field of libraries and librarianship. **Objective:** it intends to support the construction of a research and teaching agenda on the subject of machine learning through knowledge of its main applications and most used algorithms. **Methodology:** article studies references to machine learning algorithms in indexed documents in the area of Information Science in the Web of Science database. **Results and Conclusions:** evaluates 3111 identified documents and concludes that the supervised approach through classification techniques is the most used in the field, with evidence for the Support Vector Machine, Decision Tree, Random Forest algorithms.

Descriptors: Machine Learning. Information Science. Algorithms. Web of Science. Teaching. Research.

APRENDIZAJE AUTOMÁTICO Y CIENCIAS DE LA INFORMACIÓN: CONTRIBUCIONES PARA UNA AGENDA DE INVESTIGACIÓN Y ENSEÑANZA EN LA CIENCIAS DE LA INFORMACIÓN BRASILEÑA

RESUMEN
Introducción: El artículo discute y presenta los conceptos de inteligencia artificial, aprendizaje automático y discute posibles aplicaciones y estudios de técnicas de referencia en el área de las Ciencias de la Información con mayor énfasis en el campo de las bibliotecas y la biblioteconomía. **Objetivo:** pretende apoyar la construcción de una agenda de investigación y docencia en el tema de aprendizaje automático a través

del conocimiento de sus principales aplicaciones y algoritmos más utilizados. **Metodología:** el artículo estudia las referencias a algoritmos de aprendizaje automático en documentos indexados del área de Ciencias de la Información en la base de datos Web of Science. **Resultados y Conclusiones:** evalúa 3111 documentos identificados y concluye que el enfoque supervisado a través de técnicas de clasificación es el más utilizado en campo, con evidencia para los algoritmos Support Vector Machine, Decision Tree, Random Forest.

Descriptores: Aprendizaje automático. Ciencias de la Información. Algoritmos. Web de la Ciencia. Enseñando.