UNIVERSIDADE DE BRASÍLIA (UNB)
FACULDADE DE CIÊNCIA DA INFORMAÇÃO (FCI)
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
(PPGCINF)

PERSPECTIVAS DO USO DO APRENDIZADO DE MÁQUINA EM BIBLIOTECAS: uma revisão sistemática de literatura

Discente: Rafaella Carine Monterei
Orientador: Professor Dr. Dalton Lopes Martins

Brasília, 28 de setembro de 2022.

UNIVERSIDADE DE BRASÍLIA (UNB)
FACULDADE DE CIÊNCIA DA INFORMAÇÃO (FCI)
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (PPGCINF)

PERSPECTIVAS DO USO DO APRENDIZADO DE MÁQUINA EM BIBLIOTECAS: uma revisão sistemática de literatura

Dissertação de mestrado apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Ciência da Informação. Área de concentração: Gestão, Organização e Comunicação da Informação e do Conhecimento. Linha de pesquisa: Gestão, tecnologias e organização da informação e do conhecimento.

Discente: Rafaella Carine Monterei

Orientador: Professor Dr. Dalton Lopes Martins

Brasília, 28 de setembro de 2022.

Reitora da Universidade de Brasília
Profª. Drª. Márcia Abrahão Moura

Vice-Reitor da Universidade de Brasília
Prof. Dr. Enrique Huelva Unternbäumen

Diretor da Faculdade de Ciência da Informação
Prof. Dr. Renato Tarciso Barbosa de Sousa

Coordenador do Programa de Pós-Graduação em Ciência da Informação
Prof. Dr. Dalton Lopes Martins

Dados Internacionais de Catalogação na Publicação (CIP)

M778p

Monterei, Rafaella Carine.
Perspectivas do uso do aprendizado de máquina em bibliotecas [recurso eletrônico] : uma revisão sistemática de literatura ; orientador: Dalton Lopes Martins. -- Brasília, 2022.
Dados eletrônicos (1 arquivo : PDF 145 páginas).
Dissertação (Mestrado em Ciência da Informação) -- Universidade de Brasília, 2022.
Disponível em: https://bce.unb.br/bibliotecas-digitais/respositorio/

1. Aprendizado de máquina. 2. Biblioteca, inovação tecnológica, revisão sistemática de literatura. 3. Machine learning. 4. Inteligência artificial. I. Dalton Lopes Martins, orient. II Título.

CDU 027:004.8

Ficha catalográfica elaborada pela bibliotecária Rafaella Carine Monterei CRB 1/2537

MONTEREI, Rafaella Carine. Perspectivas do uso do aprendizado de máquina em bibliotecas: uma revisão sistemática de literatura. Orientador: Dalton Lopes Martins. 2022. 155 p. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2022.

Universidade de Brasília

FOLHA DE APROVAÇÃO

Título: PERSPECTIVAS DO USO DO APRENDIZADO DE MÁQUINA EM BIBLIOTECAS: uma revisão sistemática de literatura

Autor (a): Raffaella Carine Monterei

Área de concentração: Gestão, Organização e Comunicação da Informação e do Conhecimento

Linha de pesquisa: Gestão, Tecnologias e Organização da Informação e do Conhecimento

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **MESTRE** em Ciência da Informação.

Dissertação aprovada em: 28 de setembro 2022.

Presidente (UnB/PPGCINF): Dalton Lopes Martins

Membro Externo (UFES): Daniela Lucas da Silva Lemos

Membro Interno (UnB/PPGCINF): Marcio de Carvalho Victorino

Suplente (UnB/PPGCINF): João de Melo Maricato

Em 14/09/2022.

Documento assinado eletronicamente por **Dalton Lopes Martins**, **Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 03/10/2022, às 09:58, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Daniela Lucas da Silva Lemos**, **Usuário Externo**, em 05/10/2022, às 15:19, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Marcio de Carvalho Victorino**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 05/10/2022, às 18:33, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **8679928** e o código CRC **98D588F3**.

Dedico esta dissertação àqueles que estiveram ao meu lado durante essa caminhada.

AGRADECIMENTOS

Agradeço a Deus em suas mais diversas formas ou manifestações, por me permitir alcançar os meus objetivos.

Agradeço à minha mãe Kátia e à minha irmã Gabriella que apesar dos percalços e surpresas que ocorrem ao longo deste período conturbado de Covid-19, apoiaram-me a seguir em frente e acreditaram no meu potencial.

Agradeço ao meu orientador Dalton por conseguir traduzir a minha necessidade de informação e sugerir essa pesquisa.

Agradeço aos meus colegas de trabalho, em especial à Rosa e à Renata minhas chefes à época por autorizarem o meu afastamento para estudo. Alda, Dudu, Fran, Gabriela, Laila, Najla, Roberta, Rodrigo, Ronaldo e Tiago por me incentivarem a continuar essa jornada.

Agradeço aos membros da banca pelas valiosas contribuições durante a qualificação.

Agradeço à Universidade de Brasília por me proporcionar uma formação gratuita e de qualidade desde o período da graduação. Tenho muito orgulho de fazer parte desta instituição.

Agradeço também *in memoriam* à minha Tequinha por acompanhar comigo as aulas e por ter feito os meus dias mais felizes.

Agradeço ao meu pai Sandro *in memoriam* que se preocupou em me oferecer uma educação de qualidade.

Por fim, e não menos importante! Agradeço a pessoa que colocou o filtro de floresta durante a apresentação virtual da minha defesa. Saiba que com este gesto inesperado você me deixou mais tranquila!

O correr da vida embrulha tudo.
A vida é assim: esquenta e esfria,
aperta e daí afrouxa,
sossega e depois desinquieta.
O que ela quer da gente é coragem.

Guimarães Rosa

RESUMO

O presente trabalho tem por finalidade apresentar as aplicações da Inteligência Artificial, com ênfase em *machine learning*, em bibliotecas, cujo objetivo principal é mapear benefícios e impactos que o aprendizado de máquina pode oferecer para o desenvolvimento de produtos e serviços em bibliotecas. A fim de atender este objeto, o estudo se pautará em uma pesquisa de caráter qualitativo e quantitativo, com a abordagem exploratória, de natureza pura, por meio do uso da pesquisa bibliográfica. E, para realizar tal investigação, recorre-se à revisão sistemática de literatura, por meio da produção de um protocolo de pesquisa, baseado nas diretrizes propostas por Galvão e Ricarte (2020) para o campo da Ciência da Informação, complementados pelos estudos produzidos por Kitchenham (2004) e Felizardo et al (2017) para o campo da Ciência da Computação. Por fim, conclui-se que este estudo proporciona ao pesquisador refletir e identificar novos fenômenos nas relações interdisciplinares entre a Ciência da Informação e a Inteligência Artificial.

Palavras-chave: Aprendizado de máquina; Produtos e serviços de bibliotecas; Inteligência Artificial; Biblioteca, inovação tecnológica; Unidade de informação, inovação tecnológica.

RESUMEN

El presente trabajo tiene como objetivo presentar las aplicaciones de la Inteligencia Artificial, con énfasis en el aprendizaje automático, en las bibliotecas, cuyo principal objetivo es mapear los beneficios e impactos que el aprendizaje automático puede ofrecer para el desarrollo de productos y servicios en las bibliotecas. Para cumplir con este objeto, el estudio se basará en una investigación cualitativa y cuantitativa, con un enfoque exploratorio, de carácter puro, mediante el uso de la investigación bibliográfica. Y, para llevar a cabo esta investigación, se utiliza una revisión sistemática de la literatura, a través de la producción de un protocolo de investigación, basado en las directrices propuestas por Galvão y Ricarte (2020) para el campo de las Ciencias de la Información, complementado con estudios producidos por Kitchenham (2004) y Felizardo et al (2017) para el campo de las Ciencias de la Computación. Finalmente, se concluye que este estudio permite al investigador reflexionar e identificar nuevos fenómenos en las relaciones interdisciplinarias entre las Ciencias de la Información y la Inteligencia Artificial.

Palabras llave: Aprendizaje automático; Productos y servicios bibliotecarios; Inteligencia artificial; Biblioteca, innovación tecnológica; Unidad de información, innovación tecnológica.

ABSTRACT

The present work aims to present the applications of Artificial Intelligence, with emphasis on machine learning, in libraries, whose main objective is to map benefits and impacts that machine learning can offer for the development of products and services in libraries. In order to meet this object, the study will be based on a qualitative and quantitative research, with an exploratory approach, of a pure nature, through the use of bibliographic research. And, to carry out such an investigation, a systematic literature review is used, through the production of a research protocol, based on the guidelines proposed by Galvão and Ricarte (2020) for the field of Information Science, complemented by studies produced by Kitchenham (2004) and Felizardo et al (2017) for the field of Computer Science. Finally, it is concluded that this study allows the researcher to reflect and identify new phenomena in the interdisciplinary relationships between Information Science and Artificial Intelligence.

Keywords: Machine learning; Library products and services; Artificial intelligence; Library, technological innovation; Information unit, technological innovation.

LISTA DE ILUSTRAÇÃO

LISTA DE ABREVIATURAS

AAAI	Association for the Advancement of Artificial Intelligence
AI	Artificial Intelligence
AI	Inteligência Artificial
AIDA	Digital libraries, intelligence data analytics, and augmented description
ALA	American Library Association
ALIEN	Automated Library Information Exchange Network
AM	Aprendizado de máquina
AMPPD	The Audiovisual Metadata Platform Pilot Development
ANCIB	Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
ARPA	Advanced Research Projects Agency
BDBComp	Biblioteca Digital Brasileira de Computação
BDTD	Biblioteca Digital de Teses e Dissertações
BnF	Biblioteca Nacional da França
BRAPCI	Base de Dados Referencial de Artigos em Ciência da Informação
BRLR	Classificador de regressão logística
CAMPI	Computer Aided Metadata Generation of Photo Archives Initiative
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CDD	Classificação Decimal de Dewey
CDU	Classificação Decimal Universal
CFLA	Canadian Federation of Library Associations
CI	Ciência da Informação
Covid-19	Coronavirus Disease 2019
CRAI	Centro de Recursos de Aprendizagem e Investigação
CRIS	Current Research Information System
DBLP	Digital Bibliography & Library Project
DDA	Demand-Driven Acquisition

DPLA	Digital Public Library of America
E-1	Critério de exclusão 1
E-2	Critério de exclusão 2
E-3	Critério de exclusão 3
E-4	Critério de exclusão 4
E-5	Critério de exclusão 5
E-6	Critério de exclusão 6
E-7	Critério de exclusão 7
E-8	Critério de exclusão 8
E-9	Critério de exclusão 9
E-10	Critério de exclusão 10
E-11	Critério de exclusão 11
EAND	Eager Associative Name Disambiguation
EDK	Estimativa de densidade Kernel
ENANCIB	Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
ETDs	Teses e Dissertações eletrônicas
FAST	Faceted Application of Subject Terminology
FCAB	Federação Canadense de Associações de Bibliotecas
GIGO	Garbage in, garbage out
GLAMS	Galerias, Bibliotecas, Arquivos e Museus
Hamlet	How about Machine Learning Enhanced Theses
HILT	Human in the Loop
I-1	Critério de Inclusão 1
I-2	Critério de Inclusão 2
I-3	Critério de Inclusão 3
IBM	International Business Machines Corporation
IFLA	Federação Internacional de Associação de Bibliotecas e Instituições
IMLS	Instituto de Serviços de Museus e Bibliotecas
IoT	Internet das coisas
ISKO	International Society for Knowledge Organization

ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
ISTA	Information Science and Technology Abstracts
k-NN	k-Nearest Neighbors, k-Vizinhos mais próximos
LAND	Lazy Associative Name Disambiguation
LC	Library of Congress
LCSH	Library of Congress Subject Headings
LDA	Latent Dirichlet Allocation
LIBRIS	Library Information System
LISTA	Library and Information Science Abstracts
MAUI	Multi-purpose Automatic Topic Indexing
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NLTK	Natural Language Toolkit
OAI-PMH	Open Archives Initiative Protocol of Metadata Harvesting
OCLC	Online Computer Library Center
OCR	Reconhecimento óptico de caracteres
OMS	Organização Mundial da Saúde
PDF	Portable Document Format
PICO	População, Intervenção, Controle e Resultados (*Outcomes*)
PNL	Processamento de Linguagem Natural
PQDT	ProQuest Dissertations and Theses
RAND	Research and Development Corporation
RSL	Revisão Sistemática de Literatura
SAND	Self-Training Associative Name Disambiguator
SGDSVM	Stochastic Gradient Descent – Support Vector Machine
SLAND	Self-Training LAND
SQL	Structured Query Language
SVM	Support Vector Machine
TF-IDF	Frequência do termo-frequência inversa dos documentos
TICs	Tecnologia da Informação e da Comunicação
UCLA	Universidade da Califórnia em Los Angeles
UFBA	Universidade Federal da Bahia

UFMG	Universidade Federal de Minas Gerais
USP	Universidade Federal de São Paulo
WoS	Web of Science
XAI	Inteligência Artificial Explicável
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language

SUMÁRIO

1 INTRODUÇÃO

Tecnologia, segundo a sua origem etimológica[1], deriva do grego *τεχνολογία*, designa conhecimento técnico e científico e suas aplicações a um campo particular. Desta forma, tendo em vista as suas diversas possibilidades de aplicação em conformidade com a sua acepção semântica, o advento das tecnologias, desencadeou profundas transformações nas estruturas sociais, econômicas, científicas, culturais e políticas da sociedade ao longo do tempo. E, segundo Saracevic (1996, p. 58) desempenhou um importante papel na ecologia informacional (o chamado efeito de Gutemberg), bem como na evolução da sociedade contemporânea.

Assim, neste cenário de profundas transformações, as tecnologias digitais bem como a informação, objeto de estudo da Ciência da Informação, apresentam-se como elementos estruturantes da sociedade do conhecimento, além de componentes basilares da 4ª Revolução Industrial, ou Indústria 4.0, marcado pela convergência entre o digital, o físico e o biológico.

À vista disso, as aplicações em tecnologias se manifestaram em diferentes ramos do saber, inclusive na Ciência da Informação (CI) desde a sua gênese, por meio de suas relações interdisciplinares, sobretudo entre a Ciência da Computação e a Inteligência Artificial (IA). Saracevic (1996, p. 50) destaca ainda a IA como “uma das áreas chave de interesse para ambas, Ciência da computação e CI”, situando-a como uma área de estudos da Ciências Cognitivas.

Caracterizada historicamente em um período de crescimento exponen
da informação, a Ciência da Informação se viu desafiada em tornar acessív
organizável e recuperável o conhecimento humano registrado. E, neste contexto, Saracevic (1996) afirma que Vannevar Bush (1945) identificou o problema da explosão informacional, particularmente em ciência e tecnologia. E, desta forma, Bush (1945) propôs o desenvolvimento de um dispositivo denominado MEMEX capaz de armazenar um conjunto de documentos, e recuperá-los por meio de mecanismos associativos, tal como a mente humana realiza.

[1] Informações extraídas verbete tecnologia do dicionário Michaelis.

Apesar deste dispositivo não ter sido desenvolvido, a ideia de Bush ainda permanece atual, tendo em vista que a velocidade de criação de dados e informações, se expandiu em um ritmo frenético, consequência dos avanços tecnológicos em *hardware* e *software,* que ampliaram a capacidade de processamento e armazenamento dos dispositivos eletrônicos. Este cenário exponencial ainda figura como um dos desafios centrais da CI que é tornar acessível, organizável e recuperável um conjunto de informações. Todavia, em um ambiente de profusão de informações é necessário selecionar, reunir e extrair conhecimento de qualidade e com precisão, devido a um ambiente cada vez mais heterogêneo, em termos de novos formatos e suportes, e com dados cada vez mais dispersos.

Tendo em vista os novos desafios impostos pela transformação digital da sociedade, as bibliotecas são instituições que se moldam às demandas de seu público. Essas transformações foram materializadas, por meio da 5ª Lei de Ranghanathan, o qual afirma que as bibliotecas são um organismo vivo[2], ou seja, bibliotecas são instituições dinâmicas, interativas e que devem estar em permanente diálogo com o seu usuário, de forma a identificar novas necessidades e comportamentos e, assim traduzi-las em novos produtos e serviço de informação personalizados para o seu público.

Neste contexto, as bibliotecas trabalham em meio a um ambiente competitivo onde os grandes buscadores comerciais oferecem resultados rápidos e acesso instantâneo aos conteúdos, enquanto os catálogos das bibliotecas costumam ser mais lentos e exigir certas habilidades de pesquisa dos usuários (BOMHOLD, 2013, p. 431). Todavia, as bibliotecas detêm uma vantagem competitiva frente aos buscadores comerciais, pois oferecem informações de qualidade e com precisão, baseados em políticas de desenvolvimento de suas coleções, com diretrizes orientadas aos usuários e que procuram eliminam vieses ou interesses comerciais escusos.

Deste modo, para que as bibliotecas cumpram a sua missão social neste novo cenário, é necessário estimular o profissional da informação a refletir acerca das possíveis aplicações das novas tecnologias digitais em unidade de

[2] “Um organismo em crescimento absorve matéria nova, elimina matéria antiga, muda de tamanho e assume novas aparências e formas” (RANGANATHAN, 2009, p. 241).

informação, bem como discutir junto aos seus pares, as novas perspectivas de atuação frente ao novo cenário. Assim, discussões acerca das aplicações em inteligência artificial e aprendizado de máquina apresentam-se como temas relevantes para profissionais que buscam novas habilidades e áreas de atuação em bibliotecas, tendo em vista o contexto da 4ª Revolução Industrial. Além disso, o desenvolvimento de novas competências profissionais, podem auxiliar as unidades de informação a produzirem soluções inovadoras e customizadas para o seu público.

Em uma rápida pesquisa em publicações sobre IA aplicada em bibliotecas no cenário brasileiro, constatou-se que as discussões ainda se encontram em fase inicial, como um tema emergente na CI brasileira. Um exemplo destas investigações podemos encontrar em Peres (2017), o qual buscou mapear em sua dissertação, os trabalhos produzidos sobre Inteligência Artificial à luz da Ciência da Informação em eventos da ANCIB, e concluiu que a literatura sobre o tema no Brasil ainda é escassa. Outros autores que enveredaram pesquisas sobre o tema foram Silva e Nathansohn (2018, p. 121), o qual em trabalho publicado no ENANCIB, afirmaram que o campo da IA pode ser muito mais explorado pela área da Ciência da Informação, considerando a sua interdisciplinaridade entre a Ciência da Computação e a Ciência Cognitiva. Outra contribuição brasileira foi dada por Neves (2019, 2021), que observou que as pesquisas que tratam acerca da Inteligência Artificial e da computação cognitiva em unidades de informação são predominantemente estrangeiras.

Em relação ao âmbito internacional, a produção teórica e prática em IA e AM aplicada às bibliotecas, possui uma situação muito distinta da brasileira. A literatura publicada na área é substancial, e desta forma, podemos destacar algumas iniciativas, como o relatório produzido pela EUROPEANA (2021), que se propôs a investigar o impacto da Inteligência Artificial no âmbito do patrimônio cultural, em especial na análise de coleções. Outra iniciativa de destaque é uma parceria entre a Universidade de Nebraska-Lincoln e a Biblioteca do Congresso Americano, o qual desenvolveram um projeto para utilizar o aprendizado de máquina no processamento de imagens. Outro relevante projeto em IA foi produzido no Instituto de Tecnologia de Massachusetts (MIT), por meio do desenvolvimento de um sistema de aprendizado de máquina, cujo objetivo é

explorar e enriquecer as coleções de teses do MIT, desenvolvida por intermédio de uma rede neural denominada Hamlet (*How about Machine Learning Enhanced Theses*).

Estas iniciativas apresentadas, demostram o grande potencial que as pesquisas em IA/AM aplicadas em unidades de informação apresentam no desenvolvimento de novos produtos e serviços, dado que essas novas tecnologias computacionais auxiliam na classificação, no tratamento, na análise e na recuperação de grandes volumes de dados, de acordo com as necessidades e interesses de seus usuários.

E, diante do cenário brasileiro *sui generis* com aplicações quase inexistentes e poucas discussões, e dado a relevância da investigação, o interesse de pesquisa sobre este tema surgiu devido a uma sugestão do orientador em conjunto com o interesse desta pesquisadora em aprofundar discussões acerca da inteligência artificial, com ênfase em aprendizado de máquina, aplicado a ambientes de bibliotecas, em virtude de discussões iniciais sobre o tema no ambiente de trabalho.

À vista deste panorama instigante e revolucionário, é necessário que o profissional da informação desenvolva novas competências e habilidades no âmbito das transformações tecnológicas. E, tendo em vista a natureza disruptiva destas aplicações em unidades de informação e o cenário exponencial da produção de documentos nato-digitais e digitalizados, este estudo pretende apresentar uma visão panorâmica do estado da arte da Inteligência Artificial, com ênfase em aprendizado de máquina, no campo da Ciência da Informação, além de realizar um mapeamento dos seus possíveis usos e aplicações, a fim de compreender o papel do profissional da informação frente a estes novos desafios e, assim responder ao seguinte questionamento: *Como o* ***desenvolvimento de produtos e serviços*** *por meio do* ***aprendizado de máquina*** *podem impactar e beneficiar as* ***bibliotecas***?

À luz deste questionamento, o objetivo geral desta pesquisa é mapear benefícios e impactos em termos metodológicos e operacionais que o aprendizado de máquina pode oferecer para o desenvolvimento de produtos e serviços de informação em bibliotecas.

Diante disso, foram propostos os seguintes objetivos específicos, a saber: Mapear o estado da arte da inteligência artificial, com ênfase no aprendizado de máquina, no contexto da Ciência da Informação; Identificar possíveis aplicações do aprendizado de máquina em produtos e serviços de bibliotecas relatados na literatura; Identificar novas competências do profissional da informação frente ao uso de novas tecnologias; Refletir acerca das questões éticas geradas a partir do uso de novas tecnologias; Produzir um conjunto de recomendações para a aplicação de técnicas de aprendizado de máquina para bibliotecas e, por fim; Apresentar tendências futuras do campo.

Com base nestas informações, esta dissertação foi dividida em seis seções. A primeira seção apresentou um panorama introdutório acerca do cenário de transformação digital o qual as bibliotecas estão situadas. Além disso, foram apresentados a justificativa, a questão de pesquisa, bem como os objetivos para a produção desta dissertação.

Em seguida, a seção dois abordará os conceitos basilares que norteiam este trabalho, a Inteligência Artificial, o *Machine Learning*, finalizados com uma breve introdução das aplicações em bibliotecas, à luz da literatura científica.

A terceira seção versará acerca da metodologia utilizada para a construção deste trabalho, executada na quarta seção, onde serão apresentados os resultados da análise do *corpus* da Revisão Sistemática de Literatura (RSL).

A quinta seção buscará descrever as tendências e recomendações para a aplicação de técnicas de ML em biblioteca e, por fim, a última seção abordará as considerações finais da pesquisa.

Em suma, espera-se que este trabalho contribua para o desenvolvimento de discussões acerca das possibilidades de aplicação do aprendizado de máquina em bibliotecas.

A seguir serão apresentadas as linhas conceituais e históricas da inteligência artificial e do *machine learning*, bem como uma breve introdução do cenário atual da inteligência artificial aplicada as bibliotecas.

2 REFERÊNCIAL TEÓRICO

Este capítulo abordará os conceitos fundamentais para o desenvolvimento desta pesquisa, de modo a construir um arcabouço conceitual para a elaboração da Revisão Sistemática de Literatura (RSL).

2.1 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.

Nesta seara a Inteligência Artificial é definida por Russel e Norvig (2021) a partir de 4 categorias divididas em 2 dimensões, conforme dispõe o quadro abaixo:

Quadro 1: Definições da Inteligência Artificial

	Pensando como um humano	**Pensando racionalmente**
Processos de pensamento e raciocínio	"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)

	Agindo como seres humanos	Agindo racionalmente
Comportamento	"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole et al., 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: Adaptado de Norvig e Russel (2021)

Em linhas gerais, as definições reunidas por Russel e Norvig (2021) no quadro 1 estão subdivididas em processos de pensamento e raciocínio na parte superior do quadro e em aspectos relativos ao comportamento na parte inferior do quadro. Elas são fruto do movimento histórico do campo ao refletir diferentes linhas de pensamento a partir de diferentes métodos de aplicação da IA. Além disso, Russel e Norvig (2021) afirmam que a abordagem centrada nos seres humanos deve ser em parte uma ciência empírica, envolvendo hipóteses e confirmação experimental. Já a abordagem racionalista envolve uma combinação entre a matemática e a engenharia.

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), existe 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.

Ainda em relação ao primeiro evento sobre IA, Taulli (2020, p. 22-23) afirma que os pesquisadores Allen Newell, Cliff Shaw e Herbert Simon, apresentaram um programa de computador denominado de *Logic Theorist*, desenvolvido na *Reserch and Development Corporation* (RAND), este programa foi considerado o primeiro programa de IA já desenvolvido. Todavia, o evento em Dartmouth foi considerado uma enorme decepção (TAULLI, 2020, p. 23), por não apresentar nenhuma novidade (RUSSEL; NORVIG, 2021), no entanto, apesar do evento não ter suprido as expectativas da comunidade científica daquele período, a realização desta conferência foi crucial ao aproximar os principais nomes do campo.

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.

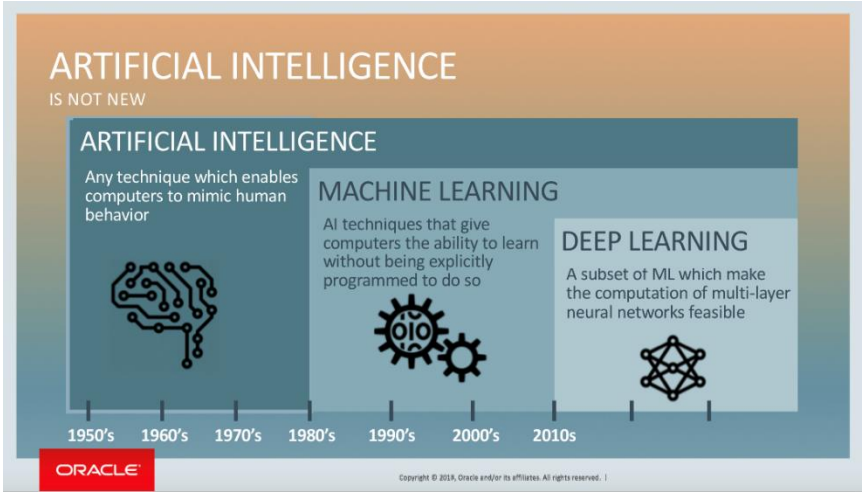
Após os anos de 1970 até 1980 perdurou um período batizado de “Inverno da IA”, o qual foi caracterizado por uma época de agravamento da crise econômica. O marco histórico do campo da IA ao longo deste período se deve a publicação do Relatório de Lighthill em 1973 pelo professor britânico Sir James Lighthill. 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

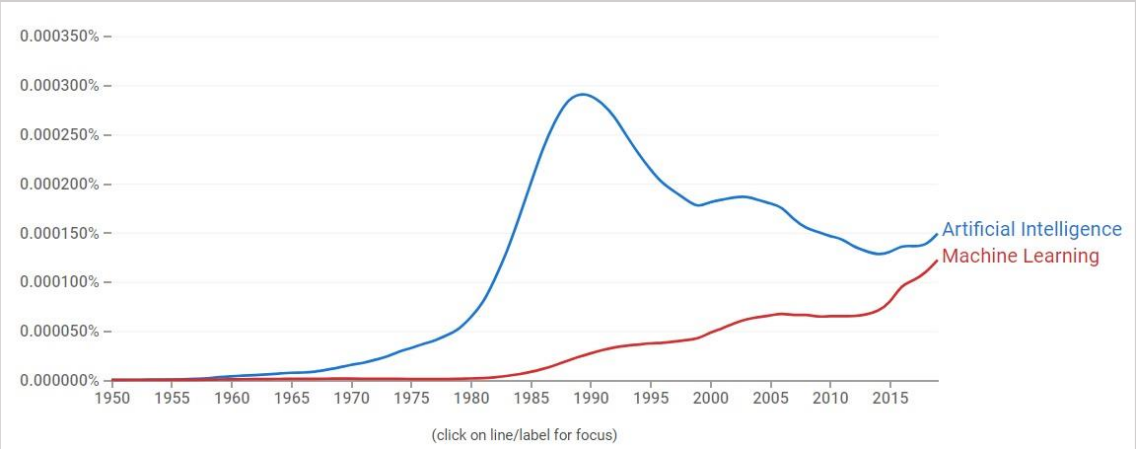


Fonte: Oracle (2018)

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.

Em consulta ao Google Ngram Viewer[3] para analisar o comportamento das expressões "*Artificial Intelligence*" e "*Machine Learning*" no *corpus* de obras disponibilizadas no Google entre o período de 1950 a 2019, observa-se que as contribuições de ambas as áreas estão caminhando para uma confluência em número de publicações, conforme imagem abaixo:

Figura 2: Comportamento das expressões AI/ML



Fonte: Google Ngram Viewer (2022a)

Tendo em vista este cenário, é necessário compreender um dos principais elementos impulsionadores de novos estudos sobre a *Artificial Intelligence* e por conseguinte do *Machine Learning*, os dados, que serão objeto de estudo da próxima seção.

2.1.1 Dados: noções básicas e categorias

Paralelamente aos avanços da tecnologia em matéria de *hardware* e *software*, houve o crescimento exponencial na produção de dados no contexto digital. Estes dados são fruto das mais diversas fontes, tais como: *web*, redes sociais, internet das coisas (IoT), banco de dados corporativos e públicos, dados biométricos, dentre outros.

[3] Google Ngram Viewer: https://books.google.com/ngrams/

Os dados, de acordo com Castro e Ferrari (2016, p. 4), são símbolos ou signos não estruturados, sem significado, como valores em uma tabela, o qual a informação está contida nas descrições, agregando significado e utilidade aos dados.

Neste sentido, a velocidade de criação nem sempre se traduz em qualidade. E para caracterizar este cenário surgiu uma máxima denominada “GIGO”, uma sigla em inglês cuja tradução em sentido literal é “lixo colocado para dentro, lixo colocado para fora”. Esta expressão de acordo com Taulli (2020, p. 53) é atribuída ao técnico da IBM George Fuechsel, o qual afirma que sistemas alimentados com dados de qualidade resultam em boas saídas, contudo, sistemas alimentados com dados ruins resultam em péssimas saídas. Desta forma, é necessário compreender as particularidades e características dos dados, elemento basilar de projetos de AI/ML.

Em relação à organização dos dados, Taulli (2020, p. 40-41) os classifica em três tipos. O primeiro deles são os dados estruturados. Eles se caracterizam por serem rotulados e geralmente armazenados em bancos de dados relacionais ou em planilhas. Ademais, são de fácil armazenamento, acesso e análise. Todavia, eles representam apenas 20% de um projeto de IA (TAULLI, 2020, p. 40). Como exemplo temos os conjuntos de metadados utilizados para a descrição de um artigo de revista em uma base de dados referencial, que incluem dados com informações sobre autor, título, fonte de publicação, assuntos dentre outros.

Já os dados semiestruturados, apresentam-se como dados híbridos, ou seja, parcialmente estruturados, o qual possuem algumas marcas internas que ajudam na categorização, o que torna a sua estrutura pouco rígida. Como exemplo temos as *tags* atribuídas em vídeos por usuários de *sites* como o Youtube. Os dados semiestruturados, representam apenas cerca de 5% a 10% de todos os dados de um projeto segundo Taulli (2020, p. 41).

Por fim, os dados não estruturados representam informações sem uma organização predefinida. Normalmente eles se referem a imagens, vídeos, sons, páginas *web*, dentre outros. Caracterizam-se por serem de difícil indexação, acesso e análise (CASTRO; FERRARI, 2016, p. 29). Além disso, essa categoria representa normalmente a maior parte dos conjuntos de dados em projetos.

Esta variedade de categorias de dados, produzidos e armazenados diariamente, fomentou um novo cenário denominado de *Big Data*. O termo *Big Data* de acordo com Taurion (2013) refere-se ao conjunto de dados extremamente grandes gerados a partir de práticas tecnológicas, tais como mídia social, tecnologias operacionais, acessos à internet e fontes de informações distribuídas. Taurion (2013) simplifica este cenário por meio da fórmula:

> ***Big Data* = volume + variedade + velocidade + veracidade, tudo agregando + valor.**

Ao apresentar essa fórmula, Taurion (2013) destacou algumas das características do *Big Data*, a saber: volume, variedade, velocidade e veracidade dos dados. Todavia, originalmente o *Big Data* é formado por três características essenciais denominado por Laney (2001) como 3Vs.

O primeiro V refere-se ao *Volume* que trata-se da ascensão ilimitada de dados no ambiente digital, que em sua maioria são formados por dados não estruturados. Já o segundo V, é a *Velocidade* resultado da rapidez com que esses dados estão sendo criados. Por fim, o terceiro V é atribuído à *Variedade*, que representa a grande diversidade de dados, provenientes de distintas fontes e formatos digitais. No entanto, o rol de características conferidas ao *Big Data* cresceu, e desta forma, foram adicionados outros Vs de acordo com a visão de especialistas do domínio, dentre os quais, se destacam a veracidade, o valor e a variabilidade de acordo com Akhtar (2018), dado ao dinamismo e complexidade deste novo cenário.

Desta forma, o *Big Data* requer uma infraestrutura robusta, o que se traduz em formas inovadoras de processamento de grandes conjuntos de dados heterogêneos. E, para isso, faz-se necessário a aplicação de técnicas, que coletem, processem, analisem e interpretem os dados de maneira ágil e inteligente.

Por fim, Taulli (2020, p. 60) afirma que ser bem-sucedido sob a perspectiva da IA significa ter uma cultura orientada por dados, logo ao tomar decisões, é

imprescindível que as instituições, sejam públicas ou particulares, analisem sobretudo os seus conjuntos de dados.

À vista deste cenário, a próxima seção apresentará os fundamentos conceituais do aprendizado de máquina (AM), como uma técnica capaz de auxiliar os gestores a identificar padrões, prever comportamentos e, assim compreender melhor as necessidades presentes e futuras de seus clientes, aqui tratados como usuários.

2.2 *Machine learning*: 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).

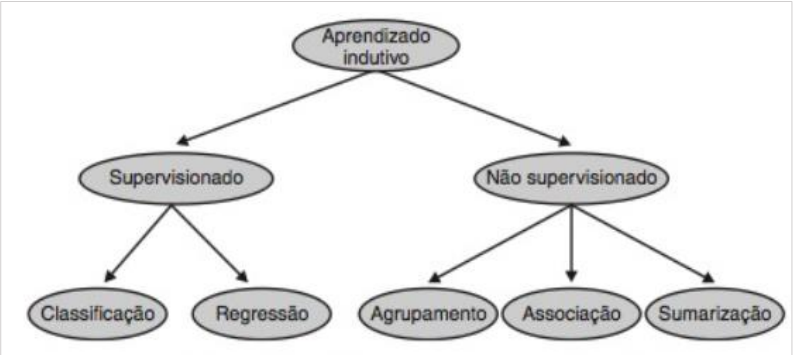
Já em 1997, o cientista e professor americano Tom M. Mitchell apresentou uma definição da área que se tornou mais popular, o qual define AM como “a área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio da experiência” (CASTRO; FERRARI, 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 Castro 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 Castro 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 colocar 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 as Bibliotecas.

2.3 A inteligência artificial aplicada a bibliotecas: análise do cenário

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, a fim de 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 (CORDELL, 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[4] (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. Foi investigado os impactos da IA e do ML no patrimônio cultural, o qual 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[5] em AI. A pesquisa concluiu 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[6], foram recuperados 8 registros com a expressão "*Artificial*

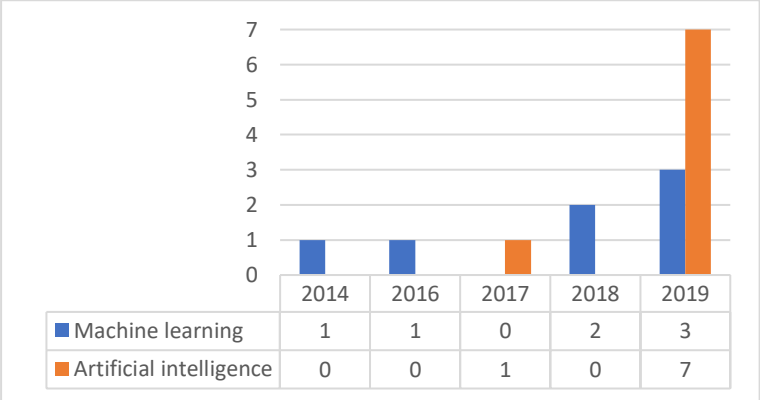
[4] A pesquisa incluiu instituições brasileiras, todavia a maior parcela de participantes se encontra na Europa.

[5] Os projetos relatados na pesquisa abrangem uma variedade de objetivos e tipos de mídia, com a maioria dos projetos voltados para digitalização e descoberta. Além disso, os principais desafios relatados estão relacionados às habilidades e conhecimentos exigidos da equipe interna e a disponibilidade de dados de treinamento devidamente anotados de acordo com a Europeana (2021).

[6] Repositório dos trabalhos produzidos em congressos da IFLA: http://library.ifla.org/cgi/search/advanced. A pesquisa foi realizada em agosto de 2021.

intelligence" e 7 registros com a expressão "*machine learning*", conforme apresentado no gráfico abaixo.

Gráfico 1: Documentos recuperados no repositório da IFLA



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*[8] (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.

[7] Chronicling America: chroniclingamerica.loc.gov
[8] By the People: crowd.loc.gov

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 Canadá, foi desenvolvido um projeto de alfabetização algorítmica para usuários de bibliotecas públicas, denominado *AI for ALL*[9]. Uma parceria entre a Biblioteca da Universidade Ryerson, a Biblioteca Pública de Toronto e a Federação Canadense de Associações de Bibliotecas (CFLA-FCAB). O objetivo é fornece uma variedade de abordagens pedagógicas para compreender os principais aspectos da IA (algoritmos), de modo a capacitar os usuários.

No âmbito do Reino Unido, a British Library é uma das principais colaboradas do projeto "*Living with machine*[10]", 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)[11] 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.

[9] AI for ALL: https://aiforall.ca/
[10] Living with machines: https://livingwithmachines.ac.uk/
[11] BnF and Artificial Intelligence: https://www.bnf.fr/en/feuille-de-route-ia

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 e 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), defendem 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 de modo a destacar 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.

Deste modo, na próxima seção serão abordados os princípios metodológicos utilizados para a construção da Revisão Sistemática de Literatura (RSL), de maneira a auxiliar na análise e investigação das aplicações do aprendizado de máquina em produtos e serviços de bibliotecas e, apresentá-las à comunidade acadêmica e ao público interessado pelo tema.

3 METODOLOGIA DE PESQUISA

Este estudo se pautará em uma pesquisa de caráter qualitativo e quantitativo, com a abordagem exploratória, de natureza pura, por meio do uso da pesquisa bibliográfica, a partir da elaboração de uma Revisão Sistemática de Literatura (RSL), para coleta, construção e análise de um *corpus* documental sobre aprendizado de máquina (AM) em bibliotecas, que será explorado à luz da metodologia de análise de conteúdo.

Quanto ao caráter qualitativo, a pesquisa buscará identificar e analisar as funcionalidades criadas, as tendências e soluções implementadas a partir da utilização AM em produtos e serviços ofertados em bibliotecas, bem como as competências dos profissionais da informação e os aspectos éticos envolvidos neste processo de transformação e inovação tecnológica. Já com relação aos seus aspectos quantitativos, será realizado uma análise estatística dos dados obtidos por meio da pesquisa bibliográfica, com o propósito de obter as seguintes informações, a saber: autoria, o que inclui o país de origem e a filiação; o título, o ano, a fonte e o local de publicação; o idioma; as palavras-chave; a aplicação (teórica ou prática); os produtos e/ou serviços analisados; a finalidade; as técnicas empregadas; o tipo de biblioteca e a área de aplicação na biblioteca e, por fim quais os critérios de inclusão e exclusão utilizados para a inclusão no *corpus* pesquisado.

No que se refere ao emprego de procedimentos exploratórios na pesquisa, haverá a reunião, a consolidação e a análise dos dados e informações acerca do *corpus* coletado, de modo a apresentar o estado da arte da temática na última década, bem como com indicar pesquisar futuras. É importante ressalta que segundo Braga (2007, p. 25) este tipo pesquisa não tem como objetivo testar uma hipótese, mas de buscar padrões. Além disso, na visão de Gil (2019, p. 26), a investigação exploratória busca proporcionar um olhar geral, do tipo aproximativo, acerca de determinado fato.

Em relação a natureza pura da pesquisa, ela tem por objetivo desenvolver um conjunto de conhecimentos, sem obrigatoriamente ter uma aplicação prática.

A respeito da utilização da pesquisa bibliográfica, é devido a pretensão de reunir e analisar os documentos publicados sobre a temática, com a finalidade

de colocar os pesquisadores e profissionais interessados em contato direto com a literatura científica produzida na última década.

E para a construção desta pesquisa bibliográfica foi utilizado o método de Revisão Sistemática da Literatura (RSL) baseada nas diretrizes propostas por Galvão e Ricarte (2020) para o campo da Ciência da Informação, complementados pelos estudos produzidos por Kitchenham (2004) e Felizardo et al (2017) para o campo da Ciência da Computação, com o objetivo de mapear as pesquisas acerca da aplicação da AM em bibliotecas.

A motivação para a produção desta RSL foi reunir pesquisas dispersas que tratam da incorporação de novas tecnologias baseadas em AM em bibliotecas, de modo a apresentar benefícios e limitações na incorporação destas novas tecnologias, bem como descrevê-las e identificá-las, de forma a contribuir para novos estudos nesta área de pesquisa.

Neste trabalho a Revisão Sistemática de Literatura é definida como:

> uma modalidade de pesquisa, que segue protocolos específicos, e que busca entender e dar alguma logicidade a um grande *corpus* documental, especificamente, verificando o que funciona e o que não funciona num dado contexto (GALVÃO; RICARTE, 2020, p. 58)

Estes estudos buscam identificar fontes documentais com uma temática pré-definida, interpretar e sintetizar e, posteriormente avaliar e analisar criticamente os dados extraídos.

À vista disso, a RSL surgiu como uma vertente mais rigorosa na condução de pesquisas. Ela se difere da revisão tradicional como um método que busca diminuir o viés em todas as etapas de produção da pesquisa, além de seguir uma metodologia mais rigorosa nos processos de planejamento e execução da pesquisa. Ademais, de acordo com Kitchenham (2004, p. 2) o que diferencia as RSL das revisões de literatura convencionais é:

> a definição de um protocolo de revisão por parte das RSL, o qual especifica a questão a ser abordada e os métodos que serão usados para realizar a revisão; A definição e a documentação das estratégias de pesquisa com o objetivo de detectar como o máximo possível da literatura relevante com rigor e integridade; E, a definição de critérios de avaliação e de qualidade explícitos para a inclusão e a exclusão de estudos primários.

As RSL são classificadas como fontes de informação secundárias, baseadas em métodos empíricos, que desempenham um importante papel na

comunicação cientifica, pois contribuem para o desenvolvimento de novas teorias, bem como na identificação de campos de estudos emergentes com alto grau de evidência e controle. Além disso, a sua natureza é iterativa, ou seja, a técnica de RSL é aplicada em todo o ciclo de vida da pesquisa, por meio de princípios, tais como a visibilidade, o rigor, a clareza, a reprodutibilidade e a auditabilidade em todo o seu processo de elaboração.

E, por fim será realizado a análise de conteúdo do *corpus* selecionado, à luz da metodologia proposta por Bardin (2021), que conceitua a análise de conteúdo como um conjunto de técnicas de análise das comunicações (BARDIN, 2021, p. 33), composta por três fases: a pré-análise, a exploração do material e o tratamento dos resultados que incluem processos de inferência e de interpretação (BARDIN, 2021, p. 121). A primeira etapa, de acordo com a autora (2021, p. 121) corresponde ao processo de seleção dos documentos que comporão o *corpus* da pesquisa, a formulação das hipóteses e dos objetivos e, a elaboração dos indicadores que fundamentam a interpretação final, materializados no quadro 2, 3, 4 e 5 desta pesquisa. Em seguida, será realizado a exploração do material reunido, ou seja, a análise propriamente dita que, de acordo com Bardin (2021, p. 127), nada mais é do que a aplicação sistemática das decisões tomadas na etapa anterior. Por fim, a última etapa refere-se ao tratamento dos resultados obtidos e a sua interpretação, os resultados são tratados de maneira significativa e válida, por meio da aplicação de análises estatísticas que auxiliam em processos de inferência e interpretação dos dados (BARDIN, 2021, p. 127).

Ademais, nesta etapa da pesquisa, será apresentado uma análise de categorias produzidas no relatório da Europeana "*AI in relation to GLAMs Task Force*[12]" (2021). O objetivo deste relatório é apresentar uma investigação acerca dos impactos da inteligência artificial e do *machine learning* no setor de patrimônio cultural. À vista disso, serão analisadas a cobertura temática no *corpus* sob a perspectiva de 9 categorias, descritas no relatório, a saber: extração do conhecimento; qualidade dos meta(dados); análise de audiência;

[12] Link para o texto completo do relatório EUROPEANA: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/AI%20in%20relation%20to%20GLAMs%20Task%20Force%20Report.pdf

crowdsourcing e *human in the loop*; visualização de coleções Glam; descoberta e pesquisa; criatividade ou engajamento, projetos e iniciativas e, por fim a tradução automática.

A seguir será apresentado o processo de planejamento para a elaboração do protocolo de pesquisa para a produção da RSL.

3.1 Planejamento da Revisão sistemática de literatura

Nesta seção serão definidas as diretrizes do protocolo de pesquisa, baseadas na metodologia produzida por Fabbri, Octaviano e Hernandes (2017, p. 18) em obra monográfica organizada por Felizardo et al (2017). A construção do protocolo desta investigação tem por objetivo definir as estratégias e os critérios para coleta, extração e análise do *corpus* documental da Revisão Sistemática de Literatura (RSL).

À vista disso, serão produzidos quadros informativos para cada etapa desta RSL de modo a descrever detalhadamente cada etapa do protocolo.

A primeira etapa do protocolo apresenta a descrição do projeto, uma breve apresentação dos objetivos (a versão detalhada encontra-se na seção 1 deste trabalho, a questão principal de pesquisa, o conjunto de critérios PICO (População, Intervenção, Controle e Resultados), bem como as possíveis aplicações da pesquisa, descritas a seguir.

Quadro 2. Protocolo de pesquisa: informações gerais

INFORMAÇÕES GERAIS	
Descrição	Revisão Sistemática de Literatura (RSL) desenvolvida como requisito parcial para a produção da dissertação de mestrado do Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília, cujo objetivo consiste na identificação das aplicações do aprendizado de máquina em produtos e serviços ofertados em bibliotecas, bem como os aspectos profissionais e éticos, sob o ponto de vista do profissional da informação, a partir do uso de tecnologias em bibliotecas.
Objetivos	Mapear benefícios e impactos em termos metodológicos e operacionais que o aprendizado de máquina pode oferecer para o desenvolvimento de produtos e serviços de informação.
Questão principal	Como o desenvolvimento de produtos e serviços por meio do aprendizado de máquina podem impactar e beneficiar as bibliotecas?
População	Publicações que descrevam produtos e serviços em bibliotecas e que utilizem o aprendizado de máquina em seu desenvolvimento, ou que aborde aspectos éticos ou laborais gerados a partir da aplicação do ML em bibliotecas.
Intervenção	Análise da literatura que cite produtos e serviço que utilizem o aprendizado de máquina em seu desenvolvimento, para executar tarefas especificas em bibliotecas.
Controle	Análise exploratória e descritiva acerca da temática em questão, considerando fontes primárias tais como relatórios técnicos, trabalhos apresentados em congressos, livros e capítulos de livros, tese e dissertações e artigos publicados em meio acadêmico e profissional.
Resultados	Pretende-se como resultado de pesquisa a reunião, análise e interpretação de dados e informações sobre a criação e o desenvolvimento de produtos e serviços em bibliotecas com o uso do aprendizado de máquina, com vistas a apresentar as tendências e soluções produzidas na literatura da Ciência da Informação sobre o assunto na última década. Além disso, será produzido um conjunto de recomendações para a aplicação de técnicas de aprendizado de máquina para bibliotecas.
Aplicação	Execução de novos estudos sobre o tema e documento de apoio à tomada de decisão para profissionais da informação na produção de novos produtos e serviços em unidades de informação.

Fonte: Elaborado pela autora (2022).

Ainda em relação ao quadro 2, o conjunto de critérios PICO, foram estruturados de acordo com a metodologia proposta por Fabbri (2017, p. 20). O qual a população refere-se a área de aplicação, a intervenção apresenta o que será investigado, o controle define o método e o procedimento utilizado e, por fim os resultados correspondem ao tipo de resultado relacionado aos possíveis efeitos da aplicação da RSL.

A seguir serão apresentados os critérios para identificação dos estudos, que envolve a definição do idioma, do corte temporal, das palavras-chave e a definição das fontes de pesquisa, delimitados a seguir:

Quadro 3. Protocolo de pesquisa: identificação de estudos

IDENTIFICAÇÃO DE ESTUDOS	
Idioma	Português; Inglês; Espanhol.
Corte temporal	Documentos publicados na última década, entre os anos de dezembro de 2010 a janeiro de 2021.
Palavras-chave	Aprendizado de máquina; *Machine learning*; *Aprendizaje automático*; Biblioteca; *Library*.
String de busca	"Aprendizado de máquina" AND Biblioteca "*machine learning*" AND *library* "*aprendizaje automático*" AND Biblioteca
Critério de seleção dos documentos	Documentos publicados no campo da Ciência da Informação, sem preferências por suporte ou formato.
Critérios de seleção das fontes de busca	Bases de dados do campo temático da Ciência da Informação disponíveis para consulta pública ou acessíveis por meio do Portal de Periódicos da CAPES e da Biblioteca Central da Universidade de Brasília.
Lista das fontes de busca	BDTD (Biblioteca Digital de Teses e Dissertações); BRAPCI (Base de Dados Referencial de Artigos em Ciência da Informação); ISTA (I*nformation Science and Technology Abstracts*); LISTA *(Library, Information Science & Technology Abstracts)* e WoS *(Web of Science).*

Fonte: Elaborado pela autora (2022).

Para o desenvolvimento dos critérios de identificação de estudos, a seleção das palavras-chave baseou-se em uma consulta ao Tesauro brasileiro de Ciência da Informação (PINHEIRO; FERREZ, 2014) e no Tesauro da base de dados LISTA. Estas consultas tiveram por objetivo selecionar termos e expressões com maior precisão e especificidade. Ademais, não houve restrições quanto à cobertura geográfica. Um outro ponto importante a ser destacado é a ausência da base LISA (*Library and Information Science Abstracts*) do rol das bases elencadas para a pesquisa. Esta ausência se deve a indisponibilidade da base para consulta tanto no Portal de Periódicos da Capes, quanto no diretório de bases de dados da Biblioteca da Universidade de Brasília no período da coleta de dados em fevereiro de 2021.

No que se refere aos critérios de seleção de documentos (quadro 4), eles foram construídos com base na metodologia proposta por Felizardo et al (2007, p. 54) segundo os objetivos desta pesquisa, relacionados no quadro 2.

Quadro 4. Protocolo de pesquisa: critérios de seleção de documentos

CRITÉRIOS DE SELEÇÃO DE DOCUMENTOS
INCLUSÃO
(I-1) Documentos que abordem as aplicações, os produtos e serviços, benefícios e impactos, bem como as dificuldades para a implantação da AM em bibliotecas;
(I-2) Documentos que tratem acerca das competências do profissional da informação em bibliotecas frente ao uso de novas tecnologias com foco na AM;
(I-3) Documentos que reflitam acerca das questões éticas geradas a partir do uso de novas tecnologias em bibliotecas com foco em AM.
EXCLUSÃO
(E-1) Documentos que não foram publicados na última década;
(E-2) Documentos que não abordem as aplicações, os produtos e serviços, impactos, benefícios e dificuldades para a implantação da AM em bibliotecas;
(E-3) Documentos que não tratem acerca das competências do profissional da informação frente ao uso de novas tecnologias com foco na AM;
(E-4) Documentos que não reflitam acerca das questões éticas geradas a partir do uso de novas tecnologias em bibliotecas com foco em AM;
(E-5) Documentos que não estejam no domínio da Ciência da Informação;
(E-6) Documentos que não estejam publicados nos idiomas: Português, Inglês e ou Espanhol;
(E-7) Documentos que não estejam disponíveis para consulta em texto integral no formato *online* ou físico;

(E-8) Documentos pré-textuais e/ou pós-textuais; (E-9) Documentos duplicados; (E-10) Documentos plagiados.

Fonte: Elaborado pela autora (2022).

Os critérios de inclusão e exclusão foram construídos com base nos objetivos da pesquisa, elencados na seção 1 e no quadro 2 deste trabalho. Ademais, no que se refere à inclusão de documentos na RSL, a publicação deverá atender a pelo menos 1 dos critérios de inclusão estabelecidos no quadro 3. Quanto à aplicação do critério de exclusão é importante destacar que os critérios E-2, E-3 e E-4, que tratam do assunto central desta pesquisa, são os principais critérios para a exclusão de documentos, pois se o documento aplicar simultaneamente os três critérios, a publicação não integrará o *corpus* da pesquisa. Já em relação aos critérios E-1, E-5, E-6, E-7, E-8, E-9 e E-10, trata-se de critérios secundários de exclusão de documentos, cuja aplicação é posterior ao resultado de aplicação dos critérios E-2, E-3 e E-4.

Com relação as estratégias para a seleção e avaliação de qualidade dos estudos, estas informações foram reunidas no quadro 5, apresentada a seguir:

Quadro 5. Protocolo de pesquisa: seleção e avaliação de estudos

SELEÇÃO E AVALIAÇÃO DE ESTUDOS	
Estratégia para seleção dos estudos	Aplicação das *strings* de busca com o auxílio dos filtros de pesquisa em cada base de dados. Após a recuperação dos documentos, procedeu-se a seleção das publicações por meio da leitura dos elementos pré-textuais e pós-textuais de cada documentos para a aplicação dos critérios de inclusão e exclusão.
Avaliação da qualidade dos estudos	Os critérios de qualidade das publicações, estão atrelados às diretrizes estabelecidos em cada base de dados, o que inclui o conjunto de critérios de avaliação produzidos pela CAPES, ora denominado QUALIS. Ademais, cada documento será avaliado por meio da aplicação de suas metodologias.

Fonte: Elaborado pela autora (2022).

Nesta etapa será observada a coerência dos estudos com os resultados alcançados, de modo a aferir a qualidade individual de cada estudo de compõe a RSL.

Por fim, os últimos dados apresentados, são relacionadas as estratégias de extração e sumarização dos dados.

Quadro 6. Protocolo de pesquisa: síntese dos dados e apresentação dos resultados

SÍNTESE DOS DADOS E APRESENTAÇÃO DOS RESULTADOS	
Estratégia de extração de dados	Serão extraídas as informações referentes aos autores, o que inclui o país de origem e a filiação; o título, o ano, a fonte e o local de publicação; o idioma; as palavras-chave; a aplicação (teórica ou prática); os produtos ou serviços analisados; a finalidade; as técnicas empregadas; o tipo de biblioteca e a sua área de aplicação. Além disso, serão extraídos também as informações sobre os aspectos éticos e profissionais elencadas nos documentos.
Estratégia de sumarização dos dados	Todas as informações serão condensadas em fichamentos e os dados extraídos serão analisados em planilhas do Excel.

Fonte: Elaborado pela autora (2022).

Logo após a execução das quatro primeiras etapas do protocolo, os dados serão selecionados, recuperados e organizados para a última fase que corresponde a sumarização e análise quantitativa e qualitativa do *corpus* documental.

Para esta etapa, será utilizada a análise de conteúdo, com o objetivo de produzir sentidos e significados em meio a uma diversidade de documentos que formam o *corpus* desta pesquisa, conforme destacado na seção anterior. Ademais, o *corpus* será analisado com base nas categorias elencadas no relatório da Europeana (2021) sobre as tendencias da IA em instituições do patrimônio cultural.

A próxima seção apresentará a análise e a discussão os dados coletados após a execução do protocolo de pesquisa.

4 REVISÃO SISTEMÁTICA DE LITERATURA (RSL)

Este capítulo apresentará uma análise geral dos dados de pesquisa recuperados após a execução do protocolo da revisão sistemática de literatura (RSL), os resumos informativos das publicações que compõem o *corpus* deste estudo e, por fim, uma análise de conteúdo de maneira a investigar o objeto de pesquisa e identificar as aplicações, as competências laborais e éticas do aprendizado de máquina em bibliotecas.

4.1 Execução e apresentação dos dados que compõem a RSL

Após uma seleção preliminar dos dados, realizada entre os dias 15 e 28 de fevereiro de 2021 nas 5 fontes de pesquisas, a saber: Base de Dados em Ciência da Informação (BRAPCI)[13], Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[14], *Information Science & Technology* (ISTA)[15], *Library & Information Science & Technology Abstracts* (LISTA)[16] e *Web of Science* (WoS)[17], acessadas através do Portal de Periódicos da CAPES, foram recuperados um total de 408 documentos, o qual 77 documentos foram pré-selecionados, a partir da análise do título, resumo e conjunto de palavras-chave, para uma análise mais aprofundada.

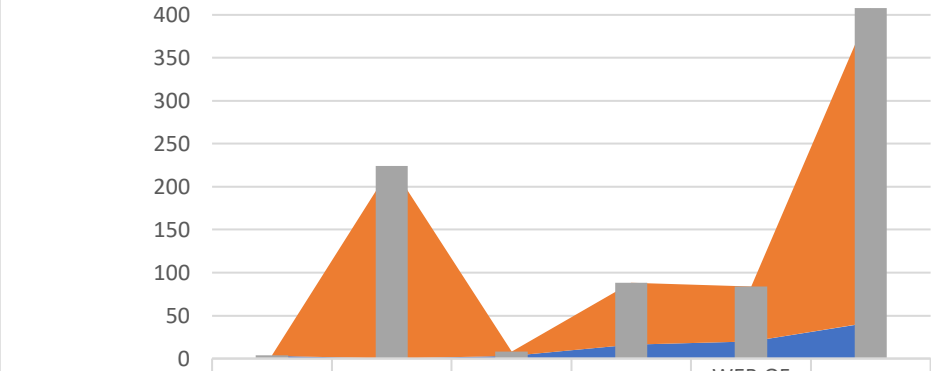
O *corpus* formado pelos 77 documentos foi analisado individualmente a partir de seus elementos pré e pós-textuais, com a aplicação dos critérios de inclusão e exclusão estabelecidos no quadro 4. E, deste modo, foram excluídas 35 publicações que não atenderam a pelo menos dos critérios de inclusão, ou que apresentaram uma análise superficial do tema, e desta forma não refletiam o escopo central da pesquisa.

[13] BRAPCI: https://brapci.inf.br/
[14] BDTD: https://bdtd.ibict.br/vufind/
[15] ISTA: https://www.ebsco.com/products/research-databases/information-science-technology-abstracts
[16] LISTA: https://www.ebsco.com/products/research-databases/library-information-science-and-technology-abstracts
[17] WoS: https://www.webofscience.com/wos/woscc/basic-search

Por fim, o *corpus* final que compõe esta RSL é constituído de 42 documentos aprovados e 366 documentos rejeitados. Ademais, o conjunto de documentos aprovados é composto por artigos, capítulos de livro e comunicações científicas fruto de eventos científicos.

Em relação aos documentos aprovados que compõem o *corpus* por base de dados, foram selecionados e aprovados: 3 documentos da Brapci, 3 documentos da ISTA, 16 documentos da LISTA e 20 documentos da WoS, perfazendo assim um total de 42 documentos, conforme a representação gráfica abaixo, que resume a fase preliminar de seleção e recuperação dos documentos.

Gráfico 2. Relação entre os documentos recuperados, os documentos aprovados e os documentos rejeitados.



	Brapci	BTDT	ISTA	LISTA	WEB OF SCIENCE	TOTAL
REJEITADOS	1	224	5	72	64	366
APROVADOS	3	0	3	16	20	42
TOTAL GERAL	4	224	8	88	84	408

Fonte: Elaborado pela autora (2022).

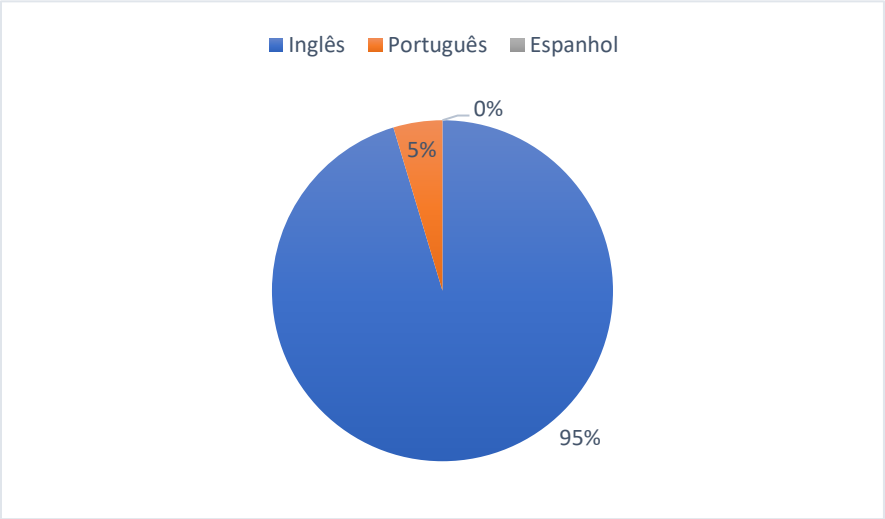
É importante ressaltar que na relação entre documentos aprovados e documentos recuperados por bases de dados, se destacou a alta revocação da base BDTD. Todavia, essa alta revocação de registros não se traduziu em precisão e, desta forma a BDTD obteve o maior número de documentos rejeitados para compor este estudo. Este fato pode ser atribuído a uma falsa recuperação de documentos, tendo em vista que o termo especificador “biblioteca” é polissêmico e, por tanto aplicável a diversas áreas do conhecimento. Ademais, infere-se ainda a partir do conjunto de registos

recuperados que o tema “aprendizado de máquina/*machine learning*” é objeto de estudo e aplicação vigente não só na Ciência da Computação (sua área matriz), como em diferentes áreas do conhecimento tais a Linguística, a Ciência Política, a Engenharia e a Odontologia dentre outros, conforme os resultados da pesquisa realizada na BDTD na última década no Brasil.

Ainda em relação aos dados recuperados na base BDTD, por se tratar de um serviço que reúne em um único portal de busca a produção científica fruto da pós-graduação brasileira (monografias de especializações, dissertações e teses), a investigação nesta base constatou que não havia publicações sobre o aprendizado de máquina no âmbito das bibliotecas entre os anos de 2010 e 2021, o que confirmou o ineditismo deste trabalho.

No que se refere aos documentos aprovados por idiomas, a análise do gráfico 3 evidência o domínio da língua inglesa em publicações científicas sobre aprendizado de máquina, o que demonstra o esforço dos pesquisadores pelo compartilhamento de pesquisas em diferentes culturas por meio de um idioma único, de modo a promover o aumento da visibilidade de suas investigações.

Gráfico 3. Idioma das publicações aprovadas para compor a RSL



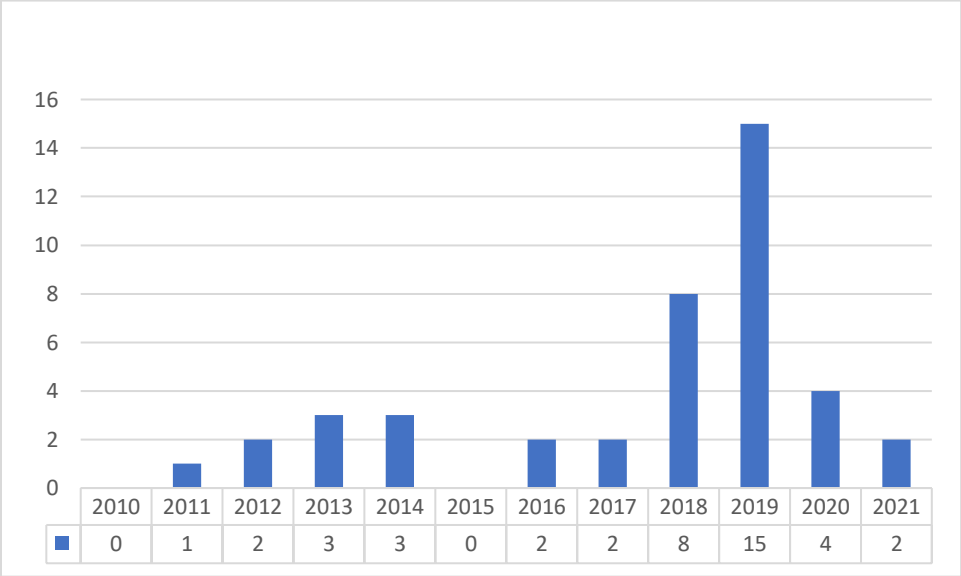
Fonte: Elaborado pela autora (2022).

Por outro lado, ao mesmo tempo que a língua inglesa facilita as comunicações entre a comunidade científica, ela pode se tornar uma barreira e

limitar o acesso ao conhecimento científico por parte da população no geral. Desta forma, conclui-se que no *corpus* desta RSL não há diversidade linguística sobre o tema, haja vista a ausência de publicações em espanhol e o baixo número de publicações em português sobre o tema (apenas dois). Todavia, observa-se, de acordo com a figura 3, que será analisada mais adiante, que apesar do baixo número de publicações em idiomas ditos "periféricos", existe um número expressivo de pesquisadores brasileiros com publicações em inglês sobre o tema em questão.

No que tange à quantidade de documentos publicados na última década, entre os anos de dezembro de 2010 a janeiro de 2021, o gráfico 4 indica uma concentração de publicações entre os anos de 2018 e 2019, e logo em seguida uma queda, conforme a imagem abaixo:

Gráfico 4. Quantidade de publicações na última década



Fonte: Elaborado pela autora (2022).

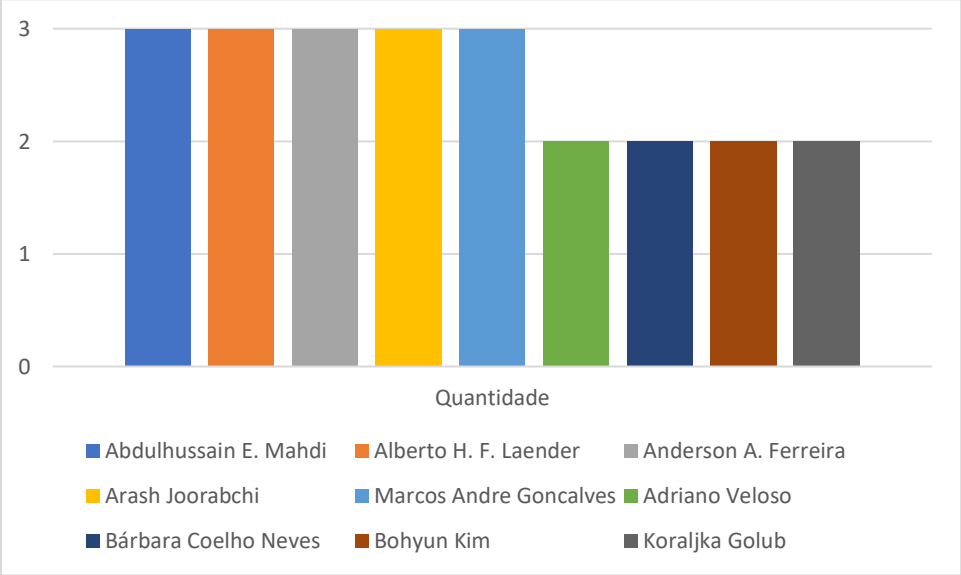
No ano de 2020 houve um recuo que pode ser atribuído ao período de pandemia[18] que afetou a produção científica não relacionada ao tema Covid-19.

[18] De acordo com a Organização Mundial da Saúde (OMS), a Covid-19 é a doença causada pelo novo coronavírus conhecido como SARS-CoV-2. A OMS soube da existência deste novo vírus em 31 de dezembro de 2019, quando foi informada de um grupo de casos de "pneumonia viral" declarados em Wuhan (República Popular da China).

Além deste contexto histórico, outro motivo para a queda justifica-se pelo fato de que as pesquisas aprovadas no ano de 2020 não terem sido finalizadas e publicadas nos repositórios institucionais das universidades durante a sua execução.

Já em relação aos dados de autoria, destacaram-se cinco autores com 3 publicações respectivamente, a saber: Marcos André Gonçalves, Arash Joorabchi, Anderson A. Ferreira, Alberto H. F. Laender e Abdulhussain E. Mahdii. É importante ressaltar que os pesquisadores brasileiros Anderson Ferreira, Marcos Gonçalves e Alberto Laender compartilharam a responsabilidade em 3 publicações, assim como os pesquisadores irlandeses Arash Joorabchi e Abdulhussain E. Mahdii que também compartilharam a autoria em 3 publicações, conforme o gráfico abaixo:

Gráfico 5. Quantidade publicações por pesquisador

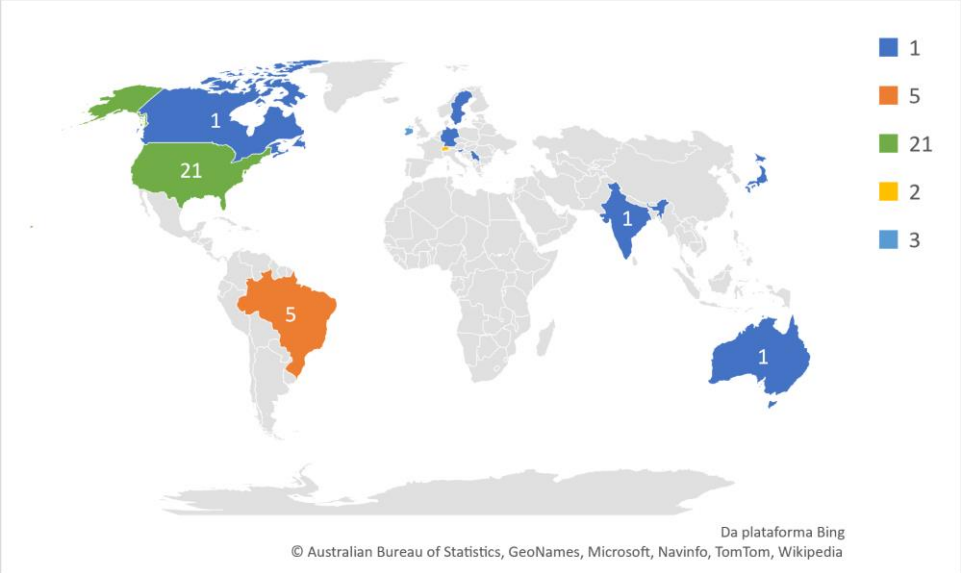


Fonte: Elaborado pela autora (2022).

Logo em seguida, com duas publicações aprovadas na RSL, os pesquisadores brasileiros Bárbara Neves e Adriano Veloso, o último publicou em coautoria com Anderson Ferreira, Marcos Gonçalves e Alberto Laender. Completam esta lista as pesquisadoras Koraljka Golub da Suíça e Bohyun Kim sul-coreana radicada nos Estados Unidos com 2 publicações cada.

No que se refere a quantidade de publicações segundo o país de origem dos pesquisadores, observa-se um domínio dos Estados Unidos com 21 publicações de pesquisadores filiados a instituições americanas na última década. Todavia, é importante ressaltar que para a coleta das informações referente à origem geográfica de cada pesquisador, foram utilizados os dados fornecidos em cada base de dados, descritos de acordo com as informações fornecidas pelos autores no momento da publicação.

Figura 4. Quantidade de publicações segundo o país de origem dos pesquisadores



Alemanha: 1	Escócia: 1	Inglaterra: 2	Singapura: 1
Austrália: 1	Eslovênia: 1	Irlanda: 3	Suécia: 1
Brasil: 5	Estados Unidos: 21	Japão: 1	Suíça: 2
Canadá: 1	Índia: 1	Servia: 1	

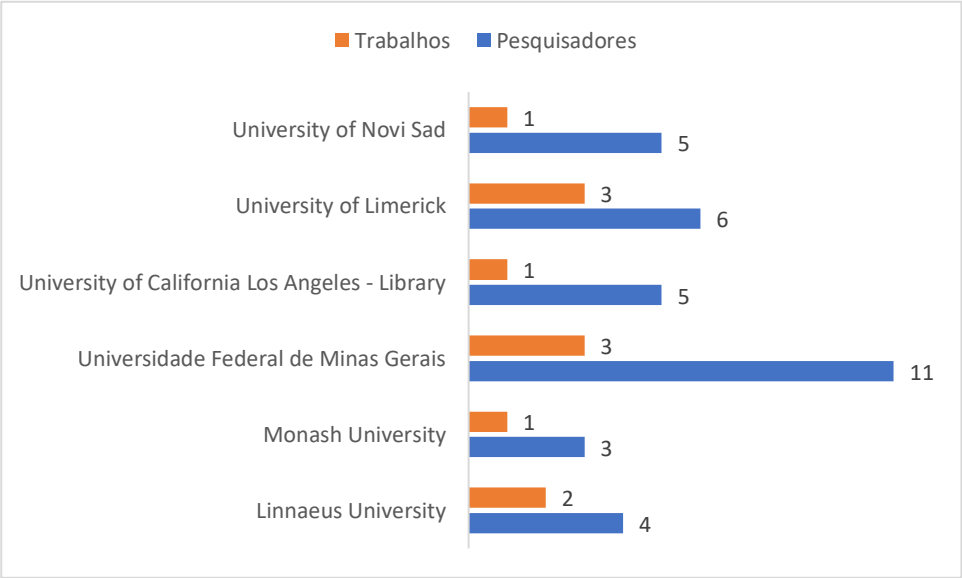
Fonte: Elaborado pela autora (2022).

Ainda em relação a quantidade de publicações segundo o país de origem dos pesquisadores, destacam-se com 5 publicações, pesquisadores de instituições brasileiras e, em terceiro lugar estão os pesquisadores vinculados a instituições da Irlanda com 3 publicações. E, por fim, no rol de países com o maior número de publicações sobre a temática central desta pesquisa, notabilizam-se também os pesquisadores da Inglaterra e da Suíça com 2 publicações cada. Em síntese, constata-se o domínio estadunidense em números absolutos se somados o quantitativo de publicações dos Estados

Unidos (vinte e uma publicações) em comparação ao número de publicação dos outros países em conjunto (vinte e uma publicações).

Quanto ao vínculo institucional destes pesquisadores, verifica-se que as instituições com maior número de pesquisadores sobre o tema, são a Universidade Federal de Minas Gerais, com 11 pesquisadores que produziram um total de 3 artigos. Em seguida, a Universidade de Limerick na Irlanda com seis pesquisadores que publicaram 3 artigos. A Universidade da Califórnia em Los Angeles (UCLA) com 5 pesquisadores que publicaram apenas 1 artigo. A Universidade sueca de Linnaeus com 4 pesquisadores e 2 publicações e, por fim, a Universidade australiana Monash com três pesquisadores e 1 artigo publicado. É importante destacar que o conjunto de pesquisadores que publicaram em coautoria se repetiram em cada publicação, para aquelas instituições com 2 ou mais publicações.

Gráfico 6. Quantidade de publicações segundo a filiação



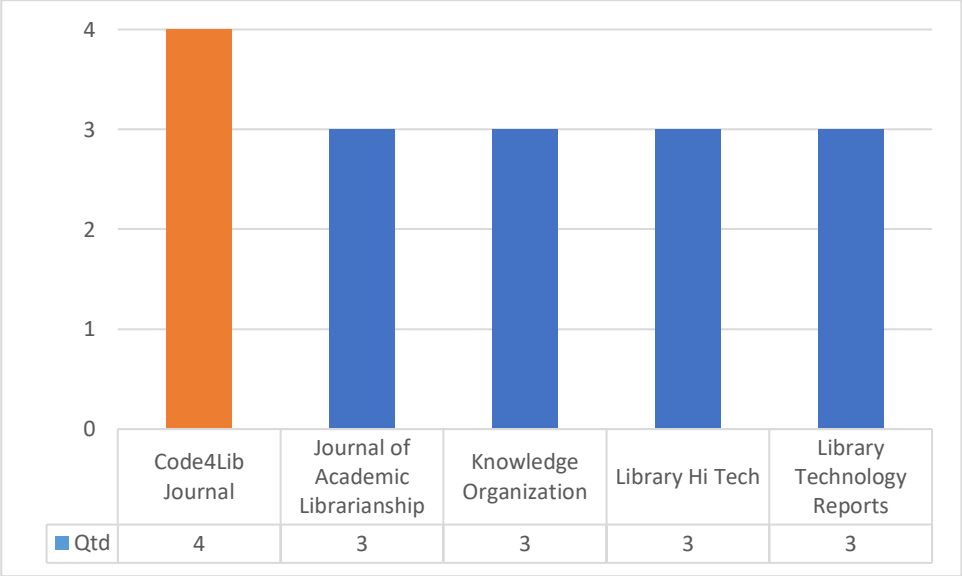
Fonte: Elaborado pelos autora (2022).

Entretanto, se analisarmos as informações deste gráfico (6) em conjunto com a Figura 3 (país de origem), causa uma certa estranheza o baixo número de instituições americanas no rol das principais instituições que mais produziram sobre Aprendizado de Máquina em Bibliotecas, ainda que em números gerais, as instituições originárias dos Estados Unidos detenham o número absoluto de

publicações, segundo a origem de seus pesquisadores. Todavia, este resultado peculiar se deve ao fato de que as pesquisas em território americano se encontram dispersas, diferentemente do que ocorre no Brasil (vide UFMG e UFBA) e na Irlanda (*University of Limerick*), por exemplo, onde as publicações sobre a temática, encontram-se concentradas em poucas instituições.

Já em relação aos periódicos que mais publicaram sobre aprendizado de máquina aplicado às bibliotecas na última década, destacou-se o periódico estadunidense *Code4Lib Journal*[19], uma revista de acesso aberto, produzida por voluntários e voltada para o público interessado em tecnologias da informação e inovações no contexto das bibliotecas, conforme o gráfico a seguir.

Gráfico 7. Quantidade de publicações segundo a fonte de publicação



Fonte: Elaborado pela autora (2021).

Logo em seguida, com 3 publicações cada, temos os periódicos americanos: *Journal of Academic Librarianship*[20], da editora Elsevier que abrange tópicos relacionados às bibliotecas acadêmicas, *Library Hi Tech*[21], cuja responsabilidade pela publicação é do grupo Emerald Insight, o qual o seu

[19] Code4Lib: https://journal.code4lib.org/
[20] Journal of Academic Librarianship: https://www.journals.elsevier.com/the-journal-of-academic-librarianship
[21] Library Hi Tech: https://www.emerald.com/insight/publication/issn/0737-8831

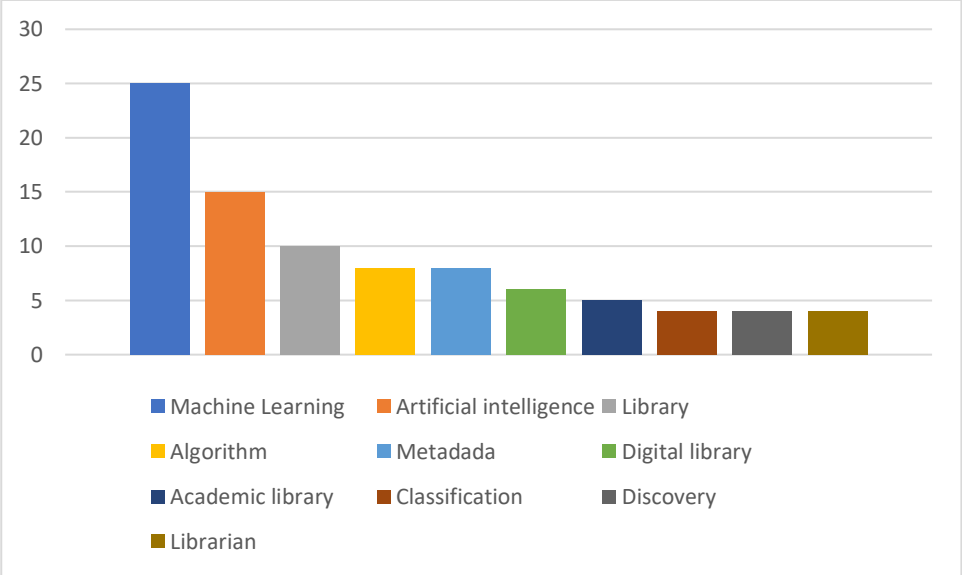
objetivo é publicar artigos, relatórios de conferências e estudos de caso sobre tecnologias em bibliotecas, e por fim o periódico *Library Technology Reports*[22] de responsabilidade da *American Library Association* (ALA), cujo objetivo é auxiliar os bibliotecários na tomada de decisão em produtos e serviços em matéria de tecnologia. Ainda no rol de publicações com expressivas contribuições no domínio do Aprendizado de máquina em bibliotecas, destacou-se o periódico alemão *Knowledge Organization*[23] publicação oficial da *International Society for Knowledge Organization* (ISKO), dedicado a assuntos relativos à representação do conhecimento com 3 contribuições.

Em síntese, os dados relativos ao quantitativo de publicações por fonte, sinalizam para um grande número de publicações periódicas especializadas no segmento de tecnologias aplicadas a bibliotecas de origem estadunidense.

Em relação ao conjunto de palavras-chave mais utilizadas pelos pesquisadores ou pelos catalogadores[24] das bases de dados para descrever os documentos, verifica-se um grande número de publicações representadas pelas expressões que compõem o universo central desta pesquisa, ou seja, "Aprendizado de Máquina" e "Inteligência Artificial". Estas expressões foram as mais utilizadas para descrever os trabalhos na última década, que traduzida em números, indicaram 25 menções atribuídas para a expressão *machine learning*, seguida de 15 menções para *artificial intelligence.* Em suma, as duas expressões revelam um percentual de 73,8% de documentos que contêm uma ou duas expressões atribuídas pelos autores ou pelos catalogadores das bases de dados no rol de palavras-chave das publicações. Todavia, 11 documentos não mencionaram essas expressões, o que representa 26,2% do universo de documentos que compõem o *corpus* desta pesquisa.

[22] Library Technology Reports: https://journals.ala.org/ltr
[23] Knowledge Organization: https://www.isko.org/ko.html
[24] Conforme mencionado, as palavra-chaves utilizadas neste estudo foram extraídas das publicações e das bases de dados. Todavia, o artigo denominado "*What can 100,000 books tell us about the international public library e-lending landscape*?" não apresentou nenhum termo para descrever o documento, e desta forma, foi realizado a indexação com o apoio do tesauro de Ciência da Informação, de modo a cobrir os assuntos do artigo e enriquecer o conjunto de dados analisados.

Gráfico 8. Quantidade palavras-chave atribuídas por publicações



Academic library - 5	Discovery - 4
Algorithm - 8	Librarian - 4
Artificial intelligence - 15	Library - 10
Classification - 4	Machine Learning - 25
Digital library - 6	Metadada - 8

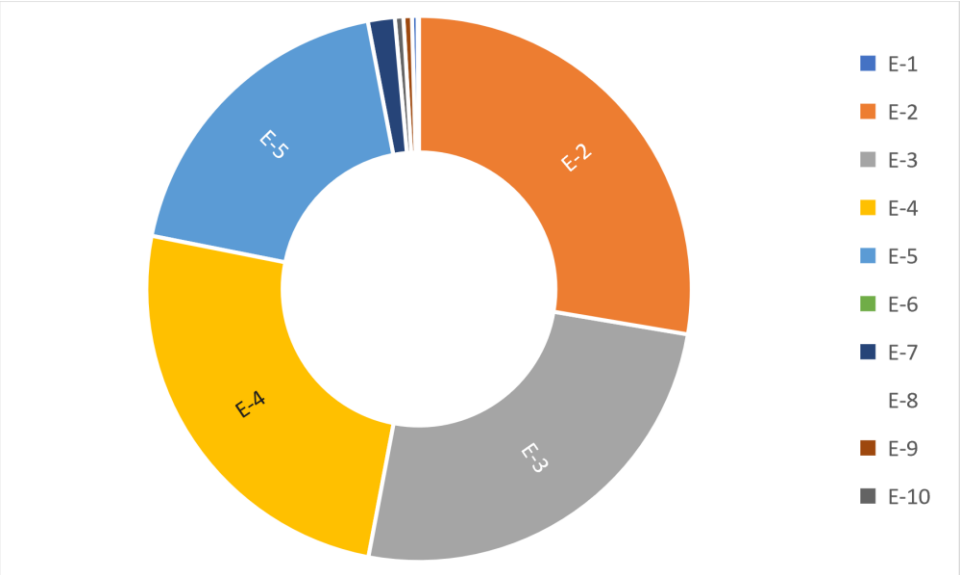
Fonte: Elaborado pela autora (2022).

O terceiro termo mais mencionado foi *library* que obteve 10 menções e, que especificou o ambiente de estudo e aplicação das técnicas em *machine learning*. Ademais, destacam-se ainda os termos *algorithm* e *metadata*, com 8 menções, *digital library* com 6 menções, *academic library* com 5 menções, e por fim, *classification*, *discovery* e *librarian* com 4 menções cada no *corpus* da RSL, entre os descritores mais citados nos trabalhos. Em suma, observa-se uma tendência das pesquisas em IA e ML serem aplicadas ao ambiente de bibliotecas digitais, haja vista o número expressivo de atribuições das palavras-chave *metadata* e *digital library*.

No que diz respeito à aplicação de um ou mais critérios de exclusão no processo de seleção do *corpus* inicial de 408 documentos, destacaram-se a aplicação de 4 critérios para a eliminação de documentos, detalhados a seguir: Critério E-2 – com a exclusão de 330 documentos que não abordaram as aplicações, os produtos e serviços, benefícios e impactos e as dificuldades para a implantação da AM em bibliotecas; Critério E-3 – com a exclusão de 302 documentos que não trataram acerca das competências do profissional da informação frente ao uso de novas tecnologias com foco na AM; Critério E-4 –

com a exclusão de 300 documentos que não refletiram acerca das questões éticas geradas a partir do uso de novas tecnologias em bibliotecas com foco em AM e, critério E-5 – com a exclusão de 225 documentos que não estavam no domínio da Ciência da Informação, conforme o gráfico abaixo:

Gráfico 9. Publicações rejeitadas segundo os critérios de exclusão



E-1	E-2	E-3	E-4	E-5	E-6
4	330	302	300	225	1
E-7	E-8	E-9	E-10	E-11	
19	0	6	6	0	

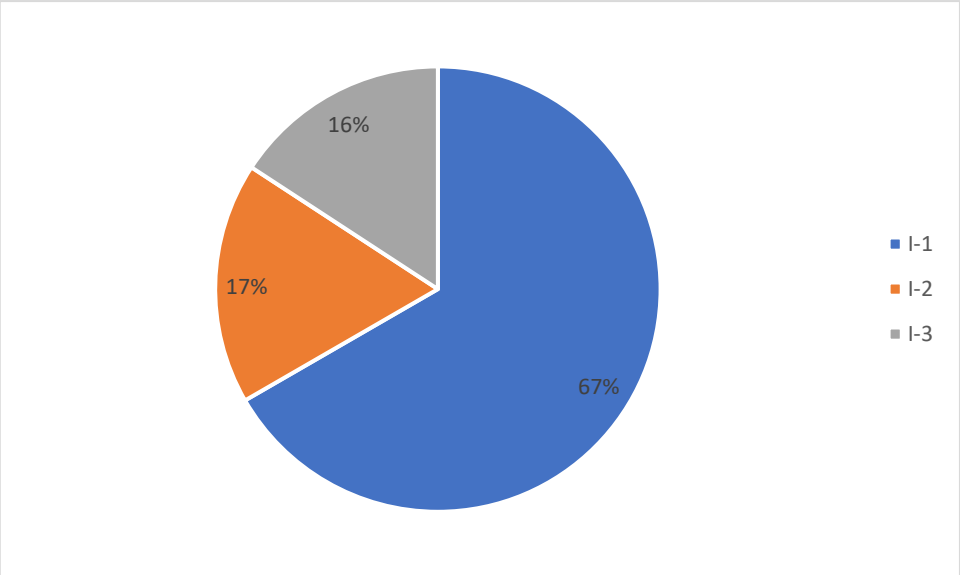
Fonte: Elaborado pela autora (2022).

Além disso, é importante ressaltar que 19 documentos não estavam disponíveis gratuitamente para consulta em texto integral (E-7), 1 documento foi publicado em alemão, fora do escopo idiomático deste trabalho (E-6), mas recuperado em uma base de dados devido a inconsistências em sua descrição e, 6 documentos estavam duplicados no conjunto de base de dados utilizadas para a seleção de documentos (E-10).

Já em relação a aplicação de um ou mais critérios de inclusão no *corpus* de documentos aprovados, 38 documentos foram aprovados segundo o critério I-1, que selecionou publicações que abordam as aplicações, os produtos e serviços, benefícios e impactos e, por fim as dificuldades para a implantação da AM em bibliotecas. Em seguida 10 publicações foram aprovadas segundo o

critério I-2, que selecionou as publicações que tratavam das competências do profissional da informação em bibliotecas frente ao uso de novas tecnologias com foco na AM, e por fim apenas 9 publicações foram selecionadas à luz do critério I-3, que recuperou publicações que apresentam reflexões acerca das questões éticas geradas a partir do uso de novas tecnologias em bibliotecas com foco em AM.

Gráfico 10. Publicações aprovadas segundo os critérios de inclusão



Critérios	Quantidade
I-1	38
I-2	10
I-3	9

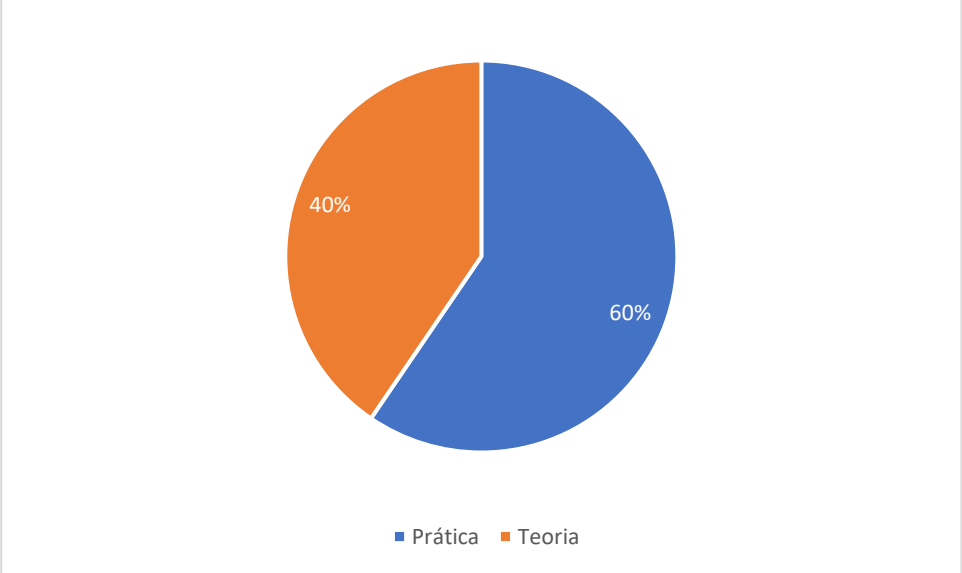
Fonte: Elaborado pela autora (2022).

Em síntese, observa-se uma tendência das publicações em AM aplicadas em bibliotecas discutirem em grande parte as aplicações em detrimento das competências profissionais dos bibliotecários e das repercussões éticas geradas diante deste novo cenário de inovação. Além disso, constatou-se que apenas 3 publicações não foram selecionadas a partir da aplicação do critério I-1, o parâmetro mais aplicado para seleção das publicações nesta RSL.

Por fim, as publicações selecionadas foram categorizadas em dois grupos a saber: publicações de natureza teórica, que buscam descrever, analisar ou explicar um determinado conhecimento, e publicações de natureza prática, que

apresentam ações concretas por meio da aplicação de conhecimentos, conforme o gráfico a seguir:

Gráfico 11. Publicações segundo a sua natureza prática ou teórica



Critério	Quantidade
Prática	25
Teoria	17

Fonte: Elaborado pela autora (2022).

Em síntese no conjunto de 42 documentos, 17 tratam-se de documentos de natureza teórica e 25 apresentaram aplicações de natureza prática, o qual relataram produtos e serviços desenvolvidos com a aplicação de técnicas de aprendizado de máquina.

A próxima seção apresentará uma análise mais aprofundada de cada subconjunto de documentos, segundo a sua natureza. Além disso, serão abordados breves resumos informativos de cada publicação que compõem esta RSL.

4.1.1 Análise das publicações de natureza teórica

O *corpus* documental cuja natureza é de cunho teórico, é composto por publicações que buscam revisar a literatura científica de um determinado período, apresentar discussões sobre projetos em ML em bibliotecas, aprimorar fundamentos teóricos e, apresentar um quadro conceitual aplicável ao momento de sua publicação.

Neste contexto, as publicações de natureza teórica, serão elencadas no quadro abaixo e, posteriormente sintetizadas de modo a apresentar um breve resumo de suas principais contribuições.

Quadro 7. Publicações de natureza teórica

Base	Título	Autor	Critério de seleção
Brapci	As perspectivas e aplicações da computação cognitiva em unidades de informação	Bárbara Coelho Neves	I-1 I-2
Brapci	Inteligência artificial e computação cognitiva em unidades de informação: conceitos e experiências	Bárbara Coelho Neves	I-1 I-2
ISTA	Machine learning for libraries and archives	Bohyun Kim	I-1
LISTA	Interactive web column: machine learning algorithms, in and out of libraries	Chana Kraus-Friedberg	I-1 I-2 I-3
LISTA	Machine learning and the library or: how I learned to stop worrying and love my robot overlords	Charlie Harper	I-1 I-3
LISTA	Explainable artificial intelligence	Michael Ridley	I-2 I-3
LISTA	Metadata automation: the current landscape and future developments	Marlee Graser; Melissa Burel	I-1 I-2
LISTA	Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization	Richard P. Smiraglia; Xin Cai	I-1
LISTA	AI and machine learning: The challenges of artificial intelligence in libraries	Jason Griffey	I-1 I-2
LISTA	Libraries in the age of artificial intelligence	Ben Johnson	I-1 I-2

LISTA	Exploring AI: how libraries are starting to apply artificial intelligence in their work	Loida Garcia-Febo	I-1 I-2
WoS	The intelligent library thought leaders' views on the likely impact of artificial intelligence on academic libraries	Andrew M. Cox; Stephen Pinfield; Sophie Rutter	I-1 I-2 I-3
WoS	Automatic subject indexing of text	Koraljka Golub	I-1
WoS	Algorithms: avoiding the implementation of institutional biases	Lori Ayre; Jim Craner	I-3
WoS	Digital libraries: the systems analysis perspective machine erudition	Robert Fox	I-1
WoS	The respective roles of intellectual creativity and automation in representing diversity: human and machine generated bias	Vanda Broughton	I-3
WoS	Thriving in the age of accelerations: a brief look at the societal effects of artificial intelligence and the opportunities for libraries	Kenning Arlitsch; Bruce Newell	I-2

Fonte: Elaborado pela autora (2022).

Na base de dados Brapci foram recuperadas e selecionadas 2 publicações de autoria de Bárbara Coelho Neves. O primeiro artigo publicado em 2019, "*As perspectivas e aplicações da computação cognitiva em unidades de informação*", aborda uma análise da literatura científica internacional e brasileira no campo da computação cognitiva (o qual abrangeu o conceito de *machine learning*) aplicada a unidades de informação, sob a perspectiva da cibercultura. A autora descreveu as principais definições da área, analisou o futuro do bibliotecário frente a este novo cenário e caracterizou o *learning analytics* e a curadoria digital. Descreveu ainda um projeto em andamento no laboratório da Universidade Federal da Bahia (UFBA), cujo objetivo é a aplicação da plataforma Watson da IBM em serviços de referência de bibliotecas. Já o segundo artigo publicado em 2021, sob o título "*Inteligência artificial e computação cognitiva em unidades de informação: conceitos e experiências*", de responsabilidade da mesma autora aborda de forma resumida as informações apresentadas na publicação anterior, porém em formato de artigo de revista.

Já na base de dados ISTA foi recuperado e selecionado o artigo "*Machine learning for libraries and archives*" de autoria de Bohyun Kim, publicado em 2021, o qual discorre sobre o *machine learning* como uma ferramenta útil para

unidades de informação. Menciona ainda algumas aplicações do ML em bibliotecas e arquivos, a saber: processamento de documentos em arquivos e bibliotecas, extração de conteúdo em documentos de texto completo, geração automática e enriquecimento de metadados, além de aplicações em conjunto de dados para treinamento em modelos de ML.

Na base de dados LISTA, foram recuperados e selecionados 8 artigos, analisados a seguir: *Interactive web column: machine learning algorithms, in and out of libraries*, por Chana Kraus-Friedberg, discorre acerca de como os algoritmos de aprendizado de máquina podem ser utilizados em bibliotecas especializadas em saúde ou acadêmicas, além disso, apresenta uma reflexão sobre as questões éticas geradas a partir da utilização de técnicas de ML em bibliotecas.

Já o artigo "*Machine learning and the library or: how I learned to stop worrying and love my robot overlords*, por Charlie Harper, reflete acerca das potenciais aplicações do ML em bibliotecas. Apresenta ainda uma terminologia básica sobre ML e da IA, bem como as possíveis questões éticas e de privacidade que as bibliotecas podem enfrentar a partir de sua utilização.

Em "Explainable artificial intelligence", de autoria de Michael Ridley, a Inteligência Artificial Explicável (XAI) é analisada. Apresenta o conceito da XAI, os objetivos, bem como as estratégias, técnicas e processos. Apresenta também as aplicações do XAI à luz do Regulamento Geral de Proteção de Dados da Europa, bem como a sua aplicação em bibliotecas.

O artigo *"Metadata automation: the current landscape and future developments"*, por Marlee Graser e Melissa Burel, discorre acerca dos impactos da automação e das tecnologias nas atividades exercidas pelos bibliotecários que trabalham com metadados. Além disso, destaca ainda aplicações do ML para a geração de metadados em imagens, bem como para o controle de qualidade.

Publicado em 2017, o artigo "*Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization"*, por Richard P. Smiraglia e Xin Cai, apresenta um estudo o qual destacaram quatro expressões no contexto da organização do conhecimento, a

saber: *clustering*, *machine learning*, indexação automática e classificação automática. Para isso realizaram um estudo baseado em citações na Bibliografia de Organização do Conhecimento publicada pela ISKO.

Em "AI and Machine Learning: the challenges of artificial intelligence in libraries", de autoria de Jason Griffey, discute os usos da inteligência artificial e do ML em bibliotecas. Destaca ainda a relevância das discussões sobre privacidade de dados tendo em vista o novo contexto tecnológico o qual as bibliotecas estão inseridas.

O artigo *"Libraries in the age of artificial intelligence"*, publicado por Ben Johnson, reflete acerca dos usos da IA em bibliotecas. Realiza uma análise comparativa entre as vantagens das bibliotecas que utilizam IA em seus serviços, com aquelas que desenvolvem seus produtos e serviços através de um bibliotecário humano. Destaca ainda alguns princípios que devem ser rediscutidos à luz da IA, a saber: acesso à informação, alfabetização informacional, privacidade dos dados e direito à liberdade de expressão.

Por fim, o artigo "*Exploring AI how libraries are starting to apply artificial intelligence in their work*", publicado por Loida Garcia-Febo, discorre acerca da implementação de aplicações de inteligência artificial em bibliotecas americanas. Discute ainda questões correlatas a acerca da implementação de IA e ML em bibliotecas tais como o direito à privacidade, à liberdade de expressão e o acesso à informação

Já na base de dados WoS foram recuperadas 6 publicações: "*The intelligent library thought leaders' views on the likely impact of artificial intelligence on academic libraries"*, publicado em 2018 por Andrew M. Cox, Stephen Pinfield e Sophie Rutter, o qual apresentam um estudo baseado em entrevistas com 33 gestores de biblioteca em 2017 sobre impactos da IA e do ML em bibliotecas universitárias.

Já o artigo "*Automatic Subject Indexing of Text*", publicado em 2019 por Koraljka Golub, explora as semelhanças e diferenças, vantagens e desvantagens do uso de sistemas de indexação automática atribuída a partir de processos de categorização, *clustering* e classificação. Destaca ainda que a

categorização de textos é talvez a abordagem de aprendizado de máquina mais difundida por geralmente apresentar bons resultados

Em *"Algorithms: avoiding the implementation of institutional biases"*, publicado em 2018 por Lori Ayre e Jim Craner, é explorado o conceito de algoritmos e as suas possíveis aplicações em bibliotecas, discute ainda as consequências éticas, bem como os *bias* (ou vieses) repercutem através da aplicação de algoritmos em bibliotecas.

O artigo *"Digital libraries: the systems analysis perspective machine erudition"*, publicado em 2016 por Robert Fox, explora o conceito de aprendizado de máquina à luz da Ciência da Informação e reflete acerca das implicações do ML em bibliotecas.

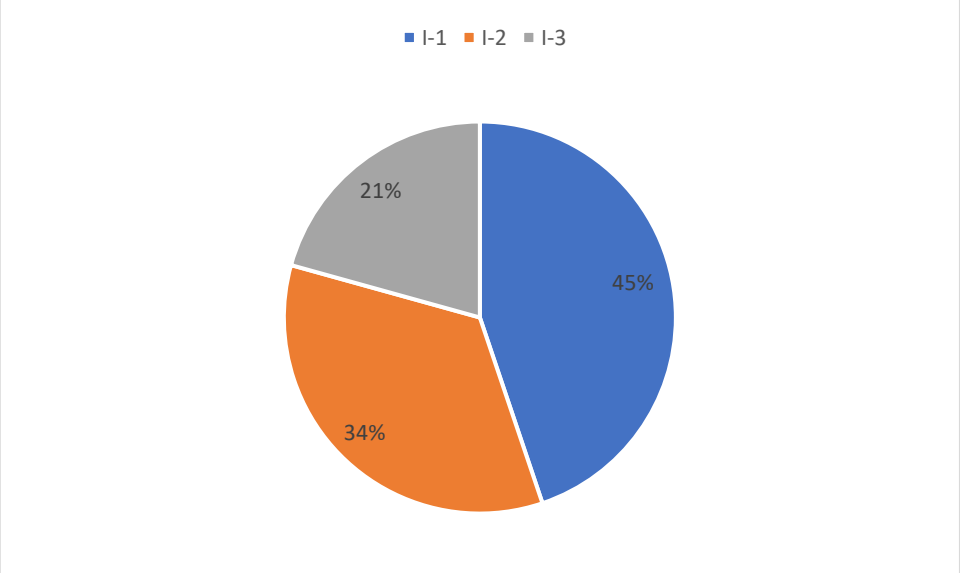
Publicado em 2019 por Vanda Broughton *"The respective roles of intellectual creativity and automation in representing diversity: human and machine generated Bias"*, discorre sobre as questões éticas em inteligência artificial, bem como métodos para resolver o problema de viés no campo da organização do conhecimento com destaque para a terminologia relacionada à religião.

Por fim, o artigo *"Thriving in the age of accelerations: a brief look at the societal effects of artificial intelligence and the opportunities for libraries"*, publicado em 2017 por Kenning Arlitsch e Bruce Newell, reflete acerca dos impactos da explosão informacional em conjunto com os efeitos da automação sobre a profissão de bibliotecário, bem como as consequências sociais e políticas, as ameaças e oportunidades para pesquisadores e bibliotecas públicas.

Em suma, o conjunto de publicações de natureza teórica é heterogêneo haja à vista que o conteúdo discutido nos documentos conjuga ao menos dois critérios de seleção em 59% das publicações que compõem o *corpus* teórico da RSL, concentradas em sua maioria na base de dados LISTA. Além disso, constata-se que os documentos que apresentam discussões acerca das competências dos profissionais de informação na perspectiva do ML (I-2), estão disponíveis apenas no subconjunto de documentos de natureza teórica,

conforme indica a representação gráfica abaixo, que detalha a aplicação de critérios de seleção no conjunto de documentos de natureza teórica.

Gráfico 12. Critério de seleção em documentos de natureza teórica no corpus da RSL



Critério	Quantidade
I-1	13
I-2	10
I-3	6

Fonte: Elaborado pela autora (2022).

Por fim, é importante destacar ainda que neste conjunto de documentos figuram duas publicações que apresentaram em suas discussões, os três critérios de inclusão simultaneamente, a saber: “Interactive web column: machine learning algorithms, in and out of libraries” e “The intelligent library thought leaders' views on the likely impact of artificial intelligence on academic libraries”.

A seguir será apresentado uma análise das publicações de natureza prática que compõem o *corpus*.

4.1.2 Análise das publicações de natureza prática

O *corpus* documental cuja natureza é de cunho prático, é composto por publicações que apresentam relatos que discutem produtos e serviços desenvolvidos através da aplicação de técnicas de *machine learning* em conjunto de dados textuais, imagéticos e/ou audiovisuais, nato-digitais ou digitalizados, em bibliotecas.

Neste contexto, assim como ocorreu com as publicações de natureza teórica, as publicações selecionadas serão elencadas no quadro abaixo e, posteriormente sintetizadas de modo a apresentar um breve resumo de suas principais contribuições.

Quadro 8. Publicações de natureza prática

Base	Título	Autor	Critério de seleção
Brapci	Mapping of ETDs in ProQuest Dissertations and Theses (PQDT) Global database (2014-2018)	Manika Lamba; Margam Madhusudhan	I-1
ISTA	An exploration of machine learning in libraries	Craig Boman	I-1 I-3
ISTA	Knowing what the patron wants: using predictive analytics to transform library decision making	Ryan Litsey; Weston Mauldin	I-1
LISTA	The democratization of artificial intelligence: one library's approach	Thomas Finley	I-1
LISTA	HAMLET: neural-net-powered prototypes for library discovery	Andromeda Yelton	I-1 I-3
LISTA	Metadata analytics, visualization, and optimization: experiments in statistical analysis of the Digital Public Library of America (DPLA)	Corey A. Harper	I-1
LISTA	AI and creating the first multidisciplinary AI Lab	Bohyun Kim	I-1 I-3
LISTA	Experimenting with a machine generated annotations pipeline	Joshua Gomez; Kristian Allen; Mark Matney; Tinuola Awopetu; Sharon Shafer	I-1

LISTA	Self-training author name disambiguation for information scarce scenarios	Anderson A. Ferreira; Adriano Veloso; Marcos André Gonçalves; Alberto H. F. Laender	I-1
LISTA	Using wavelet analysis for text categorization in digital libraries: a first experiment with Strathprints	Sándor Darányi; Peter Wittek; Milena Dobreva.	I-1
LISTA	Harnessing Apache Mahout to link content	Lim Chee Kiam; Balakumar Chinnasam	I-1
WoS	Automatic classification of older electronic texts into the Universal Decimal Classification-UDC	Matjaz Kragelj; Mirjana Kljajic Borstnar	I-1
WoS	Fusion architectures for automatic subject indexing under concept drift analysis and empirical results on short texts	Martin Toepfer; Christin Seifert	I-1
WoS	Automatic classification of Swedish metadata using Dewey Decimal Classification: a comparison of approaches	Koraljka Golub; Johan Hagelback; Anders Ardo	I-1
WoS	What can 100,000 books tell us about the international public library e-lending landscape?	Rebecca Giblin; Jenny Kennedy; Charlotte Pelletier; Julian Thomas; Kimberlee Weatherall; Francois Petitjean	I-1
WoS	Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: a machine-learning approach	Kevin W. Walker; Zhehan Jiang	I-1
WoS	Automated classification to improve the efficiency of weeding library collections	Kiri L. Wagstaff Geoffrey Z. Liu	I-1
WoS	Improving the visibility of library resources via mapping library subject headings to Wikipedia articles	Arash Joorabchi; Abdulhussain E. Mahdi	I-1
WoS	Account-based recommenders in open discovery environments	Jim Hahn; Courtney McDonald	I-1
WoS	Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts	Arash Joorabchi; Abdulhussain E. Mahdi	I-1

WoS	Combining domain-specific heuristics for author name disambiguation	Alan Filipe Santana; Marcos Andre Goncalves; Alberto H. F. Laender; Anderson Ferreira	I-1
WoS	Duplicate bibliographic record detection with an OCR-converted source of information	Shoichi Taniguchi	I-1
WoS	Classification of scientific publications according to library controlled vocabularies A new concept matching-based approach	Arash Joorabchi; Abdulhussain E. Mahdi	I-1
WoS	Cost-effective on-demand associative author name disambiguation	Adriano Veloso; Anderson A. Ferreira; Marcos Andre Goncalves; Alberto H. F. Laender; Wagner Meira Jr.	I-1
WoS	Automatic extraction of metadata from scientific publications for CRIS systems	Aleksandar Kovacevic Dragan Ivanovic Branko Milosavljevic Zora Konjovic Dusan Surla	I-1

Fonte: Elaborado pela autora (2022).

Na base de dados Brapci foi recuperado e selecionado o artigo "*Mapping of ETDs in ProQuest dissertations and theses (PQDT) global database* (*2014-2018*)", de autoria de Manika Lamba e Margam Madhusudhan, publicado em 2019, esta pesquisa apresenta um estudo sobre uma solução de modelagem de previsão e mineração em 441 teses e dissertações eletrônicas (ETDs) de texto completo do campo da biblioteconomia entre os anos de 2014 e 2018, extraídos da base *ProQuest Dissertations & Theses* (PQDT) utilizando a plataforma RapidMiner.

Já na base de dados ISTA 2 publicações foram selecionadas. A primeira publicação "*An exploration of machine learning in libraries"*, publicado em 2019 por Craig Boman, discorre acerca do uso da LDA (*Latent Dirichlet Allocation*) para a geração automática de cabeçalhos de assuntos, por meio de um estudo

de caso na coleção de *e-books* do Projeto Gutenberg. A pesquisa também descreve um fluxo para implementação de projetos de aprendizado de máquina em bibliotecas. A segunda publicação trata-se da "*Knowing what the paton wants: using predictive analytics to transform library decision making*", publicado em 2018 por Ryan Litsey e Weston Mauldin, o qual apresenta um estudo sobre o uso de um algoritmo de aprendizado de máquina, denominado *Automated Library Information Exchange Network* (ALIEN). Além disso, realiza uma análise preditiva de modo a compreender o comportamento das bibliotecas, por meio da análise de dados de circulação de itens do acervo, cujo finalidade é auxiliar os bibliotecários no desenvolvimento de coleções.

Na base LISTA foram selecionados 8 documentos, sintetizados a seguir: "*The democratization of artificial intelligence: one library's approach"*, por Thomas Finley, que relata um projeto da Biblioteca Pública de Frisco para o desenvolvimento de um programa de ensino sobre Python para os usuários, bem como o empréstimo de 150 kits para explorar a IA, dentre os quais, kits especializados em robótica, digitalização, programação e internet das coisas (IoT), disponibilizados para a comunidade local.

Em seguida, a publicação "*HAMLET: Neural-Net-Powered Prototypes for library"*, por Andromeda Yelton, apresenta um estudo acerca de um sistema de aprendizado de máquina que utiliza redes neurais, denominado Hamlet (*How about Machine Learning Enhancing Theses?*), desenvolvido pela autora. Ademais, o Hamlet utiliza o algoritmo doc2vec para alimentar as interfaces exploratórias e experimentais da coleção de teses da Massachusetts Institute of Tecnology (MIT).

O artigo "*Metadata analytics, visualization, and optimization: experiments in statistical analysis of the* Digital Public Library of America (DPLA)", de autoria de Corey A. Harper, destaca os conceitos de *analytics*, visualização e otimização de metadados. Além disso, explora o uso de técnicas de aprendizado de máquina para otimizar os metadados da Digital Public Library of America (DPLA).

Logo depois, o artigo "*AI and creating the first multidisciplinary AI Lab"*, por Bohyun Kim, discorre acerca da história da IA e o desenvolvimento do aprendizado de máquina. Destaca ainda aplicações que utilizam técnicas de IA/ML, e, por fim, apresenta as iniciativas em IA na biblioteca da Universidade de

Rhode Island, por meio da criação de um laboratório especializado em IA e ML para o desenvolvimento de pesquisas da comunidade local.

Publicado em 2020 o artigo "*Experimenting with a machine generated annotations pipeline*", de autoria de Joshua Gomez et al, apresenta um projeto de marcação automática de imagens na Biblioteca da Universidade da Califórnia em Los Angeles (UCLA), cujo objetivo é melhorar os resultados da pesquisa na biblioteca digital, por meio de metadados de serviços de marcação de imagens baseados em nuvem.

Em 2014 o artigo "*Self-Training author name disambiguation for information scarce scenarios"*, de autoria de Anderson A. Ferreira et al, discute os problemas que a ambiguidade ocasiona na descrição de nomes de autores em citações bibliográficas de documentos disponibilizados nas bibliotecas digitais da DBLP e BDBComp. Apresenta ainda uma solução para este problema, por meio de um método de autotreinamento em 3 etapas denominado SAND (self-training associative name disambiguator), com o objetivo de remover a ambiguidade no registro de nomes de autoridade.

Já o artigo "*Using wavelet analysis for text categorization in digital libraries: a first experiment with Strathprints"*, disponibilizada na base LISTA e de autoria de Sándor Darányi et al, discorre acerca dos benefícios da classificação automática em bibliotecas digitais. Apresentou um experimento piloto no repositório institucional Strathprints da Universidade de Strathclyde no Reino Unido com o auxílio de teste de Support Vector Machine (SVM) em um ambiente de aprendizado supervisionado para a indexação de objetos digitais utilizando os cabeçalhos de assunto da Biblioteca do Congresso Americana (LCSH).

Por fim, o artigo "*Harnessing Apache Mahout to link contente"*, de autoria de Lim Chee Kiam e Nalakumar Chinnasamy, discute a utilização do Apache Mahout na Biblioteca Nacional de Cingapura com o objetivo de vincular (linkar) conteúdos em várias coleções, com destaque para a coleção de artigos denominado Infopedia. Desta maneira, este projeto possibilitou a descoberta de ambas as coleções por meio da vinculação de conteúdo.

Na base de dados WoS foram recuperados e selecionados um total de 14 publicações de cunho cientifico, descritas a seguir: A primeira delas é o artigo

"*Automatic classification of older electronic texts into the Universal Decimal Classification–UDC"*, publicado em 2020 por Matjaz Kragelj e Mirjana Kljajic Borstnar, que discorre acerca da utilização de métodos de aprendizado de máquina na classificação automática de documentos antigos digitalizados, por meio da aplicação da Classificação Decimal Universal (CDU).

Em seguida, a publicação "*Fusion architectures for automatic subject indexing under concept drift: analysis and empirical results on short texts"*, publicado em 2020 por Martin Toepfer e Christin Seifert, apresentam as vantagens e desvantagens de sistemas de indexação automática em bibliotecas digitais e, discute ainda acerca das arquiteturas em sistemas de indexação automática com foco em documentos sobre economia.

O artigo *"Automatic classification of Swedish metadata using Dewey Decimal Classification: a comparison of approaches"*, publicado em 2020 por Koraljka Golub, Johan Hagelbäck e Anders Ardö, discorre sobre indexação e classificação automática com a aplicação da Classificação Decimal de Dewey (CDD), por meio de técnicas de ML, no catálogo da Biblioteca Nacional da Suécia. E, para realizar este estudo, comparou a performance de seis algoritmos de classificação.

Em *"What can 100,000 books tell us about the international public library e-lending landscape?"*, publicado em 2019 por Rebecca Giblin et al, descreve a prática do *e-lending* (empréstimo de publicações digitais) em bibliotecas, examinando as relações entre o preço do título, ano de publicação, editor dentre outras características que regem os preços das publicações, por meio de análises estatísticas que aplicam técnicas de aprendizado de máquina. Este estudo analisou os termos de licença e preços em cinco países: Austrália, Nova Zelândia, Canadá, Estados Unidos e Reino Unido, em uma plataforma agregadora de *e-books* denominada Overdrive.

Em seguida a publicação *"Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach"*, publicado em 2019 por Kevin W. Walker e Zhehan Jiang, explora como o uso do aprendizado de máquina pode ser mais eficaz para o desenvolvimento de coleções por meio da modelagem preditiva AdaBoost, de modo a estudar os padrões de aquisição em bibliotecas.

Já o artigo *"Automated classification to improve the efficiency of weeding library collections"*, publicado em 2018 por Kiri L. Wagstaff, Geoffrey Z. Liu, discorre acerca de um método para classificar automaticamente os possíveis documentos candidatos ao descarte em bibliotecas. Para isso, os autores executaram um estudo na biblioteca da Universidade de Wesleyan entre os anos de 2011 e 2014, de modo a treinar e analisar o desempenho de seis algoritmos de aprendizado de máquina com a finalidade de preverem decisões sobre a remoção ou permanência de documentos. Por fim, realizou-se uma análise estatística comparativa entre as previsões dos classificadores e as avaliações dos bibliotecários acerca do material candidato ao descarte.

A publicação *"Improving the visibility of library resources via mapping library subject headings to Wikipedia articles"*, publicado em 2018 por Arash Joorabchi e Abdulhussain E. Mahdi apresenta uma proposta para vincular descritores de registros bibliográficos do catálogo da OCLC aos artigos publicados na Wikipédia (através de *links*), com o objetivo de enriquecer e dar visibilidade aos recursos de bibliotecas, e para isto utiliza um sistema para mapeamento automático de cabeçalhos denominado FAST.

Logo em seguida o artigo, "*Account-based recommenders in open discovery environments"*, publicado em 2018 por Jim Hahn e Courtney McDonald, discorre acerca da aplicação do aprendizado de máquina em sistemas de descoberta de bibliotecas de modo a gerar recomendações personalizadas baseados nos fluxos de dados produzidos pelos próprios usuários.

Em *"Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts"*, publicado em 2014 por Arash Joorabchi e Abdulhussain E. Mahdi, apresentam uma pesquisa sobre a vinculação (*link*) entre registros de bibliotecas que compõem o catálogo da OCLC e artigos da Wikipedia por meio de um sistema automático para indexação de assuntos com o objetivo de realizar a integração entre coleções da biblioteca e artigos do Wikipedia. Para isso, treinou e implantou algoritmos genéricos de aprendizado de máquina para selecionar automaticamente os assuntos dos documentos das bibliotecas.

Já o artigo "*Combining domain-specific heuristics for author name disambiguation"*, publicado em 2014 por Alan Filipe Santana et al, apresenta um estudo sobre a desambiguação do nome do autor em citações de documentos presente nas bibliotecas digitais DBLP e BDBComp, por meio da aplicação de um conjunto de heurísticas e funções de similaridade, além de uma solução de supervisão para cada conjunto de dados específico.

O artigo "*Duplicate bibliographic record detection with an OCR-converted source of information"*, publicado em 2012 por Shoichi Taniguchi propõe um novo método para detecção de registros duplicados, por meio da utilização de reconhecimento óptico de caracteres (OCR) para registro de correspondência com o objetivo de detectar duplicações, empregando para isso técnicas de *Machine Learning*.

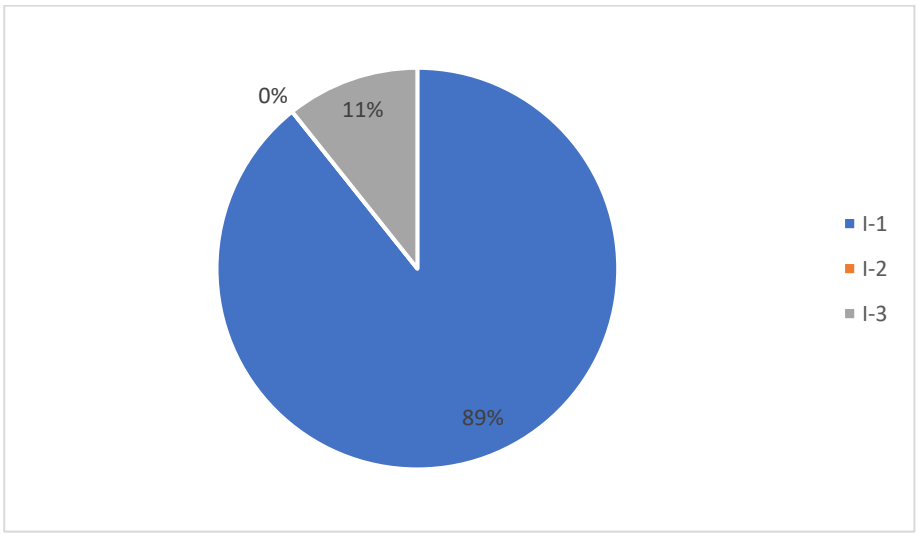
A publicação *"Classification of scientific publications according to library controlled vocabularies: a new concept matching-based approach"*, publicado em 2013 por Arash Joorabchi e Abdulhussain E. Mahdi descreve um projeto para o desenvolvimento de uma nova abordagem para sistemas automáticos de classificação e de indexação em documentos de pesquisa em bibliotecas e repositórios científicos digitais, utilizando para isso a Classificação Decimal de Dewey (CDD), o cabeçalho de assunto FAST e a enciclopédia livre Wikipedia para o mapeamento de conceitos.

Já em *"Cost-effective on-demand associative author name disambiguation"*, publicado em 2012 por Adriano Veloso et al., descreve um estudo sobre desambiguação de autoria, e propôs a comparação de três sistemas desambiguadores de nomes de autores associativos: EAND (Eager Associative Name Disambiguation), LAND (Lazy Associative Name Disambiguation) e SLAND (Self-Training LAND).

Por fim, o artigo *"Automatic extraction of metadata from scientific publications for CRIS systems"*, publicado em 2011 por Aleksandar Kovačević et al., descreve um sistema baseado em aprendizado de máquina para extração automática e classificação de metadados em artigos científicos em formato PDF para o sistema de informação de acompanhamento da pesquisa científica da Universidade de Novi Sad, denominado CRIS.

Em suma, diferente do que ocorre com o conjunto de publicações de teóricas que são de natureza heterogênea em sua maioria, os documentos de aplicação prática, buscavam contextualizar as aplicações práticas, bem como a execução de técnicas de ML em detrimento da descrição de elementos éticos ou laborais, conforme apresentado no gráfico abaixo:

Gráfico 13. Critério de seleção em documentos de natureza prática no *corpus* da RSL



Critério	Quantidade
I-1	25
I-2	0
I-3	3

Fonte: Elaborado pela autora (2022).

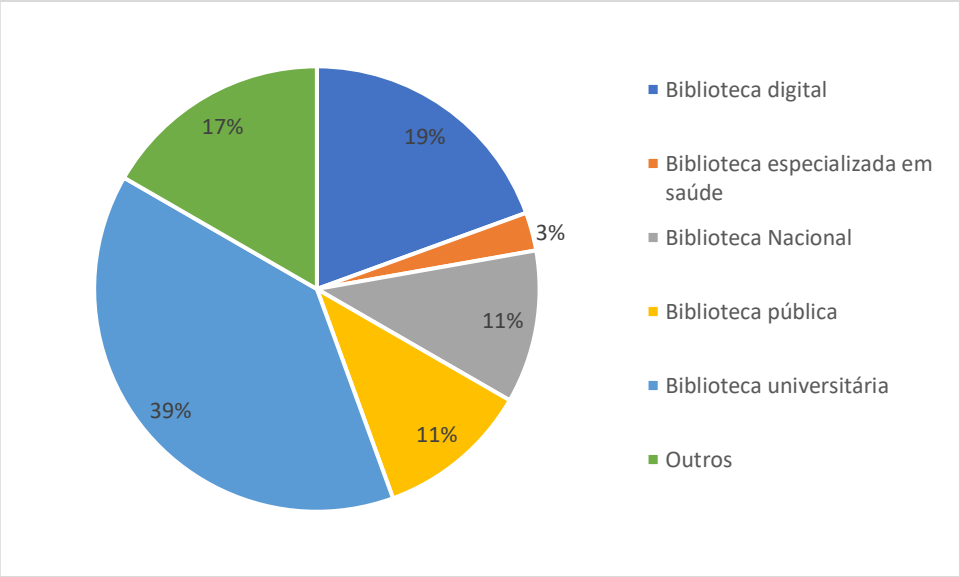
Além disso, constata-se ainda que nenhum documento de cunho prático identificou as competências ou habilidades que os profissionais da informação devem dispor neste contexto de inovação, e que em poucos documentos deste conjunto existe a preocupação de relatar questões relativas ao processo ético da execução do projeto.

A respeito da fonte de informação com o maior número de contribuições práticas, destaca-se a base *Web of Science* com 70% de suas publicações que discutiram estudos de casos à luz do critério de inclusão I-1 desta RSL.

No que se refere ao tipo de instituição em que foram aplicadas técnicas de ML, observa-se um número expressivo de bibliotecas associadas à Universidades, conforme o gráfico 13, o que é natural haja vista à proximidade

com o ambiente de inovação proporcionado por estas instituições. Logo em seguida, com 17% temos a contribuição de instituições que realizam consórcios entre bibliotecas como a OCLC, instituições sem fins lucrativos e base de dados. E em terceiro lugar, as bibliotecas digitais que são instituições propícias para a elaboração de projetos que utilizem o ML em suas técnicas de aplicação, haja vista estas instituições reunirem grandes conjuntos de dados em formato digital ou nato-digital.

Gráfico 14. Aplicação de ML em bibliotecas



Fonte: Elaborado pela autora (2022).

Destacam-se ainda, as contribuições advindas de Bibliotecas Nacionais, instituições que em sua maioria são provenientes de países europeus e, que tratam-se de grandes centros nacionais de informação bibliográfica responsáveis pelo controle bibliográfico de importantes silos de informação. Neste contexto, as bibliotecas nacionais tornam-se um ambiente ideal para a aplicação de técnicas de ML em vastos conjuntos de documentos nato-digitais ou digitalizados, haja vista a disponibilidade limitada de profissionais da informação para realizar o processamento manual destes grandes conjuntos de informação.

Já em relação às contribuições das bibliotecas públicas, tanto as tradicionais quanto às digitais, observa-se a partir de diferentes perspectivas, pois as bibliotecas públicas ditas tradicionais buscaram aplicar as técnicas de IA

e Ml por meio de projetos de Alfabetização de dados, já as bibliotecas públicas digitais aplicaram ML em projetos para o processamento e tratamento de conjunto de dados digitais.

Por fim, figurando-se em último, as bibliotecas especializadas, com contribuições apenas de cunho teórico.

4.2 Apresentação dos dados de pesquisa à luz dos critérios de inclusão

Esta seção apresentará uma análise das contribuições das pesquisas selecionadas para compor o *corpus*, de acordo com o critério de seleção dos documentos.

4.2.1 Aplicações, produtos e serviços, impactos, benefícios e dificuldades na implantação de AM em Bibliotecas

À luz do primeiro critério de seleção das publicações, observa-se na literatura que as bibliotecas têm buscado se adaptar a diferentes cenários ao longo das últimas décadas com o auxílio da tecnologia. À vista disso, e tendo em vista o atual contexto marcado pela conectividade, digitalização, automação e integração de dados, as bibliotecas têm utilizado, de maneira gradual, técnicas baseadas em inteligência artificial e aprendizado de máquina, haja vista disporem de um ambiente propício para a coleta e reunião de informações, pois em um único dia, elas serão capazes de selecionar, catalogar, classificar, indexar e armazenar grandes quantidades de informação.

Deste modo, buscou-se na literatura, mapear as aplicações em produtos e serviços, relatadas a partir do uso de técnicas de *machine learning* em projetos de bibliotecas, que serão sintetizados no quadro abaixo, segundo a sua categoria de aplicação.

Quadro 9. Aplicações, produtos e serviços relatados na literatura

Aplicações, produtos e serviços relatados na literatura	
Alfabetização de dados	Kits 3D
	Kits de circuitos sem solda
	Kits de digitalização
	Kits de Internet das Coisas (IoT)
	Kits de programação
	Kits de robótica
	Laboratório multidisciplinar de IA
Desenvolvimento de coleções e circulação	Empréstimo entre bibliotecas
	Aquisição de e-books
	Seleção – Desbaste, descarte
Imagem	Anotação automática de imagem
	Marcação automática de imagens
	Processamento automático de imagens
Texto	Análise de metadados
	Enriquecimento de metadados
	Extração automática de metadados
	Geração automática de metadados
	Visualização de metadados
	Classificação automática
	Indexação automática
	Categorização de textos
	Vinculação e mapeamento de dados
	Desambiguação de nomes de autores
	Extração de conteúdo
	Detecção de registos duplicados
Recuperação da Informação	Personalização de conteúdo
	Recomendação
	Descoberta

Fonte: Elaborado pela autora (2022).

Os primeiros elementos relatados na literatura, foram produzidos com foco na Alfabetização de dados de usuários, e que, portanto, tem por objetivo

conferir a habilidade ao usuário de explorar, entender e se comunicar com dados. Isso inclui habilidades de pensamento crítico para usar, interpretar e tomar decisões com dados, e depois, transmitir a importância e o valor deles a outras pessoas (TABLEAU, 2022).

À vista disso, na Biblioteca pública de Frisco no Texas[25], Thomas Finley (2019) apresentou um projeto o qual foram disponibilizados 20 diferentes tipos de kits produzidos para a comunidade, com o objetivo de disponibilizar ferramentas gratuitas que empregam IA e ML e, torná-las acessíveis à comunidade local. A maioria dos kits contém guias com o objetivo de demonstrar de forma rápida e simplificada as principais funcionalidades e, possuem temáticas voltadas para a programação, robótica, solda, digitalização em 3D e *internet* das coisas. Ademais, dentre as ferramentas utilizadas nos denominados *Maker Kits* podemos citar as linguagens Python, AIY Voice Project kit e Raspberry Pi.

Já a Biblioteca da Universidade de Rhode Island projetou um laboratório multidisciplinar de AI denominado AI Lab[26], com o objetivo de dar suporte aos estudantes e professores da instituição, e desta forma, torna-se um centro de inovação e aprendizado em diversos cursos que tradicionalmente aplicam a AI e ML em suas disciplinas e, também aos que exploram suas funcionalidades à luz de diferentes perspectivas, tais como a filosofia e a psicologia. O objetivo do AI Lab é integrar essas discussões entre diferentes áreas do conhecimento, de modo a promover o pensamento interdisciplinar acerca da temática. Ademais, Kim (2019, p. 19) apresentou alguns projetos que estão em processo de desenvolvimento no laboratório, tais como: programação em *deep learning* de robôs equipados com câmeras, radares e sensores; construção de algoritmos de IA para esses robôs navegarem em ambientes conhecidos e/ou desconhecidos e, acesso e análise a uma variedade de conjuntos de *big data*.

Já sob à perspectiva do desenvolvimento de coleções, várias aplicações foram elaboradas e discutidas na literatura, utilizando técnicas de ML em seus projetos, com o objetivo de analisar dados de circulação de modo a melhor

[25]Biblioteca Pública de Frisco: https://friscolibrary.bibliocommons.com/v2/search?query=Maker+Kits&searchType=smart&_ga=2.46993698.423439180.1658780896-1115834608.1658780896

[26] Lab AI: https://web.uri.edu/ai/

compreender os padrões de movimentação do acervo (LITSEY; MAULDIN, 2018; WALKER; JIANG, 2019), analisar o comportamento do usuário acerca do uso de coleções (LITSEY; MAULDIN, 2018), apresentar quais as principais características que regem os preços dos *ebooks* em diferentes países (GIBLIN et al, 2019) e, quais obras serão candidatas ao descarte em uma biblioteca universitária (WAGSTAFF; LIU, 2018).

Nesta seara, compreender o padrão de circulação dos itens de uma biblioteca de modo a melhor atender as necessidades de informação dos usuários, bem como realizar recomendações a fim de prever ações que possam auxiliar na tomada de decisão em bibliotecas, são as principais funções do *Automated Library Information Exchange Network* (ALIEN), proposto Litsey e Mauldin (2018). O ALIEN, de acordo com os autores (2018, p. 140) é a primeira ferramenta de aprendizado de máquina com análise preditiva desenvolvida com foco em bibliotecas. O seu objetivo é investigar o comportamento de uma biblioteca por meio da análise de diferentes tipos de dados de modo a recomendar e prever serviços e ações para a instituição. Por exemplo, o sistema pode reconhecer uma mudança no comportamento dos usuários e, a biblioteca a partir das previsões do sistema, pode adaptar as coleções de maneira a se adequar às novas necessidades de informação dos usuários. À vista disso, Litsey e Mauldin (2018) aplicaram o ALIEN em um estudo acerca dos empréstimos entre bibliotecas do Texas Tech University, com o objetivo de compreender o comportamento desse serviço no contexto desta instituição.

Já Giblin et al (2019) investigou a disponibilidade de *e-books* para empréstimo eletrônico em bibliotecas de cinco países: Austrália, Canadá, EUA, Nova Zelândia e Reino Unido, por meio do agregador de conteúdos Overdrive. Para este estudo foi utilizado um algoritmo de modo a examinar as relações entre o preço do título, ano de publicação, termos de uso, jurisdição, editora e tipo de editora, aplicando análises estatísticas e aprendizado de máquina em algumas fases da pesquisa, a fim de compreender as principais características que regem o preço dos *e-books* entre os países anglófonos. Para o desenvolvimento deste estudo foi utilizado a árvore de regressão e o cálculo de distância de Levenshtein.

Uma outra aplicação no campo do desenvolvimento de coleções foi descrita na literatura por Walker e Jiang (2019). Os pesquisadores utilizaram o algoritmo de aprendizado AdaBoost de modo a realizar uma análise preditiva para a aquisição de livros na modalidade de Aquisição Orientada por Demanda (DDA - *Demand Driven Acquisitions*) em bibliotecas universitárias. A DDA, de acordo com Walker e Jiang (2019), baseia-se em uma premissa bastante simples, pague apenas pelos títulos que os seus usuários utilizam. Desta maneira, o estudo explora como o aprendizado de máquina pode auxiliar o desenvolvimento de coleções por meio de estratégias de gerenciamento que utilizam a modelagem preditiva em dados de uso de coleções, de modo a compreender os padrões de compra em bibliotecas universitárias.

Por fim, uma outra iniciativa foi o emprego da classificação automática para o descarte de obras do acervo da biblioteca da Universidade de Wesleyan. Para isso, o estudo conduzido por Wagstaff e Liu (2018), buscou avaliar empiricamente métodos para classificação automática de obras candidatos ao descarte, comparando o desempenho de seis algoritmos de classificação (Nearest-neighbor classifier, Naive Bayes classifier, Decision tree, Random forest, Support vector machine), em um conjunto de 80.346 itens da biblioteca. Para este experimento foi utilizado Python em conjunto com o Scikit-learn.

No domínio da aplicação de técnicas de ML em conjunto de dados imagéticos, foram citadas três aplicações nas publicações selecionadas na RSL, a saber: processamento automático de imagens, anotação automática de imagens e, marcação automática de imagens. O processamento automático de imagens de acordo com Thomé (2004, p. 6) consiste na manipulação de uma imagem através de algoritmos implementados em computador de modo que a entrada e a saída do processo sejam imagens ou informações extraídas da imagem. No que se refere à anotação automática de imagens e a marcação automática de imagens, as duas expressões são consideradas sinônimas por Siddiqui, Mishra e Verma (2015, p. 27), o qual trata-se de um processo por sistema de computador que atribui automaticamente metadados na forma de palavras-chave a uma imagem digital, aplicada em sistemas de recuperação de imagens de modo a organizar e recuperar imagens em um banco de dados.

Neste contexto, a Biblioteca da Universidade da Califórnia realizou um experimento conduzido pelo Labs Team (GOMEZ et al, 2020) em sua biblioteca digital, com o objetivo de melhorar os resultados de pesquisa de imagens através de quatro serviços de marcação de imagens baseado em nuvem: AWS Rekognition, Clarafai, Google Vision e Microsoft Azure CV. A finalidade do projeto é *taggear* imagens de três coleções de fotografias jornalísticas que datava entre as décadas de 1920 e 1960 e, determinar se essas *tags* poderiam aprimorar os metadados já existentes.

Já Bohyun Kim (2021) citou em seu trabalho o projeto Aida (Image Analysis for Archival Discovery), uma parceria entre a University of Nebraska–Lincoln e a Library of Congress, o qual um dos objetivos é aplicar uma série de métodos de zoneamento e segmentação de imagens em ML com o propósito de identificar e classificar o conteúdo gráfico em uma coleção de documentos manuscritos.

No que tange a aplicação de técnicas de ML em metadados de bibliotecas, houve a menção na literatura de técnicas de análise de metadados (LAMBA; MADHUSUDHAN, 2019), enriquecimento de metadados, extração automática de metadados (KOVAČEVIĆ et al, 2011; LAMBA; MADHUSUDHAN, 2019), geração automática de metadados (KIM, 2021; BOMAN, 2019), automação de metadados (GRASER; BUREL, 2018), visualização de metadados (HARPER, 2016) e vinculação e o mapeamento de dados (JOORABCHI; MAHDI, 2017, 2014, 2013; KIM; CHINNASAMY, 2013).

A análise de metadados tem como objetivo investigar os padrões de uso, a estrutura e/ou a qualidade dos metadados utilizados em projetos de pesquisa em bibliotecas digitais. Já o enriquecimento de metadados, segundo Lira (2014, p. 30), trata-se do processo de atribuir maior significado aos metadados e dados por intermédio da aplicação de recursos auxiliares, objetivando facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. A extração automática de metadados ou a geração automática de metadados são técnicas que realizam a extração de informações em campos específicos, agrupando e armazenando os resultados de forma estruturada em banco de dados ou arquivos XML, de modo a permitir que possam ser pesquisados e analisados (CORTEZ; SILVA, 2010). A visualização de metadados

refere-se a uma técnica que permite compreender padrões e apresentar as características dos metadados em um determinado contexto (HARPER, 2016). Por fim, a vinculação e o mapeamento de recursos, de acordo com Lira (2014, p. 30), consiste em descobrir *links* entre as combinações semânticas dos dados e metadados com outros recursos na *web*.

Bohyun Kim (2021, p. 40) elencou dois projetos que aplicam técnicas em ML na geração automática de metadados. O primeiro deles é AMPPD (The Audiovisual Metadata Platform Pilot Development) realizado pela Universidade de Indiana com o apoio da Fundação Andrew W. Mellon e em colaboração com a Universidade do Texas, a empresa de consultoria digital AVP e a Biblioteca Pública de Nova Iorque, cujo objetivo é gerar e gerenciar metadados para materiais audiovisuais em escala para bibliotecas e arquivos com o objetivo de auxiliar os catalogadores humanos na tarefa e não os substituí-los. Outro projeto destacado por Bohyun (2021, p. 40) é o CAMPI (Computer Aided Metadata Generation for Photo Archives Initiative), projeto desenvolvido com foco na coleção de fotografias da Universidade Carnegie Mellon. A equipe da Universidade elaborou um projeto para geração de metadados a partir de um protótipo de aplicativo que utiliza a visão computacional.

Craig Boman (2019) desenvolveu um estudo para a geração automática de cabeçalhos de assuntos na coleção de *ebooks* do Projeto Gutemberg[27], com a aplicação de Latent Dirichlet Allocation (LDA) na modelagem de tópicos de *ebooks*. Já Harper (2016) desenvolveu um estudo na Digital Public Library of America (DPLA) com o propósito de aplicar técnicas baseadas em análises estatísticas a fim de analisar, visualizar e otimizar metadados. A análise foi conduzida com o objetivo de verificar as relações entre os metadados de assunto e as consultas dos usuários, bem como para construir modelos de regressão utilizando dados de uso e metadados como preditores das estatísticas de uso do item.

A extração automática de metadados em publicações científicas da Universidade de Novi Sad, é o foco do artigo de autoria de Kovačević et al (2011). O sistema é baseado em aprendizado de máquina e realiza extração e

[27] Github repository for Gutemberg project - https://github.com/craigboman/gutenberg

classificação automática de metadados em oito categorias pré-definidas, a saber: título, autores, filiação, endereço, e-mail, resumo, palavras-chave e notas de conteúdo. A tarefa de extração é realizada por meio de um processo de classificação utilizando os algoritmos Decison Tree, Naive Bayes, K-nearest Neighbours e Support Vector Machines, a fim de avaliar o desempenho de cada algoritmo classificador.

Graser e Burel (2018) realizaram um estudo teórico com o objetivo de apresentar uma série de ferramentas que podem ser utilizadas na automação de metadados em fluxos de trabalho do profissional da informação, dentre as quais se destacam o XML, Python, OAI-PMH, Openrefine e o MarcEdit. Além disso, os autores (2018, p. 9) enfatizam que os bibliotecários de metadados devem possuir conhecimentos não apenas em descrição, aplicação e gerenciamento de metadados em vários esquemas, como também em ferramentas de código aberto para limpeza de dados e uma série de outras habilidades técnicas necessárias para gerenciar a automação de metadados.

Já no estudo *Mapping of ETDs in ProQuest dissertations and theses* (PQDT) global database (2014-2018), Lamba e Madhusudhan (2019) apresentaram uma solução para o gerenciamento e extração do conhecimento em um conjunto de 442 teses e dissertações em texto completo no domínio da biblioteconomia, em língua inglesa, extraídas da base de dados ProQuest. Para isso, o estudo contou com uma análise de metadados a fim de determinar as universidades e departamentos de destaque, tipos e níveis de graus e a localização geográfica das teses e dissertações. Em seguida foi realizado anotação por meio de modelagem de tópico utilizando a plataforma RapidMiner, e por fim foi aplicado um modelo de previsão utilizando o classificador Support Vector Machine (SVM).

Joorabchi e Mahdi (2017; 2014; 2013) apresentaram um projeto com o objetivo de melhorar a visibilidade dos recursos da biblioteca por meio da vinculação e mapeamento de cabeçalhos de assunto da biblioteca em artigos da enciclopédia livre Wikipédia de modo a conectar os conjuntos de registros. Este estudo utilizou dados desta enciclopédia e do catálogo WorldCat do consórcio OCLC. Destaca-se ainda que o Wikipédia é utilizado como vocabulário de modo a enriquecer os metadados de assunto dos registros.

Um outro estudo de caso semelhante, foi conduzido por Kiam e Chinnasamy (2013), o qual aplicaram o Apache Mahout com o objetivo de vincular conteúdos entre a Infopedia collection e o catálogo da Biblioteca Nacional de Singapura.

Já no domínio da representação do conhecimento houve a contribuição de diversos estudos práticos e trabalhos de natureza teórica (SMIRAGLIA; XIN, 2017) que buscaram discutir e/ou aplicar a classificação automática (GOLUB; HAGELBÄCK; ARDÖ, 2020; JOORABCHI; MAHDI, 2013, 2017; KRAGELJ; BORSTNAR, 2020), a categorização de textos (GOLUB, 2019; DARÁNYI; WITTEK; MCPHERSON, 2012) e a indexação automática (GOLUB, 2019; JOORABCHI; MAHDI, 2014, 2017) em conjuntos de documentos digitais ou nato digitais.

Smiraglia e Xin (2017) apresentaram a evolução de quatro conceitos utilizados no domínio da organização do conhecimento, o qual serão destacados nesta seção: o *clustering*, a indexação automática e a classificação automática. O *clustering*, de acordo Smiraglia e Xin (2017), envolve a análise de dados com o objetivo de agrupá-los de acordo com alguma medida de distância. Os documentos de uma coleção são vistos como pontos em um espaço e são categorizados como membros de grupos de acordo com a distância relativa (SMIRAGLIA; XIN, 2017, p. 216). Já a classificação automática refere-se a uma aplicação que cria modelos que associam documentos a categorias semanticamente semelhantes" (SALLES et al. 2016, p. 2). E, por fim, a indexação automática, que de acordo com Golub (2019, p. 106), denota processos não intelectuais, baseados em máquina, de indexação de assuntos, para fins de recuperação aprimorada de informações.

Nesta seara da representação do conhecimento, Golub, Hagelbäck e Ardö (2020) relataram o projeto de classificação automática Scorpion de iniciativa da OCLC, cujo objetivo é utilizar a Classificação Decimal de Dewey (CDD) como metodologia na atribuição automática de assuntos para itens eletrônicos (THOMPSON; SHAFER; VIZINE-GOETZ, 1997). Ademais, o Scorpion cria um conjunto de *clusters* de referência para classes CDD e implementa uma medida de distância termo-frequência de modo a encontrar o *cluster* relevante (e

consequentemente a classe CDD) para o documento a ser classificado (JOORABCHI; MAHDI, 2013, p. 726)

Em 2014 e 2017 Joorabchi e Mahdi desenvolveram um sistema automático para classificação e indexação por assunto de registros de metadados de bibliotecas com conceitos elaborados pela Wikipédia, por meio de uma abordagem baseada em correspondência de conceitos. Neste projeto foram utilizadas as ferramentas Weka, Wikipedia-Miner e, foram avaliados os desempenhos dos algoritmos classificadores *Logistic regression*, *Bayes Network*, *Multilayer Perpeptron*, *Decision Tree* e *Random Forest*, com o objetivo de classificar cada conceito candidato como "correspondente" ou "não correspondente" (JOORABCHI; MAHDI, 2017, p. 57).

Kragelj e Borstnar (2020) relataram um estudo cujo foco é o desenvolvimento de um modelo para classificação automática de textos antigos digitalizados e disponibilizados por meio da Biblioteca Digital da Eslovênia, utilizando para isso métodos de aprendizado de máquina e a Classificação Decimal Universal (CDU). A etapa de treinamento do estudo utilizou um *corpus* de 70.000 textos acadêmicos processados por bibliotecários. Além disso, foi realizada uma análise de *clustering* com a aplicação do algoritmo k-means. Por fim, para a implementação do modelo foram testados os algoritmos de aprendizado supervisionado *Linear regression*, *Naive Bayes*, *Support Vector Machine*, K-*Nearest Neighbours* e *Multilayer Perceptron*.

Golub, Hagelbäck e Ardö (2020) desenvolveram também um estudo de classificação automática, de modo a aplicar a Classificação Decimal de Dewey (CDD) em coleções digitais, a fim de avaliar o desempenho de seis algoritmos de aprendizado de máquina, bem como um algoritmo de correspondência de *strings* baseado nas características da CDD. No estudo foram utilizados registos catalográficos do catálogo mantido pela Biblioteca Nacional da Suécia denominado LIBRIS[28]. À vista disso, foram aplicados e avaliados o desempenho dos algoritmos: *Support Vector Machine, Multinomial Naïve Bayes, Simple linear network, Standard neural network, 1D convolutional neural network* e *Recurrent neural network.*

[28] LIBRIS: https://libris.kb.se/?language=en

No domínio da categorização de textos, que de acordo com Golub, Hagelbäck e Ardö (2020, p. 19) é frequentemente empregada para classificação automática de texto livre. As abordagens de categorização de texto podem ser divididas em *hard* e *soft*, de acordo com Golub (2019, p. 109) apud Sebastiani (2002), na abordagem *hard*, decide-se se o documento pertence ou não a uma categoria; já na abordagem *soft*, uma lista classificada de categorias de candidatos é criada para cada documento. Neste contexto, Darányi, Wittek e McPherson (2012) conduziram um estudo no Repositório de Strathprints da Universidade de Strathclyde e, para isso utilizaram a análise Wavelet para a categorização de textos no Repositório de Strathprints, cujo objetivo é demonstrar que a categorização de textos pode ser aplicada para analisar a cobertura temática em repositórios digitais. Neste estudo, foi utilizado o algoritmo Support Vector Machine de modo a reproduzir a classificação do conjunto teste de 6.000 objetos digitais indexados por meio da aplicação do cabeçalho de assunto Library of Congress Subject Headings (LCSH).

Por fim, no âmbito da representação do conhecimento, a indexação automática que, de acordo com Golub (2019, p. 104), pode ser utilizada tanto para enriquecer registros de metadados existentes, como para estabelecer conexões entre recursos de vários metadados e coleções, além de melhorar a consistência dos metadados. Gil Leiva (2017, p. 140) definiu indexação automática à luz de três perspectivas, a saber:

> a) Programas de computador que auxiliam no processo de armazenamento de termos de indexação, uma vez obtidos intelectualmente (ou seja, indexação assistida por computador durante o armazenamento); b) Sistemas que analisam documentos automaticamente, mas os termos de indexação propostos são validados e publicados, se necessário, por um profissional (indexação semiautomática); e, c) projetos desenvolvidos sem programas sem validação, (ou seja, os termos propostos são armazenados diretamente como descritores desse documento).

Neste contexto, Godby e Reighart apresentaram o projeto da OCLC denominado WordSmith. A ideia central deste projeto, de acordo com Golub (2019, p. 112) foi desenvolver um *software* para extrair frases nominais significativas de um documento, cujo objetivo era alcançar a precisão da indexação automática, através de um classificador que apresentasse uma lista dos sintagmas nominais mais significativos em um determinado conjunto de

documentos. Para isso, foi implementado uma série de filtros estatísticos a fim de identificar o vocabulário descritivo em coleções de texto em inglês (GODBY; REIGHART, 2008).

Uma aplicação outra relacionada ao uso de indexação automática foi produzida por Toepfer e Seifert (2018) o qual buscaram realizar uma análise detalhada das arquiteturas associativas, lexicais e de fusão apoiadas através de estudo empírico em textos de conteúdo econômico, cuja fonte é a Biblioteca Nacional de Economia da Alemanha, considerando para isso a dinâmica entre termos e conceitos. Para este experimento os autores destacaram o uso de duas abordagens: o MAUI (*Multi-purpose Automatic Topic Indexing*) para produzir predições com conhecimento lexical e, abordagens relacionadas ao SGDSVM (*Stochastic Gradient Descent – Support Vector Machine*) para predição de forma associativa (2018, p. 11). Além disso, foram utilizados o BRLR (Classificador de regressão logística), RHACK, MONQ, Python e Scikit-learn.

Já no domínio da representação da informação, Bohyun Kim (2021) afirma que muitas tarefas que atualmente são executadas por profissionais qualificados podem ser automatizadas por aplicativos de AI e ML. Kim (2021, p. 39) ainda destaca que as técnicas de ML podem melhorar o trabalho de processamento de modo a trazer maior eficiência e profundidade em escala na descrição de materiais tornando as coleções mais detectáveis.

À vista disso, Ferreira et al (2014), Santana et al (2014) e Veloso et al (2012) desenvolveram um trabalho com o objetivo de eliminar a ambiguidade no registro de nomes de autores em um conjunto de registro de citação, aqui entendido, de acordo com Ferreira et al (2014, p. 1257), como um conjunto de características bibliográficas, como nomes de autores e coautores, título da obra e local de publicação, de uma publicação específica. A ambiguidade pode ser causada por diversos motivos, incluindo a falta de padrões e práticas comuns na descrição de registros e a geração descentralizada de conteúdo (por exemplo, por meio de coleta automática). Para isso os autores utilizaram um conjunto de heurísticas e *clusters* aplicados no acervo do *Digital Bibliography & Library Project* (DBLP) e da Biblioteca Digital Brasileira de Computação (BDBComp) de maneira a apresentar o EAND (*Eager Associative Name Disambiguation*), SAND (*Self-taining Assocative Name Disambiguator*), SLAND (*Self-Training* LAND) e o

LAND (*Lazy Associative Name Disambiguation*), que foram comparados com outros métodos, a saber: *Support Vector Machine, Naive Bayes e Cosine*

Já em "*Duplicate bibliographic record detection with na OCR-coverted source of information*" Taniguchi (2012) desenvolveu um estudo para detecção de registro bibliográfico duplicado por meio de um método que usa a fonte de informação convertida por meio de reconhecimento óptico de caracteres (OCR). No experimento foram utilizados dois conjuntos de dados, o primeiro conjunto trata-se de 225 registros bibliográficos da Biblioteca da Universidade de Tsukuba e de sua respectiva fonte de informação tratada com OCR. Já o segundo conjunto de dados conta com registros bibliográficos provenientes dos catálogos da WorldCat da OCLC e da *Libraries* Australia (um serviço gerenciado pela Biblioteca Nacional da Austrália). O experimento foi conduzido utilizando *machine learning* para a detecção de duplicatas, por meio da aplicação de aprendizado supervisionado com a ferramenta Weka em conjunto com alguns algoritmos, a saber: Decision Tree, Naive Bayes, Random Forest, Support Vector *Machine* com *linear kernel*, *AdaBoost* com *Decision Stumps* e *Bagging* com *REPTree*. Esse método foi adotado pelas Bibliotecas da Universidade da Califórnia.

Por fim, as aplicações do *machine learning* produzidas no campo da recuperação da informação, que de acordo com Cunnigham, Littin e Witten (1997), ocorreram somente a partir do final dos anos 80. Os algoritmos de aprendizado de máquina utilizam atributos e valores, que os sistemas de recuperação de informação podem fornecer em abundância, além disso, processos de indexação e classificação automáticas por meio de algoritmos de aprendizado de máquina auxiliam em processos de descoberta e recomendação de conteúdo.

Yelton (2019) idealizou um sistema baseado em rede neural denominada HAMLET[29] (*How about Machine Learning Enhancing Theses*?). Em sua concepção o Hamlet utiliza o algoritmo doc2vec para estimar a similaridade entre diferentes documentos. Ademais, o HAMLET[30] possui três interfaces (YELTON, 2019, p. 12) em seu protótipo: um mecanismo de recomendação, que permite

[29] Github HAMLET: https://github.com/thatandromeda/hamlet
[30] Site que hospeda o HAMLET: https://hamlet.andromedayelton.com/

pesquisar teses por autor ou título e informa quais outras teses são conceitualmente semelhantes; uma interface para carregar arquivos, cujo objetivo é semelhante a primeira interface, e por fim, uma interface voltada para a revisão de literatura, com a finalidade de descobrir quais trabalhos foram citados por teses idênticas ao conteúdo pesquisado.

Hahn e McDonald (2017) desenvolveram um protótipo de serviço de recomendação baseado em aprendizado de máquina para ambientes abertos de descoberta como VuFind, utilizando para isto os dados de *checkout* transacional de biblioteca. Para este estudo foram utilizados o *software* de recomendações Minrva e a ferramenta Weka. O estudo piloto reúne dados dos usuários de modo a gerar recomendações e, em seguida, analisa as interações do usuário com as recomendações a fim de avaliar o *software*.

Em número absoluto de aplicações observa-se uma tendência das pesquisas em bibliotecas, publicadas na última década, se dedicarem à conjuntos de dados textuais com a aplicação de ferramentas para o seu tratamento. Estas pesquisas representam 57% do total de publicações que compõem o *corpus* da RSL, em detrimento de conjuntos imagéticos ou audiovisuais, que somados, foram objeto de pesquisa em 14% das investigações.

No que se refere aos benefícios e impactos produzidos por aplicações do ML e relatados da literatura, Kim (2019, p. 40) destaca que todos os benefícios do ML se aplicam apenas a objetos digitalizados ou nato-digitais. A autora ainda afirma que

> a digitalização pode parecer menos interessante e inovadora em comparação com a aplicação de técnicas de ML. Mas o primeiro é de fato, um pré-requisito para o último. Bibliotecas e arquivos devem continuar a expandir seus esforços de digitalização, se eles planejam fazer uso de técnicas de ML para melhor curar e preservar suas coleções. (KIM, 2019, p. 40)

Desta forma, a criação de grandes silos de dados digitalizados com descrições inexistentes ou superficiais tornam esses objetos digitais com pouca ou nenhuma possibilidade de descoberta e recuperação, haja vista as instituições disporem de pouco recurso laboral para a produção manual de metadados. E, devido à falta de mão de obra, muitos arquivos e coleções

especiais em bibliotecas têm grandes acúmulos de materiais não processados (KIM, 2019, p. 39). Neste contexto, segundo Kim (2019), o Ml pode avançar na descoberta, navegação e agrupamento desses materiais gerando metadados mais detalhados.

Como consequência, os fluxos de trabalho do profissional da informação são remodelados, à medida em que novas tecnologias são incorporadas aos processos de trabalho. Nesta circunstância, Graser e Burel (2018) afirmaram que a aplicação da automação em fluxos de trabalho indica uma mudança da função tradicional de criação de metadados para a coleta e o gerenciamento de metadados, tornando os objetos digitais acessíveis, detectáveis e utilizáveis.

Além disso, de acordo com Graser e Burel (2018, p. 11), a automação oferece uma excelente oportunidade para profissionais da biblioteca criarem grandes quantidades de metadados robustos e com qualidade de modo a auxiliar na acessibilidade das coleções da biblioteca.

Dentre as implicações práticas, Kragelj e Borštnar (2020) afirmaram que modelos de classificação desenvolvidos com ML podem fornecer recomendações aos bibliotecários durante o trabalho de classificação e, além disso, podem ser implementados como um complemento à pesquisa de texto completo nos bancos de dados da biblioteca. Ainda de acordo com os autores, em relação às implicações sociais, a classificação automática auxilia o bibliotecário a economizar tempo em sua tarefa diária de classificação e possibilita tornar disponível o conhecimento produzido em textos antigos.

No que se refere a atividade de descarte, Wagstaff e Liu (2018, p. 246), afirmaram que o principal obstáculo nestes projetos é a falta de tempo, contudo a aplicação de métodos de classificação automática pode reduzir a quantidade de esforço humano nesta atividade.

Quando ao enriquecimento de metadados gerados em escala via métodos de ML, Kim (2021, p. 40) afirma que depois de revisados e ampliados por catalogadores humanos, os metadados podem melhorar as opções de pesquisa em materiais audiovisuais, facilitando assim a navegação e a identificação de diversos elementos nestes materiais.

Ainda no que tange a geração automática de metadados, Boman (2019, p. 21) afirmou que um desafio específico é a melhoria da geração automática de metadados em bibliotecas, permitindo a catalogação e a indexação não apenas aumentar a velocidade de geração de metadados, mas também melhorar a profundidade e a amplitude dos termos do assunto.

Em relação a vinculação bidirecional entre catálogos de bibliotecas e enciclopédias livres (JOORABCHI; MAHDI, 2018; 2013), esta atividade pode enriquecer a qualidade dos artigos das enciclopédias livres e ao mesmo tempo, aumentar a visibilidade dos recursos da biblioteca, criando um novo fluxo de informações entre os usuários de bibliotecas e os usuários de enciclopédias livres. Desta forma, os pesquisadores Joorabchi e Mahdi (2018, p. 57) destacaram como vantagens da vinculação de coleções: a possibilidade de permitir a criação de um *link* bidirecional entre a Wikipédia e os catálogos das bibliotecas; e segundo, eliminar a necessidade de indexação de cada registro de biblioteca recém-criado individualmente.

A respeito dos impactos e benefícios do ML aplicados a processos que incluem os resultados de pesquisa, Cox, Pinfield e Rutter (2019, p. 422) em entrevista a 33 profissionais que trabalham com informação, destacaram em algumas percepções acerca da implementação de IA/ML colhidas no estudo que os sistemas de recomendação “atingiriam uma melhora e a execução em segundo plano poderia até mesmo substituir a necessidade de pesquisar, porque esses sistemas “anteciparam as necessidades”, e desta forma os resultados da pesquisa chegariam de forma muito mais proativa para ao usuário. Um outro ponto destacado é que essas recomendações também podem ser altamente personalizadas.

Ainda em relação aos sistemas de recomendação baseado em ML, Hahn e McDonald (2018, p. 70) ressaltaram que as

> recomendações personalizadas podem aumentar o acesso, uso e o impacto dos investimentos em conteúdo digital e coleções de pesquisa. Um “recomendador” resultará verdadeiramente útil se novos conjuntos de dados inexplorados forem extraídos e utilizados para fornecer recomendações futuras para usuários. As recomendações extraídas de forma inteligente oferecem novos *insights* sobre as necessidades de informação e os melhores recursos disponíveis em bibliotecas digitais. E, além disso, as interações do usuário por meio das recomendações fornecidas ajudarão o algoritmo a filtrar as preferências.

Além disso, ao extrair e explorar os dados gerados por bibliotecas os gestores conhecem melhor os seus usuários e adaptam os serviços de acordo com as suas necessidades.

Por fim, Robert Fox (2016) afirma que as bibliotecas podem contribuir para essa evolução usando a vasta gama de dados disponíveis. A prática tradicional de usar dados para descrever informações usando análise estruturada e vocabulários controlados pode ser aprimorar através do uso de algoritmos autônomos, de modo a atender efetivamente às necessidades públicas e acadêmicas por meio da aplicação de técnicas de aprendizado de máquina.

No que se refere às dificuldades compartilhadas nos estudos, os pesquisadores relataram problemas de escassez de conjuntos de dados de treinamento (KIM, 2021; KRAGELJ; BORŠTNAR, 2020; GOLUB; HAGELBÄCK, ARDÖ, 2020; JOORABCHI; MAHDI, 2018), o que pode provocar distorções nos resultados de pesquisa, devido a pouca diversidade de cobertura dos dados em quantidade satisfatória. Dificuldade relativas à etapa de pré-processamento de dados, ou seja, na etapa de limpeza de dados (KRAGELJ; BORŠTNAR, 2020, GOLUB; HAGELBÄCK, ARDÖ, 2020; HARPER, 2018), que demanda tempo e esforços dos pesquisadores. E, Kim (2021) ainda destaca a dificuldade em obter ou estabelecer a verdade fundamental.

Uma outra questão referente aos conjuntos de dados, foi relatada por Harper (2018) o qual afirma que os conjuntos de dados de treinamento de imagem precisam incluir representações fenotípicas e demográficas mais amplas. Em conjunto de dados textuais, Kragelj e Borštnar (2020) relataram que um dos principais desafios enfrentados na pesquisa foi a evolução da escrita o que levou os pesquisadores a buscarem dicionários com o objetivo de traduzir termos e expressões para uma linguagem mais atual e, desta forma, possibilitar um melhor desempenho dos algoritmos de classificação.

Em projetos de indexação automática Toepfer e Seifert (2018) relataram vários desafios que podem ser enfrentados no desenvolvimento desses projetos. O primeiro deles refere-se às restrições legais que podem limitar o uso de documentos em texto completo ou resumo de publicações, e desta forma, prejudicar o desempenho do projeto. Em segundo lugar, Toepfer e Seifert (2018)

destacaram que conjunto de dados com pouca uniformidade na distribuição de conceitos podem distorcer os dados de treinamento.

Graser e Burel (2018) afirmam que em projetos onde a recuperação de imagens é baseada em conteúdo, a lacuna semântica é um problema significativo. Graser e Burel (2018) ainda destacam que essa lacuna é a disparidade entre o conteúdo de baixo nível nas imagens e os conceitos semânticos de alto nível que elas podem representar para um usuário.

Outro obstáculo relatado é a disponibilidade limitada de profissionais para projetos de Ml em bibliotecas (KRAGELJ; BORŠTNAR, 2020) e a falta de competência técnica para o acompanhamento das atividades.

Por fim, Cox, Pinfield e Rutter (2019) destacam que os custos provavelmente serão uma barreira para a implementação de projetos que incluam IA no desenvolvimento. Uma outra questão que pode dificultar a implementação de projeto em bibliotecas é o domínio de provedores comerciais, que podem reduzir o papel das bibliotecas (COX; PINFIELD; RUTTER, 2019).

A seguir serão apresentadas as competências laborais do profissional da informação no contexto do aprendizado de máquina.

4.2.2 Competências do profissional da informação em bibliotecas frente ao uso de novas tecnologias

À luz da literatura científica que compõe está RSL, as competências do profissional da informação estão se modificando à medida em que novas habilidades são demandadas no contexto da 4ª Revolução industrial. Nesta perspectiva, é requerido um novo conjunto de conhecimentos e competências do profissional da informação de maneira a se adequar ao crescente cenário de inovação, a fim de responder às necessidades de informação de sua comunidade.

David Lankes citado por Ridley (2019, p. 36), alerta para uma nova divisão digital com “uma classe de pessoas que podem usar algoritmos e uma classe usada por algoritmos” e argumenta que “os bibliotecários precisam se tornar bem

versados nessas tecnologias e participar de seu desenvolvimento, e não simplesmente descartá-las". Ridley (2019, p. 36-37) acrescenta que o objetivo não é apenas demonstrar falhas onde elas existem, mas estar pronto para oferecer soluções. Soluções baseadas em nossos valores e nas comunidades que atendemos.

Ademais, Arlitsch e Newell (2017) e Boman (2019) afirmam que é importante que as bibliotecas preparem a sua própria equipe para esta nova realidade, e se tornem centros de educação continuada para as suas comunidades.

Bárbara Neves (2019, p. 15; 2021, p. 200), após análise da literatura cientifica internacional, afirmou que as maiores perspectivas da computação cognitiva, aqui tratada como aprendizado de máquina são

> a transformação do trabalho dos profissionais, a melhoria do processamento de informações, o apoio à pesquisa, técnicas do aprendizado de máquina incorporados aos sistemas de automação, atuação dos profissionais como especialistas capazes de intermediar os usuários com a IA e a apropriação das ferramentas e recursos da computação cognitiva.

Todos os elementos apresentados pela autora corroboram para a transformação dos fluxos de trabalho do bibliotecário, tendo em vista que a aplicação de técnicas de aprendizado de máquina poderá padronizar processos, eliminar ou reduzir atividades repetitivas, prever e se antecipar a demandas, facilitar a tomada de decisão e desta forma, transformar o papel do profissional da informação que passará a se dedicar-se, por exemplo, a processos de curadoria de dados. Essas mudanças, de acordo com Cox, Pinfield e Rutter (2019, p. 433) constituem a extensão das implicações do paradigma do que se poderia chamar a biblioteca inteligente e, elas representam um desafio significativo para o futuro posição dos profissionais da informação.

No que tange às bibliotecas digitais, ambientes propícios para a aplicação de técnicas de ML em suas coleções, Graser e Burel (2018, p. 9) destacaram que os bibliotecários de metadados

> estão sendo chamados a ter conhecimento e experiência não apenas em descrição, aplicação e gerenciamento de metadados em vários esquemas, mas também em XML, ferramentas de código aberto para limpeza de dados, XSLT, Python, OAI-PMH, MarcEdit e uma série de outras habilidades técnicas necessárias para gerenciar a automação de metadados.

Já em relação à curadoria de dados, Cox, Pinfield e Rutter (2019, p. 433) afirmam que os bibliotecários devem adquirir novas competências relacionadas a dados: aquisição e licenciamento, gerenciamento de dados, controle de qualidade, curadoria e administração. Arlitsch e Newell (2017) ressaltam que os profissionais da informação devem adquirir habilidades quantitativas e analíticas para aprender o valor do *big data* e como ele pode ser manipulado, visualizado e analisado. Em suma, os bibliotecários devem encontrar maneiras de fazer as máquinas trabalharem a seu favor.

Acerca da alfabetização de dados, tópico destacado em 11% das pesquisas que compõem o *corpus* da RSL, García-Febo (2019, p. 4) afirma que felizmente, os bibliotecários estão analisando a IA a partir de várias perspectivas. Alguns estão utilizando a IA para ensinar alfabetização informacional e habilidades de pensamento crítico de modo a ajudar os usuários a formularem perguntas para esses dispositivos e aprenderem a como avaliar as respostas. Cox, Pinfield e Rutter (2019) destacam que os bibliotecários também devem instruir o uso responsável da IA, procurando entender os funcionamentos de sua face algorítmica, como indivíduo que tem o dever de se apropriar da competência tecnológica, de modo a evitar negligenciar armadilhas e práticas não autorizada das leis. Desta forma, Ridley (2019, p. 39) afirma que assim como a alfabetização informacional fornece aos usuários habilidades e perspectivas para avaliar recursos, a alfabetização algorítmica é uma estratégia de “explicabilidade” que permite aos usuários navegar e utilizar ferramentas e serviços algorítmicos.

Neste contexto, é necessário, que o bibliotecário reflita acerca dos potenciais problemas éticos e de privacidade que as bibliotecas enfrentarão à medida que o aprendizado de máquina permeia a sociedade. Desta maneira, Ridley (2019, p. 39) afirma que o Instituto de Serviços de Museus e Bibliotecas (IMLS) e a Universidade Estadual de Montana estabeleceram a “consciência algorítmica” como uma “nova competência”, cujo objetivo é “encontrar transparência para a lógica invisível incorporada em nossas interações de *software*”. Além disso, é importante que o bibliotecário incorpore em suas práticas de alfabetização de dados, orientações sobre a proteção de dados pessoais, a privacidade e o respeito aos direitos autorais. Deste modo, as bibliotecas devem se posicionar como proponentes ativos para o

desenvolvimento de projetos que busquem capacitar os seus usuários em IA/ML à luz de uma perspectiva crítica.

À vista deste cenário e, com o objetivo de sintetizar as contribuições relatados na literatura, a figura 4 apresenta as principais competências laborais requeridas aos profissionais da informação no âmbito da inteligência artificial e do *machine learning*. É importante ressaltar que esses elementos não formam um rol taxativo, mas apenas refletem as contribuições do conjunto de documentos analisados.

Figura 5. Competências do profissional da informação em ambientes de ML/IA

Fonte: Elaborado pela autora (2022).

O primeiro elemento destacado é o pensamento crítico, onde é requerido do profissional da informação a habilidade analisar objetivamente fatos e situações, sob diferentes perspectivas, fundamentados em informações precisas e de qualidade. Concatenado a esta ideia, dois importantes fatores se destacam: o posicionamento ético, o qual deve ser pautada no respeito aos direitos humanos e na liberdade intelectual, e a consciência algorítmica, que permite compreender os efeitos, as funcionalidades, bem como as consequências das aplicações dos algoritmos em diversos âmbitos da vida.

Em seguida, o segundo elemento realça a importância das habilidades técnicas que compreendem conhecimentos básicos em linguagem de

programação, ferramentas, metodologias e técnicas que englobem a inteligência artificial e o aprendizado de máquina. Além disso, associados a este elemento temos a alfabetização de dados e a capacidade de gestão de dados. O primeiro inclui a habilidade de explorar, entender e se comunicar com dados (TABLEAU, 2022). Essa habilidade deve ser analisada sob duas perspectivas, a primeira como uma competência laboral e, segundo como uma prática a ser adotada e difundida por meio de projetos de transferência de conhecimento para usuários de bibliotecas. No que se refere à curadoria de dados, o profissional da informação deverá selecionar, coletar, organizar, processar, preservar e reutilizar os dados, por meio da aplicação de técnicas de IA/ML, com a finalidade de descobrir e recuperar informações em grandes conjuntos de dados.

Por fim, o pensamento analítico que de acordo com Bortoletti (2021) busca desenvolver a competência de analisar os dados e as informações de forma racional por meio do uso de técnicas de análise estatística. Outra característica do pensamento analítico é que ele compartimentaliza as ideias e problemas a fim de entender o todo (BORTOLETTI, 2021). Assim, para o pleno desenvolvimento desta competência é necessário ter raciocínio lógico e realizar uma análise aprofundada dos dados, a fim de, por exemplo identificar padrões e prever comportamentos.

A seguir serão analisados os elementos éticos que permearam as pesquisas analisadas no *corpus* documental.

4.2.3 Questões éticas geradas a partir do uso de novas tecnologias em bibliotecas com foco em AM

À medida em que a Inteligência artificial e o Aprendizado de máquina ganham terreno e se fundem aos mais diversos aspectos da sociedade, algoritmos e códigos de programação que alimentam sistemas automatizados, dominam cada vez mais os inúmeros segmentos da vida moderna.

Neste sentido, é necessário compreender e engajar a comunidade de bibliotecas acerca das implicações éticas no contexto de produção,

disseminação e preservação do conhecimento (KENNEDY, 2019, p. 3), haja vista que, de acordo com Broughton (2019, p. 598), agentes inteligentes não são criados espontaneamente, mas requerem algum grau de participação humana, e nenhum sistema de aprendizado de máquina pode evitar o uso de informações que foi em algum estágio processado por humanos.

Sob outra perspectiva, Broughton (2019, p. 597) afirma que embora existam algumas discussões no século XX sobre a possibilidade de ações morais – ou imorais – das máquinas, o campo realmente começou a emergir a partir do Simpósio AAAI de Ética das Máquinas de 2005, onde o problema é claramente discutido e nomeado por Anderson et al. (2005):

> Pesquisas anteriores sobre a relação entre tecnologia e ética concentraram-se amplamente no uso responsável e irresponsável da tecnologia por seres humanos, com algumas pessoas interessadas em como os seres humanos devem tratar as máquinas. Em todos os casos, apenas os seres humanos se engajaram no raciocínio ético. Acreditamos que chegou a hora de adicionar uma dimensão ética a pelo menos algumas máquinas. O reconhecimento das ramificações éticas do comportamento envolvendo máquinas, bem como os desenvolvimentos recentes e potenciais na autonomia das máquinas exigem isso. Exploramos essa dimensão por meio da investigação do que tem sido chamado de ética da máquina. (BROUGHTON, 2019, p. 597-598 apud ANDERSON et al, 2005)

Todavia, neste cenário as discussões identificadas no *corpus* da RSL se concentraram nas implicações éticas acerca da responsabilidade algorítmica que os humanos assumem por suas ações. Ayre e Craner (2018, p. 346) afirmam que como seres humanos, todos nós temos preconceitos implícitos. E à medida que construímos novos sistemas, estamos desenvolvendo-os à nossa própria imagem, ou seja, com preconceitos incorporados. Ridley (2019, p. 41) sugere que o papel do arquivista e do bibliotecário deve, de fato, incluir a responsabilidade algorítmica devido ao posicionamento central em suas práticas laborais.

Não obstante, Harper (2018, p. 7) destaca que o ML não remove a subjetividade e o erro humano da tomada de decisão e da previsão. Em vez disso, ele pode reforçar o comportamento humano de maneiras complexas e imprevisíveis por meio dos dados nos quais um modelo é treinado. Desta

maneira, é importante ressaltar que as máquinas não criam preconceitos, elas apenas herdam e replicam arquétipos culturais, sociais e/ou institucionais.

Discussões recentes sobre viés concentram-se no domínio da discriminação de gênero. Em estudo recente, Criado-Perez (2019) citado por Broughton (2019, p. 598), revela que os próprios dados são muitas vezes tendenciosos, porque a amostra é de alguma forma falha. A principal preocupação de Criado-Perez (2019) é com o desequilíbrio de gênero, e fica claro que uma perspectiva feminina é muitas vezes omitida, porque os dados são derivados de estudos que trataram apenas do sexo masculino. Harper (2018, p. 7) reforça a mesma preocupação e complementa que se os dados de treinamento se inclinarem para homens brancos, a precisão do seu modelo também será distorcida.

Portanto a ausência de diversidade em conjunto de dados de treinamento e em equipes de projetos de AI/ML, mesmo que realizados de maneira não intencional podem estimular vieses e sedimentar as preferências dos desenvolvedores. Em suma, Griffey (2019, p. 47) afirma que se os dados forem tendenciosos, contiverem maus exemplos de tomada de decisão ou forem simplesmente coletados de uma maneira que não represente o conjunto completo de problemas, o sistema produzirá resultados incorretos, não representativos ou ruins. Além disso, complementando essa discussão, Zhou (2018) citado por Ayre e Craner (2018, p. 344), enfatiza que alguns desses vieses programáticos podem ser incorporados aos *softwares* propositalmente enquanto outras vezes, os vieses encontram seu caminho nos algoritmos “acidentalmente”.

Ainda em relação ao viés, Caliskan et al. (2017) citado Broughton (2019, p. 599) afirma que não apenas qualquer incompletude ou distorção na amostra de dados é repassada para sistemas inteligentes, mas destaca que vieses semânticos também podem ser transferidos. À vista disso, Broughton (2019, p. 597) afirma que os fatores que exacerbam o viés semântico no campo da organização do conhecimento incluem:

> provisão desigual de terminologia ou (em um sistema codificado como uma classificação) distribuição desigual de notação; falha em nomear certos grupos ou perspectivas; e linguagem que

> tenha um forte sabor de uma perspectiva ou cultura favorecida em particular.

Broughton (2019, p. 599) complementa que a existência de tal viés semântico tem implicações consideráveis para recuperação da informação porque classificadores constroem estruturas baseados em *corpora* textuais na suposição de que estes apresentam um caráter neutro e objetivo do mundo.

No que se refere aos reflexos das distorções em conjuntos de dados pouco diversificados ou heterogêneos, esse cenário pode levar ao fenômeno denominado "filtro bolha[31]", o qual de acordo com Ayre e Craner (2018, p. 344), as pessoas são colocadas em um silo com pouca exposição a pontos de vista contrários. À vista disso, conjunto de dados pouco representativos associados a uma capacidade crítica de informação limitada apresenta-se como ambientes propícios para a reprodução de discursos de ódio. Johnson (2018, p. 15) afirma que a mesma lente crítica da alfabetização informacional deve ser aplicada à IA e, para isso, precisaremos de uma lente muito mais poderosa.

Em relação à transparência no desenvolvimento de algoritmos para a aplicação de técnicas de AI/ML, Ridley (2019) defende a importância da aplicação da inteligência artificial explicável (doravante XAI). O XAI, segundo Ridley (2019, p. 28), refere-se a um conjunto diversificado de estratégias, técnicas e processos que tornam os sistemas de IA interpretáveis e responsáveis. Além disso, Ridley (2019, p. 29) adiciona a expectativa de que os sistemas de IA "terão a capacidade de explicar sua lógica, caracterizar seus pontos fortes e fracos e transmitir uma compreensão de como eles se comportarão no futuro".

O XAI é composto por dois pilares: a confiança e a responsabilidade. Ademais, Ridley (2019) destaca que o XAI visa permitir que usuários humanos "compreendam, confiem adequadamente e gerenciem efetivamente a produção emergente de parceiros artificialmente inteligentes".

[31] De acordo com Cecatto (2020) O termo "filtro bolha" foi cunhado por Eli Pariser em 2011 e refere-se aos resultados dos algoritmos que ditam o conteúdo que chega até nós pela "web". Estes algoritmos criam um ambiente exclusivo de informações para cada usuário, alterando completamente a maneira com que as informações nos alcançam. São seleções de conteúdos personalizadas, baseadas no histórico de navegação, idade, sexo, localização e outros dados do usuário.

No âmbito das bibliotecas, Ridley (2019, p. 31) ressalta que

> à medida que as bibliotecas adquirem e desenvolvem cada vez mais sistemas e serviços algorítmicos de tomada de decisão em apoio às comunicações acadêmicas e à operação de bibliotecas, elas devem fazê-lo de uma maneira que insista na interpretabilidade e na explicação. Fazer menos do que isso é uma delegação inconsciente à tecnologia e uma ab-rogação do rigor acadêmico.

Além disso, Ridley (2019, p. 40) apresentou um exemplo interessante e instrutivo do papel do XAI em bibliotecas:

> a Springer Nature publicou um livro de acesso aberto denominado “AI: Lithium-Ion Batteries[32]: A Machine-Generated Summary of Current Research”. O autor, identificado como “Beta Writer”, categorizou e resumiu algoritmicamente mais de 150 principais publicações de pesquisa selecionadas entre aproximadamente 1.000 publicações entre 2016 e 2018, sintetizando assim um grande e complexo *corpus* da literatura de pesquisa atual sobre o tema. Os processos algorítmicos que criaram este livro, usaram uma combinação de várias ferramentas de processamento de linguagem natural (NLP) “de prateleira”, incluíram o pré-processamento dos documentos para tratar de várias normalizações linguísticas e semânticas; agrupar documentos por similaridade de conteúdo (ou seja, o conteúdo dos capítulos e seções do livro); gerar resumos, introduções e conclusões; e finalmente a saída do XML como um manuscrito completo.

Em face do exposto, Henning Schoenenberger (2019), destaca que o objetivo do projeto é “iniciar uma ampla discussão, junto com a comunidade de pesquisa e especialistas do domínio, sobre as futuras oportunidades, desafios e limitações desta tecnologia”.

Bibliotecas precisam considerar não apenas como os dados podem ser expostos em sistemas algorítmicos, mas também questões relativas às novas obrigações em relação à privacidade e reutilização de dados. Desta forma, essas instituições devem prezar pela proteção de dados pessoais, de forma a não fornecerem dados sensíveis ou rastrearem seus usuários. Todavia, é importante destacar que os gestores devem considerar que algoritmos presentes em sistemas informacionais podem ser capazes de reidentificar dados supostamente anônimos.

[32] Link para o texto completo: https://link.springer.com/book/10.1007/978-3-030-16800-1

Desta forma, à medida em que mais bibliotecas e fornecedores comerciais passam a desenvolver sistemas com IA e aprendizado de máquina, gerentes de biblioteca devem estar sensíveis às implicações de privacidade na coleta, manipulação e armazenamento de dados necessários para treinar e atualizar esses sistemas.

Neste sentido, bibliotecas que contratam fornecedores, precisam estar cientes dos mecanismos pelos quais esses dados são protegidos e como eles podem ser compartilhados por outras pessoas. Além disso, devem estar atentos ao sigilo comercial, as questões de propriedade intelectual e a privacidade de dados, elementos esses que não devem se sobrepor e tornarem-se justificativas para a pouca transparência nos processos de desenvolvimento de ferramentas e sistemas informacionais. Desta forma, é primordial que os gestores pressionem fornecedores por mais informações sobre algoritmos e códigos utilizados antes de adquirir novos produtos e serviços.

Neste ponto é fundamental a aplicação da “explicabilidade”, pois o fornecedor deverá apresentar explicações do produto de maneira holística e, desta forma, de acordo com Ridley (2019, p. 32), ele deverá ser capaz de abordar “como” (entradas, saídas, processo), “por que” (justificativa, motivação), “o quê” (consciência de que existe um sistema algorítmico de tomada de decisão) e o “objetivo” (design, manutenção) em projetos que apliquem a AI/ML. Além disso, Ridley (2019, p. 31) complementa que uma visão mais holística do processo incluiria explicações que considerassem os dados usados para treinamento e tomada de decisão, o ambiente computacional utilizado, o contexto do projeto e implantação algorítmica e os responsáveis por sua operação e uso (ou seja, uma análise).

Nesta perspectiva as discussões que envolvem as aplicações éticas em projetos de AI/ML em bibliotecas concentram-se à luz de quatro importantes aspectos: a responsabilidade algorítmica, a liberdade intelectual, a privacidade e o acesso à informação, elementos esses que devem adotados e aplicados por profissionais da informação em sua atuação diária.

Figura 6. Elementos decorrentes da aplicação da ética em sistemas baseados em AI/ML em bibliotecas.

Fonte: Elaborado pela autora (2022).

O primeiro elemento destacado é a responsabilidade algorítmica, que deve assegurar a execução de projetos auditáveis e transparentes, minimizando tomadas de decisões opacas e pouco fundamentadas.

Em relação à liberdade intelectual, Johnson (2018, p. 16) esclarece que essa expressão reflete o direito de cada indivíduo de buscar e receber informações sob todos os pontos de vista sem restrição. E, de acordo com a IFLA (2020, p. 4) ela pode abranger múltiplas dimensões, como a liberdade para formar e manter opiniões sem interferência, liberdade de expressão, acesso à informação, bem como autodeterminação individual mais ampla.

Já em relação à privacidade trata-se do princípio responsável por proteger e controlar o compartilhamento de informações pessoais em sistemas informatizados e, desta forma evitar o uso não autorizado de informações de modo a garantir o sigilo de dados pessoais. Os elementos correlatos a aplicação deste princípio são a governança de dados e a segurança da informação.

Por fim, o acesso à informação, direito fundamental expresso na Constituição federal e elemento basilar para o cumprimento da missão das

bibliotecas. O acesso à informação amplia o exercício da cidadania e confere subsídios para o pleno exercício de direitos e deveres do cidadão.

A reunião de todos esses princípios em conjunto com fundamentos aprovados por governos[33] e instituições de pesquisa, conferem ao profissional responsabilidade na tomada de decisão algorítmica de modo a não contribuírem para o desenvolvimento de sistemas opacos e que reforcem vieses sistêmicos.

A próxima seção abordará as principais técnicas, metodologias e ferramentas empregadas na construção de produtos e serviços de bibliotecas.

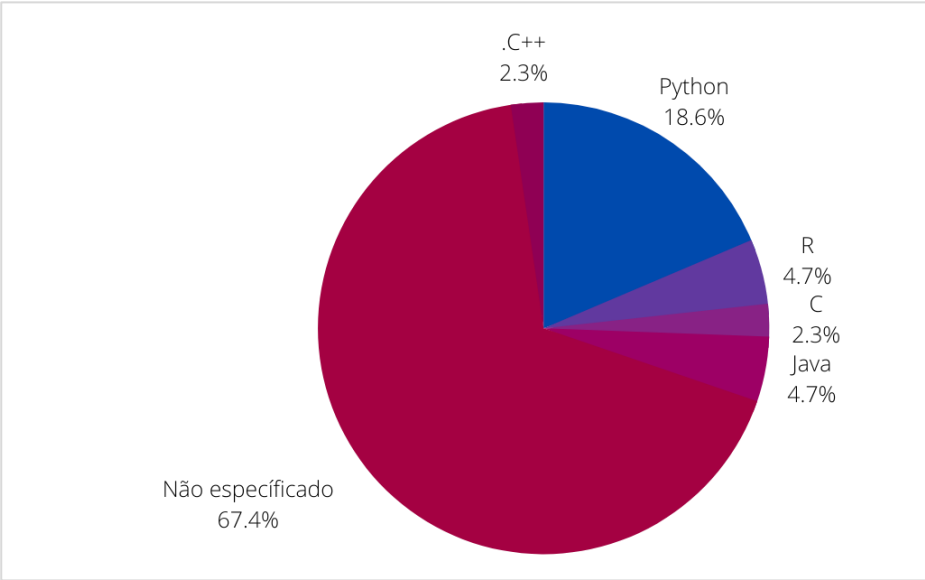
4.3 Técnicas, metodologias e ferramentas utilizadas em projetos de *machine learning* em Bibliotecas

Nesta seção serão apresentadas as principais técnicas aplicadas em projetos de aprendizado de máquina em bibliotecas com destaque para a descrição de linguagens de programação, algoritmos, ferramentas e métodos citados ao longo dos trabalhos que compõem a RSL.

O primeiro elemento analisado são as linguagens de programação que são um método padronizado utilizado para comunicar instruções para um computador e que empregam uma série de conjuntos de regras sintáticas e semânticas utilizados para definir um programa de computador (IFSC, 2020). À vista disso, foram mencionados no *corpus* da RSL cinco tipos de linguagens de programação: Python, R, C, C++ e Java. Ademais, cabe destacar que 67,4% das pesquisas não descreveram nenhuma linguagem de programação na construção de seus projetos.

[33] O Grupo de Especialistas de Alto Nível em Inteligência Artificial da Comissão Europeia estabeleceu sete fundamentos para alcançar uma inteligência artificial confiável. Os fundamentos são: agência e supervisão humana; robustez e segurança; privacidade e governança de dados; transparência; diversidade, não discriminação e justiça; bem-estar social e ambiental; e responsabilidade.

Gráfico 15. Linguagem de programação mencionadas no *corpus*



Fonte: Elaborado pela autora (2022).

A linguagem Python[34] foi a mais utilizada não apenas no desenvolvimento de aplicações, como também em projetos de alfabetização de dados em bibliotecas. Ela foi mencionada em 18,6% das pesquisas.

O Python foi criado no início dos anos 90 por Guido van Rossum no *Stichting Mathematisch Centrum* na Holanda como sucessor de uma linguagem denominada ABC (PYTHON SOFTWARE FOUNDATION, 2022b). Ele é caracterizado por Graser e Burel (2018, p. 2) como uma linguagem de *script* gratuita e orientada a objetos nos campos da análise e manipulação de dados. Além disso, o Python é considerado por Finley (2019, p. 11) como uma ótima ferramenta para programadores iniciantes visto que ela é mais fácil de aprender e executar do que outras linguagens de programação. Dentre as suas inúmeras aplicações, o Python foi utilizado por Harper (2016, p. 4) com o objetivo de analisar a origem dos dados da Digital Public Library of America (DPLA), além de contar as ocorrências de cada perfil de aplicação de metadados. Gomez et al (2020, p. 4) aplicaram o Python no desenvolvimento de um projeto para marcação automática de imagens com rótulos descritivos. Além disso, a

[34] Python: https://www.python.org/

linguagem foi utilizada no desenvolvimento de kits para alfabetização de dados na Biblioteca Pública de Frisco (FINLEY, 2019, p. 11).

Em seguida, destacam-se o uso das linguagens Java e R com respectivamente 4,7% das menções nas pesquisas. O Java[35] é uma linguagem de programação e plataforma de computador que surgiu em 1995 desenvolvida pela empresa Sun Microsystems (ORACLE, 2022a). Ela é orientada a objeto e elaborada de modo a permitir que desenvolvedores criem uma plataforma continua, além disso o Java possui versão gratuita tanto para uso pessoal como para uso profissional. Atualmente a empresa responsável pelo seu desenvolvimento é a Oracle Corporation que adquiriu a linguagem em 2008. Santana et al (2014, p. 180) utilizaram o Java na implementação de métodos de treinamento e classificação em conjunto com a linguagem C em um projeto cujo objetivo é combinar heurísticas especificas de domínio para remover a ambiguidade em registros de autores.

Já a linguagem R[36] é uma linguagem e ambiente para computação estatística e gráficos. Castro e Ferrari (2016, p. 348) afirmam que em virtude de sua origem estatística, o R foi adotado por diferentes grupos de pesquisadas que acabaram por criar pacotes de algoritmos para as mais diversas tarefas de mineração de dados. Desta forma, o R fornece uma ampla variedade de técnicas estatísticas tais como: modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento etc. (R FOUNDATION, 2022). Além disso, um de seus pontos fortes, é a facilidade com que gráficos podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário (R FOUNDATION, 2022). Walker e Jiang (2019, p. 207) aplicaram a linguagem R em todas as análises realizadas para prever a probabilidade de compra de títulos por meio de programas de aquisição orientada por demanda.

Por fim, foram citadas as linguagens C[37] e C[38]++ com 2,3% das menções cada. A linguagem C foi criada em 1970 pelo estadunidense Dennis Ritchie, por meio da evolução da linguagem B. A sua filosofia básica é que os programadores

[35] Java: https://www.java.com/pt-BR/
[36] R: https://www.r-project.org/
[37] C: https://www.iso.org/standard/74528.html
[38] C++: https://www.iso.org/standard/68564.html

devem estar cientes do que estão fazendo, ou seja, supõe-se que eles compreendam o que estão ordenando o computador fazer e, desta forma explicitem completamente as suas instruções (MARTINS, 2011, p. 1). Atualmente ela é padronizada por meio da ISO/IEC 9899/2018.

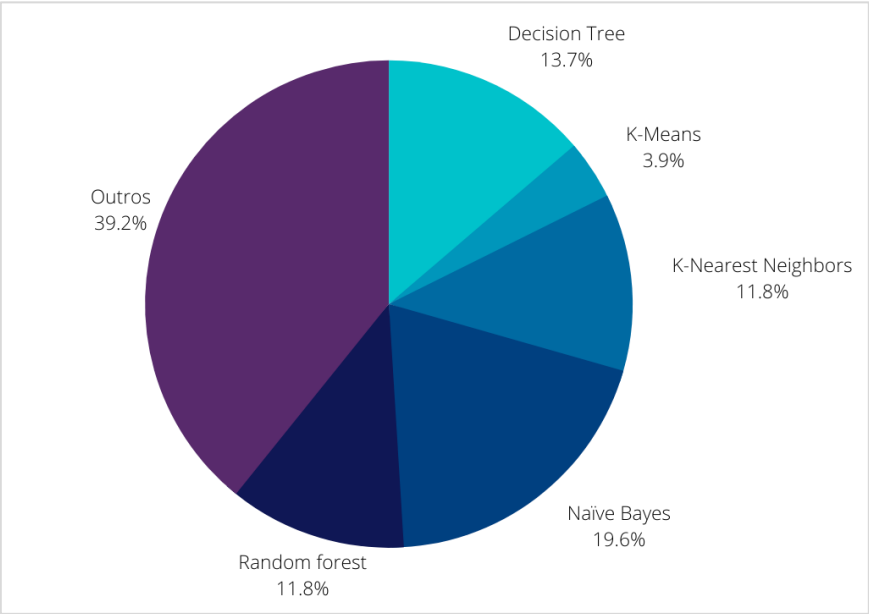
Já a linguagem C++ é uma linguagem de programação criada por volta de 1980 pelo dinamarquês Bjarne Stroustrup. Essa nova linguagem trouxe vantagens em relação à anterior, justamente porque surgiu da necessidade de corrigir algumas das limitações da linguagem C (CARDOSO, 2020). Iniciando o processo com a inserção de classes, o produto final desse projeto se tornou uma linguagem com suporte total à Programação Orientada a Objetos (POO). Além disso, ela também foi normalizada por meio da ISO/IEC 14882:2020. Walker e Jiang (2019, p. 207) utilizaram a linguagem C++ para a programação do pacote fast Adaboost na aplicação de análises preditivas na aquisição de obras orientadas por demandas.

No que tange à aplicação de algoritmos[39], estes são definidos pela Enciclopédia Britânica (2022) como um procedimento sistemático que produz, em um número finito de passos, uma resposta a uma pergunta ou solução. Os algoritmos de aprendizado de máquina são preditivos e, de acordo com Kraus-Friedberg (2019, p. 112), indica que eles usam padrões baseado em dados históricos para prever ações ou necessidades futuras.

No conjunto de pesquisas foram citados mais de 20 diferentes tipos de algoritmos, no entanto foram selecionados 5 para um detalhamento daqueles com o maior número de menções na literatura, a saber: *Random Forest, Decision Tree, K-means, k-Nearest Neighbors* e *Naïve* Bayes, destacados no gráfico a seguir:

[39] O termo algoritmo é derivado da tradução latina de "*Algoritmi de numero indorum*", do tratado de aritmética do matemático mulçumano Al-Khwarizmi do século IX "Al-Khwarizmi Concerning the Hindu Art of Reckoning".

Gráfico 16. Algoritmos mencionadas no *corpus*



Fonte: Elaborado pela autora (2022).

É importante destacar ainda que a maioria das pesquisas buscaram avaliar o desempenho de dois ou mais algoritmos e métodos de classificação elencados nesta seção, para o desenvolvimento de projetos.

Nesta perspectiva, destacaram-se os trabalhos de Kragelj e Borštnar (2020) em processos de classificação de textos utilizando a CDU. Golub, Hagelbäck e Ardö (2020) aplicaram seis algoritmos com o objetivo de analisar o desempenho da atribuição da CDD em coleções de objetos digitais suecos. Wagstaff e Liu (2018) aplicaram classificadores para o descarte de coleções. Santana et al (2014), Ferreira et al (2014) e Veloso et al (2012) utilizaram os algoritmos em projeto de remoção da ambiguidade em nomes de autoridade. Taniguchi (2012) empregou classificadores na detecção de registros bibliográficos duplicados. E, por fim, Joorabchi e Mahdi (2013) aplicaram algoritmos na classificação de publicações cientificas. Ademais, é importante destacar que 32% das pesquisas que compõe o corpus, descreveram a aplicação de algum algoritmo classificador em seus projetos.

À vista deste cenário, 19,6% dos trabalhos são provenientes de métodos probabilísticos o qual lidam com tarefas preditivas (FACELI et al., 2019, p. 73), por meio da aplicação de algoritmos *Naïve Bayes*. Os algoritmos *Naïve Bayes*

são classificadores estatísticos fundamentados no Teorema de Bayes e usados para predizer a probabilidade de pertinência de um objeto a determinada classe (CASTRO; FERRARI, 2016, p. 186). Eles possuem desempenho comparável a redes neurais artificias e árvores de decisão para alguns problemas, além disso apresentam alta acurácia e velocidade de processamento quando aplicados a grandes bases de dados (CASTRO; FERRARI, 2016, p. 186).

Wagstaff e Liu (2018, p. 240) afirmam que os seus pontos fortes são: a probabilidade fundacional, por isso, naturalmente, fornece uma probabilidade posterior para cada previsão que é realizada; a possibilidade de aceitar entradas numéricas ou categóricas; e, por fim, não exige a especificação de parâmetros.

Já em relação às suas fraquezas, Wagstaff e Liu (2018, p. 240) destacam que para que as suas previsões se generalizem, a distribuição de classes em conjuntos de dados de treinamento deve ser consistente com as probabilidades verdadeiras de essas classes serem observadas em novos dados. Além disso, Oguri (2007, p. 25) afirma que este classificador é denominado ingênuo (naive) por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro.

Os classificadores Naïve Bayes são, de acordo com Taulli (2020, p. 81), comumente utilizados em análise de texto para a detecção de *spam* por *e-mail*, segmentação de clientes, análise de sentimentos, diagnóstico médico e previsões meteorológicas. A razão para esses usos é que essa abordagem é útil na classificação de dados com base em características-chave e padrões.

Em seguida, com 13,7% notabilizou-se o uso de algoritmos baseado em métodos simbólicos, segundo Faceli et al (2022), denominados *Decision Tree* ou árvore de decisão. Em métodos simbólicos a representação do conhecimento extraído dos dados pode ser realizada por meio de estruturas simbólicas, que possibilitam uma interpretação mais direta por seres humanos (FACELI et al, 2022). Castro e Ferrari (2016, p. 170) definem a árvore de decisão como uma

> estrutura em forma de árvore na qual cada nó interno corresponde a um teste de um atributo, cada ramo representa um resultado do teste e os nós folhas representam classes ou distribuições de classes. O nó mais elevado da árvore é conhecido como nó raiz, e cada caminho da raiz até um nó folha corresponde a uma regra de classificação.

Uma árvore de decisão procura resolver problemas em um espaço possível de soluções e, desta forma, ela produz uma série de testes organizados hierarquicamente. À vista disso, Wagstaff e Liu (2018, p. 240) apresentam como pontos fortes em sua aplicação: a geração de um modelo fácil de entender que possa explicar como cada predição foi feita através do simples esboço de uma árvore, e a possibilidade deste algoritmo aceitar entradas numéricas ou categóricas.

Em relação a sua principal fraqueza, Wagstaff e Liu (2018, p. 240) afirmam que as suas estimativas de probabilidade posterior geralmente não são muito confiáveis porque eles são frequentemente calculados a partir da fração de exemplos que atingem um dado nó folha, que podem ser uma amostra muito pequena.

Logo após, destaca-se o uso do algoritmo baseado em distância, de acordo com Faceli et al (2019, p. 61), k-Nearest Neighbors (k-NN) ou “k-Vizinhos mais próximos” com 11,8% em aplicações, que buscam analisar a proximidade entre os dados na realização de predições. Segundo Faceli et al (2019, p. 63) representa um dos paradigmas mais conhecidos do aprendizado de máquina indutivo: objetos com características semelhantes pertencem ao mesmo grupo. O k-NN é um algoritmo simples de calcular e, de acordo com Taulli (2020, p. 83), e ele é conhecido por ser um algoritmo de aprendizagem preguiçosa porque não há nenhum processo de treinamento com os dados.

Dentre os pontos fortes do k-NN, de acordo com Wagstaff e Liu (2018, p. 240) destacam-se a rapidez em sua construção, já que nenhum modelo explícito precisa ser treinado; a ausência de suposições sobre a distribuição de classes no espaço de características; e a facilidade de as previsões exibirem os k exemplos que foram usados para rotular o novo item.

Já em relação a sua maior fraqueza Wagstaff e Liu (2018, p. 240) destacam que o tempo necessário para classificar um novo item aumenta com o tamanho do conjunto de dados de treinamento, uma vez que todos os exemplos devem ser considerados para encontrar os k exemplos mais semelhantes.

Logo após com 11,8% de aplicação em projetos, destacou-se o uso do algoritmo *Radom Forest* ou florestas aleatórias, que de acordo com Faceli et al

(2022) é baseado no método de injeção de aleatoriedade. Desta forma, o modelo gera várias árvores de decisão, cujas previsões são combinadas por votação uniforme (FACELI et al, 2022). O algoritmo usa a amostragem de exemplos com reposição do *bagging* combinada com a seleção aleatória de atributos (FACELI et al, 2022). Neste contexto, Wagstaff e Liu (2018, p. 240) afirmam que as decisões coletivas feitas por meio de uma floresta aleatória são mais confiáveis do que aquelas feitas através de uma árvore de decisão única (o efeito conjunto).

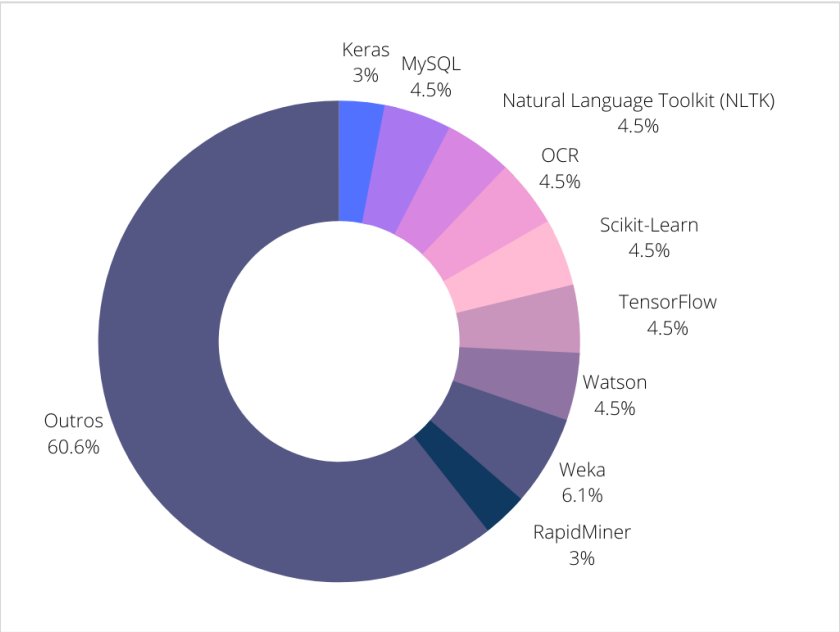
Dentre os pontos fortes na aplicação de algoritmos *Radom Forest* verifica-se que as decisões são mais robustas devido à natureza do seu conjunto, o que pode gerar uma confiança posterior nos valores (WAGSTAFF; LIU, 2018, p. 240). Já em relação a sua principal fraqueza, Wagstaff e Liu (2018, p. 240) destacam o fato dele ser composto por muitos modelos o que pode tornar a sua interpretabilidade diminuída em comparação com uma única árvore de decisão.

Por fim, o algoritmo K-means ou k-médias se sobressaiu ao ser aplicado em 3,9% do conjunto de pesquisas. Faceli et al (2019, p. 213) categoriza o k-means como um algoritmo particional baseado em erro quadrático. E, de acordo com Taulli (2020, p. 89) o K-means é eficiente em grandes conjuntos, pois coloca dados semelhantes não rotulados em diferentes grupos. Todavia, de acordo com Faceli et al (2022, online) ele não apresenta um bom desempenho em conjunto de dados ruidosos.

À vista deste cenário, Golub (2019, p. 111) afirma que o primeiro passo para a aplicação do K-means é criar aleatoriamente um número k de clusters e, em seguida, novos documentos são adicionados aos diferentes *clusters* com base em sua similaridade. Desta forma, o algoritmo particiona o conjunto de dados em k *clusters*, em que o valor de k é fornecido pelo usuário (DUDA et al., 2001 apud FACELI et al, 2022).

No que tange ao uso de ferramentas nos projetos de AI/Ml em bibliotecas, foi verificado a aplicação de mais de 40 tipos de ferramentas, aplicadas em diversas etapas dos projetos. Todavia, 9 ferramentas se destacaram com 41,8% de aplicações, e desta forma foram selecionadas para um maior detalhamento de seus usos.

Gráfico 17. Ferramentas mencionadas no *corpus* da RSL



Fonte: Elaborado pela autora (2022).

No conjunto das 9 (nove) ferramentas selecionadas, sete apoiam pesquisas nas áreas do *Machine Learning* e da Inteligência artificial, a saber: Weka, TensorFlow, Keras, Watson, RapidMiner, Natural Language Toolkit (NLTK) e Scikit-Learn.

O primeiro deles em número de menções é o Weka[40], presente em 6,1% das pesquisas, um *software* de código aberto com uma coleção abrangente de algoritmos de aprendizado de máquina para tarefas de mineração de dados (UNIVERSIDADE WAIKATO, 2022). De acordo com Castro e Ferrari (2016, p. 348) o Weka foi desenvolvido em Java e mantido pela Universidade de Waikato. O *software* possui interface gráfica que permite ao usuário, segundo Castro e Ferrari (2016, p. 348), realizar tarefas de pré-processamento, classificação, regressão, agrupamento e visualização dos dados, assim como planejar e executar análises ou experimentos mais complexos por meio da construção de fluxogramas que encadeiam as tarefas de mineração de dados.

O Weka foi aplicado por Taniguchi (2012, p. 165) em projeto de detecção de registros duplicados na etapa de implementação de algoritmos de

[40] Weka - https://www.cs.waikato.ac.nz/~ml/weka/

aprendizado. Joorabchi e Mahdi (2018; 2013) utilizaram o Weka em processos de vinculação de registros de bibliotecas e artigos do Wikipédia para determinar a eficácia de cada uma das características definidas para conceitos candidatos no projeto. Hahn e McDonald (2018, p. 72) aplicou o Weka na execução de processo de aprendizado de máquina off-line em projeto para a implementação de um recomendador em ambiente de descoberta.

Logo em seguida, com 4,5% das aplicações em projetos, destacou-se o uso do TensorFlow[41]. O TensorFlow é uma plataforma de código aberto para criar modelos de aprendizado de máquina (GOOGLE, 2022a). Ela apresenta um ecossistema abrangente e flexível de ferramentas, bibliotecas e recursos da comunidade que permite aos pesquisadores criarem e implantarem aplicativos com ML (GOOGLE, 2022a). O TensorFlow foi aplicado em grande parte das pesquisas em conjunto com o Keras.

O Keras[42] é um aplicativo de rede neural desenvolvido em Python e executada na plataforma de aprendizado de máquina TensorFlow (GOOGLE, 2022b). E, de acordo com a Microsoft (2022a) ele é capaz de executar redes neurais em alto nível junto a outras estruturas de DNN populares para simplificar o desenvolvimento.

À vista disso, o TensorFlow foi empregado em conjunto com o Keras na etapa de aplicação do método Latent Dirichlet Allocation (LDA) para a geração automática de metadados na coleção de *e-books* do Projeto Gutemberg desenvolvido por Boman (2019, p. 24). Além disso, Kim (2019) destacou a disponibilização do TensorFlow e do Keras, como ferramentas populares para o desenvolvimento de *frameworks* em projetos AI/ML, no Laboratório AI Lab para o uso de pesquisadores da Universidade de Rhode Island.

Ainda nesta categoria destaca-se as discussões em torno do uso do Watson[43] em 4,5% dos projetos. O Watson é uma plataforma *multicloud* que permite automatizar o ciclo de vida da Inteligência artificial e que possui a tecnologia *Machine Learning* em sua composição (IBM, 2022). De acordo com Fox (2016, p. 66) o Watson possui duas funções principais:

[41] TensorFlow: https://www.tensorflow.org/?hl=pt-br
[42] Keras: https://keras.io/
[43] Watson - https://www.ibm.com/br-pt/watson

> A primeira refere-se ao fato dele realizar o processamento de linguagem natural com um alto grau de precisão, para que possam ser feitas perguntas que correspondam ao modelo linguístico do pesquisador. Em segundo lugar, o Watson pode processar dados não estruturados para revelar *insights*.

Neste sentindo, o Watson foi projetado para consumir informações, processá-las de forma abrangente e depois interpretá-las (Shah, 2011 apud FOX, 2016, p. 66). E, Fox (2016, p. 66) complementa que as informações referenciadas não são metadados, mas artigos em texto completo, *feeds* de notícias, relatórios, mídias sociais e outras formas de dados institucionais.

À luz dessa perspectiva, Neves (2019, p. 7) relatou em seu artigo um projeto na Universidade da Bahia cujo objetivo é implementar uma solução de teste, por meio da aplicação do Watson, como proposta para apoiar o atendimento do bibliotecário de referência.

Outras duas pesquisas realizaram apenas relatos da importância da aplicação do Watson em ambientes de bibliotecas: Cox, Pinfield e Rutter (2019) e Fox (2016) já relatado. Cox, Pinfield e Rutter (2019) abordaram discussões a partir de relatos colhidos em entrevistas a profissionais da informação e, afirmam que o Watson está sendo incorporado em ambientes virtuais de aprendizado de modo a fornecer um tipo de experiência de aprendizado adaptativo.

Com aplicações voltadas para o campo da linguística computacional, a plataforma Natural Language Toolkit (NTKL) [44], obteve 4,5% das menções. Essa plataforma escrita em Python e distribuída sob a licença de código aberto, se dedica ao processamento de textos para classificação, tokenização, lematização, marcação, análise e raciocínio semântico de acordo com o *site* oficial (NLTK PROJECT, 2022). Esse kit de ferramentas foi originalmente desenvolvido em 2001 como parte do curso de linguística computacional da Universidade da Pensilvânia (NLTK PROJECT, 2022), com quatro objetivos principais: a simplicidade, a consistência, a extensibilidade e a modularidade de suas aplicações.

Kragelj e Borštnar (2020, p. 7) utilizaram o NLTK em conjunto com o Python nas etapas de pré-processamento de dados que inclui a limpeza, a

[44] Natural Language Toolkit - https://www.nltk.org/

lematização e a remoção de *stop words* no projeto para a classificação automática de texto antigos. Boman (2019, p. 21) também utilizou o NLTK em etapa de preparação e limpeza de dados.

Aplicado em projetos de Ciência de dados, o RapidMiner[45] foi citado em 3% dos projetos de ML/AI em Bibliotecas. Essa ferramenta foi desenvolvida pela Universidade de Dortmund em 2001, e trata-se de um *software* que atua em processos de mineração de dados e possui versões gratuitas e pagas (CASTRO; FERRARI, 2016, p. 349). Ademais, os autores (2016, p. 349) complementam que o RapidMiner permite a construção visual, por meio de blocos e fluxogramas, de processos complexos de análise e mineração de dados, podendo conectar-se a diferentes fontes de dados, tais como arquivos e diferentes SGBDs.

Lamba e Madhusudhan (2019, p. 171) aplicaram o RapidMiner em etapa de análise de previsão, o que inclui o pré-processamento dos documentos, a divisão em subconjuntos, o treinamento da base de teste, a aplicação de classificadores e a avaliação de desempenho do modelo. Já Kovacevic et al (2011, p. 388) aplicou o RapidMiner em experimento de classificação para avaliar modelos de algoritmos em projeto de extração automática de metadados em publicações científicas.

Por fim, os projetos de pesquisa integrantes do *corpus* da RSL mencionaram o uso da biblioteca de aprendizado de máquina Scikit-learn[46]. O Scikit-learn é módulo Python que integra uma gama de algoritmos de aprendizado de máquina para problemas supervisionados e não supervisionados (PEDREGOSA et al, 2011, p. 2826). O projeto foi iniciado em 2007 no *Google Summer of Code* por David Cournapeau (SCIKIT-LEARN, 2022).

O Scikit-learn foi aplicada em pesquisa desenvolvida por Harper (2016, p. 16), em conjunto com o Pandas e vários algoritmos de classificação para tentar prever a variável de destino da análise binária em projeto na Digital Public Libray of America (DPLA). Toepfer e Seifert (2018) e Wagstaff e Liu (2018) aplicaram o

[45] RapidMiner: https://rapidminer.com/
[46] Scikit-learn: https://scikit-learn.org/stable/

Scikit-learn para a implementação de algoritmos preditivos respectivamente em projetos de indexação automática e descarte de obras em bibliotecas.

Não menos importante, destacou-se a aplicação das ferramentas de apoio aos projetos OCR e MySQL. Essas ferramentas não possuem como objetivo principal a aplicação de técnicas avançadas de ML/AI, mas auxiliam em etapas de limpeza e pré-processamento de dados.

As ferramentas de reconhecimento óptico de caracteres popularmente conhecidas através de seu acrônimo OCR, “são capazes de extrair de forma completamente automática o texto (impresso ou manuscrito) contido em uma imagem digital” (OLIVEIRA et al, 2014, p. 1) e, desta forma tornando essa informação editável e pesquisável. Apesar de serem consideradas APIs simples, tratam-se de importantes ferramentas no contexto de bibliotecas, haja vista o grande volume de dados que podem ser produzidos por meio de processos de digitalização e posteriormente convertidos em formato de texto por meio do uso desta ferramenta.

Kim (2021) destacou a aplicação do OCR em processos de conversão de conteúdo de material digitalizado, o que de acordo com Cordell (2021) citado por Kim (2021, p. 39), resultou em aumento significativamente no acesso a coleções em grande escala. Taniguchi (2012) aplicou o OCR para converter a folha de rosto e o seu verso com o objetivo de utilizá-los em processos de detecção de registros duplicados no catálogo de bibliotecas.

Por fim, no rol de ferramentas com expressivo destaque na literatura o MySQL[47], que trata-se de um servidor robusto de banco de dados que utiliza linguagem SQL (*Structured Query Language*) (ORACLE, 2010, p. 1). O MySQL é caracterizado como rápido, multi-tarefa e multi-usuário. Em relação às licenças de acesso, a ferramenta possui versão gratuita e licença comercial (2010, p. 16), ambas gerenciadas pela Oracle Corporation. E, de acordo com a Oracle (2022b), o MySQL é o banco de dados de código aberto mais popular do mundo, aplicado aos mais diversos tipos de negócios. O MySQL foi utilizado por Golub, Hagelbäck e Ardö (2020, p. 24) na etapa de estruturação e organização de dados.

[47] MySQL: https://www.mysql.com/

No que tange a aplicação de métodos, foram computados aproximadamente o uso de 30 (trinta) métodos aplicados em projetos para o desenvolvimento de serviços e produtos em bibliotecas. É importante destacar que os pesquisadores aplicaram simultaneamente duas ou mais soluções em etapas de suas pesquisas. E devido à natureza heterogênea das aplicações, esses métodos e técnicas foram reunidos em conjuntos que compartilham características e finalidades em comum. Neste âmbito, destacam-se a aplicação de análises preditivas, técnicas estatísticas, redes neurais, *clustering*, mineração de dados e de texto e processamento de linguagem natural, detalhados a seguir.

A predição, segundo Castro e Ferrari (2016, p. 8) é uma terminologia utilizada para se referir à construção e ao uso de um modelo de modo a avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de um dado objeto.

No contexto de bibliotecas a aplicação da análise preditiva, segundo Litsey e Mauldin (2018, p. 141), pode auxiliar na compreensão das trocas de nível de transação e fornecer modelagem da experiência do cliente com a biblioteca. Desta forma, Taylor (2011) citado por Litsey e Mauldin (2018, p. 141) destaque que

> a análise preditiva funciona melhor ao prever três configurações. Primeiro, o desenvolvedor pode se basear no comportamento anterior para prever riscos futuros. Segundo, pode utilizar as interações passadas para prever crescimento futuro. Finalmente, pode ser usado na análise de fraudes passadas para prever a exposição futura à fraude.

Para a aplicação de métodos de predição foram aplicados algoritmos K-NN, Naïve Bayes e Árvore de decisão, descritos anteriormente. Destaca-se ainda a aplicação em 25% das pesquisas de técnicas de Support Vector Machines (SVM).

Support Vector Machines (SVMs) ou Máquina de vetores de suporte, de acordo com Lorena e Carvalho (2007, p. 43), trata-se de uma teoria que busca estabelecer uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida a partir da capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

Wagstaff e Liu (2018, p. 240) identificaram como pontos fortes do SVM a eficiência computacional especialmente para grandes conjuntos de dados e o bom desempenho em generalizações. Já em relação a sua principal franqueza, Wagstaff e Liu (2018, p. 240) destacaram a falta de interpretabilidade para as previsões, deste modo, as SVMs podem converter-se em caixas pretas que geram previsões sem qualquer explicação ou justificativa.

Neste cenário, a análise preditiva foi aplicada por Litsey e Mauldin (2018, p. 140) para compreender proativamente o comportamento da biblioteca. Além disso, destaca-se ainda a criação da primeira ferramenta de análise preditiva desenvolvida com técnica de Ml, de acordo com Litsey e Mauldin (2018, p. 142), denominada *Automated Library Information Exchange Network* (doravante ALIEN), com o objetivo de analisar, recomendar e prever serviços e ações que podem ajudar a melhorar a função geral da biblioteca.

Em relação a aplicação de técnicas estatísticas, tratam-se de importantes procedimentos que visam explorar conjuntos de dados. Neste contexto, a estatística descritiva, uma subárea da estatística destacada por Faceli et al (2022), resume de forma quantitativa as principais características de um conjunto de dados. No conjunto de pesquisas que compõe do *corpus* destacaram-se a aplicação de medidas de frequência, redes bayesianas, estimativa de densidade por Kernel e o TF-IDF.

As medidas de frequência medem a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados (FACELI et al, 2022). Já as redes bayesianas tratam-se de um tipo de modelo gráfico probabilístico que usa inferência bayesiana para cálculos de probabilidade. Em relação ao F-measure indica uma medida da precisão de um teste. A estimativa de densidade por Kernel (EDK) calcula uma forma não-paramétrica para estimar a função densidade de probabilidade de uma variável aleatória. E, por fim, sobressaiu a aplicação do TF-IDF, abreviatura para Frequência do termo-frequência inversa dos documentos, utilizada em 14% dos projetos de pesquisa.

Neste contexto, Faceli et al (2022) citando Salton et al (1975) afirma que o peso do termo é igual ao produto da sua frequência pela frequência inversa dos documentos. Neste âmbito, Kiam e Chhinnasamy (2013, p. 4) esclarecem que o TF mede a frequência de um termo em um documento enquanto o IDF

mede a raridade do termo em toda a coleção. À vista disso, Kragelj e Borštnar (2020, p. 9) utilizaram o TF-IDF para modelar a matriz de vetores de palavras que aparecem em textos.

No domínio das redes neurais artificiais, Castro e Ferrari (2016, p. 320) esclarecem que tratam-se de ferramentas de processamento de informações, inspiradas na operação do sistema nervoso humano que fazem uso de um processamento massivamente paralelo e distribuído de informação o que lhes conferem grandes capacidades de realizar mapeamentos não lineares de elevada complexidade.

De acordo com Yelton (2019, p. 11) durante o treinamento, a rede neural recebe registros do conjunto de dados de treinamento, um de cada vez. Para cada registro, ele compara a saída final da rede com algum tipo de valor esperado. Yelton (2019, p. 11) complementa a explicação com um exemplo: se as entradas forem fotografias, a saída pode ser uma decisão binária: “gato” ou “não é um gato”.

> A rede neural então avalia o quão errado estava e atualiza um pouco todos os parâmetros de todas as suas funções, em qualquer direção que a torne menos “errada”. Com o tempo, à medida que ela treina em um grande número de registros, a rede neural se torna cada vez mais precisa.

Como exemplo de aplicação, destaca-se o trabalho de Andromeda Yelton (2019) no treinamento de uma rede neural denominada HAMLET, cujo objetivo é explorar as teses produzidas no programa de pós-graduação do MIT.

No que tange a aplicação de métodos de agrupamento ou *clustering*, Castro e Ferrari (2016, p. 8-9) definem esse método como o processo de separar (particionar ou segmentar) um conjunto de objetos em grupos (do inglês *clusters*) de objetos similares. De acordo com Kragelj e Borštnar (2020, p. 8) trata-se de uma técnica não supervisionada, no qual um algoritmo busca semelhanças em um conjunto de dados sem que o supervisor atribua ou desconsidere rótulos. Desta forma, o *clustering*, por exemplo, agrupa em conjuntos por determinada característica uma *corpora* de documentos.

Castro e Ferrari (2016, p. 9) esclarecem que diferentemente da tarefa de classificação, o agrupamento de dados considera dados de entrada não

rotulados, ou seja, o grupo (classe) ao qual cada dado de entrada (objeto) não pertence não é conhecido a *priori*. Neste cenário, essa técnica foi aplicada ou discutida em 22% dos trabalhos.

À vista disso, destaca-se a aplicação de *clusters* no conjunto de trabalhos de Ferreira et al (2014) e Santana et al (2014) para a identificação de *clusters* de dados mais representativos selecionados para servir como dados de treinamento para a terceira etapa de atribuição de autor supervisionado. Giblin et al (2019) também aplicou *clusters* em seu trabalho, todavia o objetivo era identificar licenças regidas pelas mesmas regras em relação ao preço, na investigação sobre licenças de *ebooks* na plataforma Overdrive em países de língua inglesa.

Outra metodologia empregada nos trabalhos é a mineração de dados, expressão esta que, de acordo com Castro e Ferrari (2016, p. 4), foi cunhada em alusão ao processo de mineração de minerais valiosos, uma vez que explora uma base de dados (mina) utilizando algoritmos (ferramentas) adequados para obter conhecimento (minerais preciosos). Faceli et al (2019, p. 331) complementa que a mineração de dados consiste em extrair ou "minerar" conhecimento a partir de grandes quantidades de dados.

Smiraglia e Xin (2017, p. 216) destacam que "mineração de dados", é muitas vezes é confundida com o aprendizado de máquina. No entanto, os algoritmos de aprendizado de máquina não são usados apenas para resumir os dados e descobrir padrões ocultos como ocorre na mineração de dados, mas também podem servir como ferramentas para descoberta e para fazer previsões (RAJARAMAN; ULLMAN, 2011 apud SMIRAGLIA; XIN, 2017). Um exemplo de aplicação foi identificado no estudo de Wagstaff e Liu (2018, p. 239). Os pesquisadores utilizaram a mineração de dados em estudo experimental para o descartar obras da Biblioteca da Universidade de Wesleyan

Já a mineração de texto de acordo com Pezzini (2016, p. 58) pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural. Faceli et al (2019, p. 331) complementa que como os dados em texto são apresentados de uma forma não estruturada, eles precisam ser convertidos para o formato atributo-valor antes de serem utilizados.

Lamba e Madhusudhan (2019, p. 173), aplicaram a mineração de texto por meio da modelagem de tópico, processo que auxiliou nas etapas de processamento, organização, gerenciamento e extração de conhecimento de um *corpus* de documentos acadêmicos recuperados na base ProQuest.

Por fim, destaca-se a aplicação de técnicas de processamento de linguagem natural (PNL). O PNL, de acordo com Vieira e Lopes (2010, p. 183) é uma área da Ciência da Computação que investiga o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais. Neste âmbito, o PNL utiliza *machine learning* (ML) para revelar estruturas e significados do texto (GOOGLE, 2022). Dentre as suas aplicações destacam-se a análise de sentimentos, a sumarização de conteúdos e a tradução automática de documentos.

Neste contexto, Lamba e Madhusudhan (2019) e Boman (2019) aplicaram em seus projetos o PNL através do Latent Dirichlet Allocation (LDA) ou Alocação de Dirichlet Latente, técnica muito utilizada no processamento de linguagem natural. De acordo com Lamba e Madhusudhan (2019, p. 2019), o LDA trata-se de um modelo probabilístico desenvolvido por Blei et al em 2003, que foi empregado na pesquisa na etapa em que buscou-se inferir automaticamente o tópico discutido no conjunto de documentos.

Em suma, observa-se que os pesquisadores apresentaram poucas discussões acerca das aplicações de métodos e técnicas no desenvolvimento de seus projetos, identificados através descrições sintéticas empregadas nas etapas de desenvolvimentos de novos produtos e serviços.

A seguir será apresentada uma análise dos dados de pesquisa à luz dos critérios elencados em relatório da Europeana.

4.4 Análise dos dados de pesquisa à luz do relatório da Europeana

Esta seção apresentará uma análise holística dos dados do *corpus* da RSL de acordo com critérios de pesquisa apresentados no relatório final da Europeana, sobre o uso de tecnologias baseadas em inteligência artificial em bibliotecas, arquivos, museus e galerias (doravante Glams) publicado em 2021.

À vista disso, o quadro abaixo resume os critérios apresentados no relatório de acordo com a cobertura temática de cada fonte de informação utilizada na pesquisa.

Quadro 10. Uso de tecnologias AI/ML em pesquisas que compõem o *corpus* da RSL

	BRAPCI	ISTA	LISTA	WoS
Extração de conhecimento	33%	0%	0%	0%
Qualidade dos meta(dados)	0%	33%	13%	10%
Análise de audiência	0%	0%	6%	0%
Crowdsourcing e Human in the loop	0%	0%	0%	0%
Visualização de Coleções GLAM	0%	0%	0%	0%
Desenvolvimento de Coleções	0%	33%	19%	15%
Descoberta e pesquisa	100%	67%	31%	20%
Criatividade ou engajamento, projetos e iniciativas	100%	100%	56%	65%
Tradução automática	0%	0%	6%	0%

Fonte: Adaptado de Europeana (2021, p. 7)

O primeiro elemento a ser analisado são as publicações que discutiram ou aplicaram técnicas de extração de conhecimento em suas pesquisas. A

extração de conhecimento é um conjunto de técnicas para a geração de conhecimento a partir da extração de dados e informações disponíveis em documentos de texto completo ou em base de dados. E, de acordo Castro e Ferrari (2016, p. 4) é um processo que integra a mineração de dados.

Neste domínio, a única pesquisa que relatou o uso da extração de conhecimento foi desenvolvida por Lamba e Madhusudhan (2019) e disponibilizada na base de dados Brapci. Todavia, é importante ressaltar que outras pesquisas aplicaram a técnica de mineração de dados, mas sem revelar se a finalidade de fato era a extração de conhecimento.

No que se refere à qualidade dos dados ou dos metadados, esse é um importante elemento de discussão haja vista que a geração de dados imprecisos e instáveis podem converter-se em elementos que comprometem a recuperação da informação e que disseminem conteúdos eivado de vícios. Neste tópico destacam-se que as todas as contribuições analisadas eram de natureza teórica. Já em relação às contribuições por fonte de informação, a base de dados LISTA contribuiu com mais pesquisas em números absolutos (duas publicações) em comparação com a base ISTA (apenas 1 publicação), apesar desta apresentar uma cobertura percentual superior.

Neste contexto, a importância da qualidade dos dados foi destacada e discutida por Kim (2021) em processos de geração automática de metadados. Já em pesquisas disponibilizadas na base de dados LISTA, Harper (2016, p. 7) destacou a importância da mensuração de metadados de “qualidade”. Graser e Burel (2018, p. 12) afirmaram que à medida em que as instituições utilizam métodos de automação que permitem que os profissionais da informação reutilizem ou extraiam metadados de diferentes fontes, a sua qualidade e o contexto se tornam cada vez mais importantes. Na base de dados WoS, Cox, Pinfield e Rutter (2019) relataram a partir de entrevistas certa preocupação dos profissionais da informação com implicações referentes à qualidade e segurança dos dados advindos do emprego de aplicativos de análise de aprendizado. Já Golub (2019) discutiu brevemente a qualidade de termos de indexação no contexto da indexação automática.

Em relação à análise de audiência, buscou-se verificar nas pesquisas preocupações relativas a experiência e a satisfação do usuário na aplicação de

ML em produtos e serviços de bibliotecas. Deste modo, apenas uma pesquisa desenvolvida por Gomez et al (2020) relatou a aplicação de testes para avaliar o serviço de marcação automática de imagens em três coleções digitais da Universidade da Califórnia. Todavia, os testes foram realizados com voluntários externos à instituição.

É importante destacar que apesar do baixo número de trabalhos que externalizaram contribuições no âmbito da experiência e da satisfação do usuário em projetos de ML/AI em bibliotecas, foram relatadas nas pesquisas preocupações referentes à privacidade (LITSEY; MAULDIN, 2018; HAPER, 2018; GRIFFEY, 2019; GARCÍA-FEBO, 2019; JOHNSON, 2018; COX; PINFIELD; RUTTER, 2019; HAHN; MCDONALD, 2018) e a proteção de direitos autorais (JOHNSON, 2018; COX; PINFIELD; RUTTER, 2019).

No que se refere ao emprego de *Crowdsource* e *Human in the loop* nas pesquisas. O primeiro de acordo com Holley (2010) trata-se de uma atividade que utiliza técnicas de engajamento social para ajudar um grupo de pessoas a alcançar um objetivo compartilhado, geralmente significativo e grande, trabalhando de forma colaborativa como um grupo. O *crowdsourcing*, ainda de acordo com Holley (2010), também geralmente envolve esforço, tempo e contribuição intelectual de um indivíduo para uma comunidade.

Já o "*Human in the loop*" (HILT) em *machine learning* (ML) é o processo o qual humanos e algoritmos trabalham juntos para resolver problemas com mais eficiência e precisão do que cada um poderia realizar sozinho (AVERKAMP et al (2021, p. 5). Ou seja, baseia-se na premissa de que uma pessoa pode ensinar, treinar e testar um sistema com o objetivo de auxiliá-lo a apresentar resultados mais precisos e confiáveis.

Neste tópico, nenhuma pesquisa relatou ou discutiu acerca o uso ou aplicação de *Crowdsource* e *Human in the loop* no desenvolvimento de seus projetos.

No que tange ao desenvolvimento de coleções, trata-se de um procedimento que busca gerenciar de maneira sistêmica as atividades de seleção, aquisição, desbaste e descarte em coleções físicas e/ou digitais de uma biblioteca, de acordo com as necessidades de uma determinada comunidade.

À vista disso, foram identificadas 7 pesquisas que discutiram as implicações da AI/ML em processos de desenvolvimento de coleções. Destacaram-se as contribuições de Giblin et al (2019) que investigou a disponibilidade de e-books para empréstimo eletrônico em cinco países através da análise de termos de licença e preços. Walker e Juang (2019) que aplicaram o algoritmo AdaBoost na aquisição de obras por demanda. E, por fim, Wagstaff e Liu (2018) que desenvolveram um estudo acerca do descarte de obras por meio do uso de algoritmos de classificação.

Em relação ao tópico de pesquisa e descoberta, foram analisadas publicações que aplicaram ou discutiram o emprego de AI/ML nos processos de descoberta e pesquisa de produtos e serviços de bibliotecas. Este tópico foi o segundo mais discutido no *corpus* da RSL. Além disso, constatou-se que todas as publicações da base Brapci abordaram tanto processos de descoberta quanto de pesquisas. Porém em números absolutos a base LISTA foi a base que mais apresentou contribuições, com 6 publicações.

Neste domínio, Kiam e Chinnasamy (2013) aplicaram técnicas para vincular conteúdos de coleções digitais com a coleções de artigos da Infopedia de Cingapura o que aumentou a capacidade de descoberta de ambas as coleções. A pesquisa desenvolvida por Jim e McDonald (2017) tem a finalidade de desenvolver um recomendador baseado em ML para ambientes de descoberta. Já Yelton (2019) treinou uma rede neural com mecanismo de recomendação que permite pesquisar em teses por autor ou título e informar quais outras teses são conceitualmente mais semelhantes. Gomez et al (2020) utilizaram serviços de marcação automática de imagens com o objetivo de melhorar os resultados de pesquisa na biblioteca digital.

O tópico criatividade ou engajamento, projetos e iniciativas, revelou-se o mais aplicado nas pesquisas. Nas bases Brapci e ISTA obtiveram 100% das discussões em pesquisas. Todavia, em números absolutos a base de dados Web of Science contribuiu com 14 publicações. No que se refere às publicações que não aplicaram nenhum elemento deste tópico, constatou-se que em grande parte, tratam-se de pesquisas que buscaram revisar a literatura, em grande medida alinhadas ao domínio da representação do conhecimento.

Por fim, o último tópico a ser analisado é a tradução automática que de acordo com a Microsoft (2022b) são aplicações ou serviços *online* que utilizam tecnologias de ML para traduzir grandes quantidades de texto para outro idioma. Neste contexto Kim (2019) apresentou a única discussão que analisou brevemente a tradução automática no campo do *Deep learning*.

Em suma, destaca-se que a base LISTA apresentou maior diversidade no conjunto de pesquisas de acordo com as categorias elencadas para análise, em comparação com a base Brapci que apresentou pesquisas que se concentraram apenas em 3 tópicos: extração de conhecimento, descoberta e pesquisa e criatividade e engajamento, projetos e iniciativas.

5 TENDÊNCIAS E RECOMENDAÇÕES PARA A APLICAÇÃO DO *MACHINE LEARNING* EM BIBLIOTECAS

Bibliotecas são instituições que tradicionalmente reúnem grandes conjuntos de informação, todavia esse cenário não se traduz em acesso instantâneo a estes documentos, haja vista que grande parte desses conjuntos de dados estão armazenados em formatos analógicos. Aliado a este desafio, as instituições apresentam limitações no quadro de trabalho, com pouca mão-de-obra especializada no tratamento da informação. Por esse motivo, à medida que a inteligência artificial se incorpora ao cotidiano popular, técnicas de *machine learning* transformam o ambiente de biblioteca e incorporaram soluções inovadoras no tratamento e na descoberta de informação.

À vista disso, seis aplicações de ML em bibliotecas se destacaram na literatura científica internacional, e algumas dessas atividades também foram objeto de discussão por Ryan Cordell[48] e tratadas como aplicações promissoras de MI em bibliotecas. Essas tendências, desenvolvidas e documentadas em pesquisas, já são uma realidade em instituições internacionais.

Neste âmbito, se destacaram as seguintes aplicações, a saber: geração automática de metadados, classificação e *clustering* (ou agrupamento), vinculação de coleções, anotação de dados, descoberta de conhecimento e alfabetização de dados.

[48] Cordell (2020) produziu um relatório encomendado pela Library of Congress, o qual apresentou um panorama do aprendizado de máquina em bibliotecas.

Figura 7. Bibliotecas + *Machine Learning*.

Fonte: Elaborado pela autora (2022).

Todavia, antes de iniciar as discussões em torno das seis tendências em aplicações do MI em bibliotecas, é de suma importância uma análise centrada no conjunto de dados a ser explorado. Um dos grandes desafios de bibliotecas e instituições de memória é o tratamento de grandes coleções de documentos produzidos no formato análogo. À vista disso, é necessário que estas instituições convertam esses documentos para formatos digitais, simultaneamente com a aplicação de ferramentas de reconhecimento óptico de caracteres ou de informação manuscrita. O cumprimento desta etapa torna viável a aplicação de ferramentas de ML facilitando o acesso à informação e a descoberta de novos conteúdos.

Neste âmbito, no que tange a construção de conjuntos de dados para a aplicação de técnicas de ML, recomenda-se que este conjunto contenha dados bem distribuídos, representativos, pouco ruidosos e livre de viés. Ademais, é importante documentar a procedência desses dados de modo a tornar o projeto transparente.

Todavia, observa-se que os conjunto de dados muitas vezes podem conter lacunas (ausência de valores), duplicações de registros, ambiguidades e informações discrepantes (*outliers*). Desta forma, é necessário na etapa de pré-

processamento de dados a aplicação de ferramentas de limpeza de maneira a reduzir essas inconsistências e padronizar os valores.

A partir desta perspectiva, a primeira aplicação que se apresenta como uma tendência em bibliotecas é a geração automática de metadados. Um método que, de acordo com Cordell (2020), busca reconhecer automaticamente palavras significativas, como nomes próprios, nomes geográficos e marcadores temporais de dados de texto. Esse método pode ser implementado como complemento à entrada manual de metadados, visto que a descrição manual é uma atividade morosa, propensa a erros e que pode apresentar custos mais elevados. Além disso, este método pode apoiar bibliotecários no enriquecimento de metadados já existentes, acrescentando novas informações de modo a preencher lacunas semânticas e criar novos pontos de acesso para documentos.

A respeito da aplicação de processos de classificação e *clustering*, muito discutidos na literatura cientifica internacional, Cordell (2020) afirma que eles são frequentemente apontados como técnicas valiosas por causa de sua capacidade de identificar conexões entre materiais, integrantes de grandes conjuntos de dados, que podem não ser aparentes para humanos, devido às limitações de atenção e memória humana.

Outra tendência é a aplicação de técnicas que permitam a vinculação e a integração entre diferentes fontes de informação no contexto da *web* semântica. A aplicação destas técnicas associadas ao movimento do *Linked Open Data*, auxiliam no enriquecimento da qualidade de metadados ao mesmo tempo em que amplia a visibilidade e a descoberta de recursos além das fronteiras das bibliotecas.

Já a anotação de dados é um método que procura rotular imagens com alta eficiência e baixa subjetividade de modo a explorar semelhanças a partir de seus conteúdos semânticos. A aplicação desta técnica permite tornar grandes coleções de imagens recuperáveis.

No que tange a descoberta de conhecimento, Cordell (2020) destaca que todas as aplicações anteriormente citadas corroboram para processos de descoberta de conhecimento, de modo a auxiliar na pesquisa e acessibilidade de documentos. Este cenário apresenta-se como ambiente propicio para a

serendipidade, princípio definido por Cunha e Cavalcanti (2018, p. 332) como a circunstância em que, acidentalmente, quando se está a procurar algo, descobre-se, inesperadamente outra coisa, que nos convém e traz satisfação.

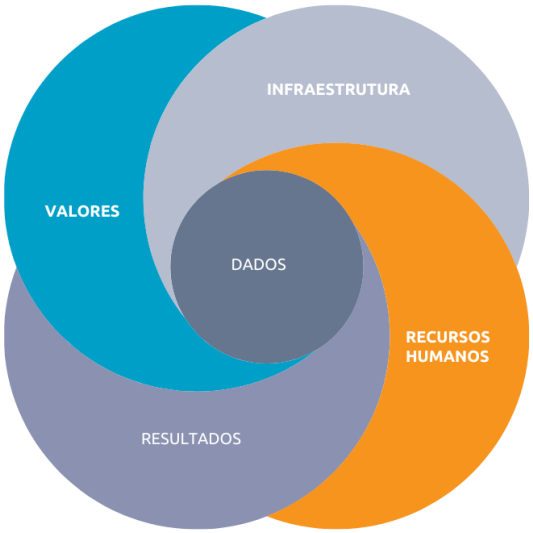
Por fim, e não menos importante, a medida em que a IA/ML estimulam a geração e o acesso à informação é necessário fomentar esforços educacionais bem como o pensamento crítico no ensino de técnicas e métodos de IA, tornando os usuários e os profissionais da informação agentes mais ativos contra sistemas crescentes de vigilância e injustiça algorítmica (CORDELL, 2020, p. 31). Ademais, é importante reforçar o compromisso das bibliotecas com a privacidade e com a liberdade intelectual.

A aplicação de todas essas tendências, corroboram com o pensamento de Feigenbaum (1989, p. 298), o qual as bibliotecas do futuro abarcarão coleções de documentos ativos que realizam novas conexões, associações e analogias, sem que os usuários precisem definir claramente as suas necessidades de informação.

Neste ponto é necessário elucidar algumas recomendações gerais para a implementação de projetos de ML em bibliotecas. Essas recomendações são baseadas em contribuições de Ryan Cordell (2020) e Thomas Padilla[49] (2019). Posto isso, destacam-se cinco categorias: valores, recursos humanos, infraestrutura, dados e resultados, representadas na imagem abaixo:

[49] Thomas Padilla produziu um relatório para a OCLC que delimita as operações responsáveis em Ciência de dados, aprendizado de máquina e IA em bibliotecas

Figura 8. Categorias de recomendações para projetos de ML em bibliotecas



Fonte: Elaborado pela autora (2022).

No que tange aos valores, a Library of Congress (CORDELL, 2020, p. 15) recomenda o desenvolvimento de uma declaração o qual o objetivo principal é explicitar os valores e princípios que nortearão o uso, a aplicação e o desenvolvimento do aprendizado de máquina na biblioteca. Neste âmbito, é importante também o fomento à aplicação responsiva do ML fundamentada em princípios éticos, considerando a responsabilidade algorítmica, a representação justa e a mitigação de vieses em todas as etapas do projeto. Além disso, atrelados aos valores e padrões éticos, todo projeto deve ser passível de auditabilidade e explicabilidade por meio de avaliações do impacto algoritmo e interfaces explicáveis, aplicados e avaliados por meio de processos de auditoria ao longo de todo o projeto.

Desta forma, baseado nas recomendações de Cordell (2020), o quadro abaixo sintetiza os principais aspectos que devem ser observados na categoria valores:

Quadro 11. Recomendações – valores

RECOMENDAÇÕES - VALORES
1. Elaborar uma declaração de valores para projetos de ML em biblioteca que reflita padrões éticos passíveis de avalições por auditorias.
2. Caso a declaração de valores não tenha sido desenvolvida, buscar por modelos de declarações de instituições similares que possam ser adotados ou adaptados.
3. Desenvolver avaliações de impacto algorítmico e um plano para ML explicáveis;
4. Analisar das consequências ambientais da aplicação do projeto, bem como de suas compensações.
5. O plano deve possibilitar a revisão (se necessário) e a adaptação de valores à medida que o projeto se desenvolve.

Fonte: Adaptado de Cordell (2020)

Em seguida, é necessário criar e definir o escopo do projeto de ML para a biblioteca de maneira a realizar uma análise interna da instituição com relação às necessidades de sua comunidade e os objetivos de sua aplicação.

A segunda categoria de recomendações refere-se aos recursos humanos disponíveis para a implementação de projetos. É necessário a formação de uma equipe multiprofissional, com experiencia e conhecimento sobre técnicas de ciências de dados, inteligência artificial e aprendizado de máquina. Padilla (2019) recomenda para os profissionais da informação a realização de simpósios, a formação de grupos de trabalho e a promoção de parcerias entre instituições.

Em suma, as recomendações abaixo buscam orientar gestores acerca das principais ações quanto aos recursos humanos em projetos de Ml em bibliotecas.

Quadro 12. Recomendações – recursos humanos

RECOMENDAÇÕES - RECURSOS HUMANOS
1. Verificar se a equipe possui experiência e formação adequada para elaborar o projeto de ML. É necessário conhecimento para gerenciar a aquisição, limpeza e gerenciamento de dados; anotação de dados de treinamento específico do domínio; desenvolvimento e implantação de ML; desenvolvimento de interface e visualização; avaliação de resultados.
2. Caso a equipe não tenha a experiência necessária, é necessário realizar o treinamento da equipe para atender às necessidades do projeto.

3. Buscar oportunidades de colaboração com outras instituições para auxiliar nas demandas do projeto e/ou trocar experiências.
4. Verificar se a equipe reflete a diversidade de opiniões, origens e tipos de pensamento.
5. Elaborar um memorando de entendimento descrevendo os deveres de todos os colaboradores e os resultados previstos, tanto coletivos quanto individuais, para o trabalho do projeto.
6. Garantir que as habilidades e conhecimentos de todos os participantes do projeto sejam valorizados, bem como definir baseado nos resultados que todos os colaboradores sejam reconhecidos e recompensados de forma justa por suas contribuições.

Fonte: Adaptado de Cordell (2020)

É importante destacar que fomentar a capacitação de profissionais exigirá investimentos financeiros tanto em infraestrutura como em conhecimento.

No que tange à infraestrutura técnica, a instituição deverá dispor de *hardwares* e *softwares* adequados para o treinamento de modelos de ML. Cordell (2020, p. 39) destaca que os custos de *hardware*, de acordo com acadêmicos independentes, são uma barreira significativo ao seu trabalho.

Quadro 13. Recomendações – infraestrutura técnica

RECOMENDAÇÕES - INFRAESTRUTURA TÉCNICA
1. Verificar se a biblioteca já possui *hardware* e *software* adequados para treinar modelos de ML e aplicá-los nas coleções desejadas.
2. Inspecionar se os ambientes de *software* estão corretamente configurados para executar algoritmos de ML.
3. Verificar se existem estruturas adequadas para o armazenamento de dados.
4. Caso a instituição não disponha de estrutura de *hardware* e *software* adequados é necessário verificar se a instituição possui orçamento para comprar ou alugar equipamentos e espaço necessário para o desenvolvimento do projeto.

Fonte: Adaptado de Cordell (2020)

Em relação à categoria de conjunto de dados, é necessário realizar uma análise holística desses elementos de modo a compreender e avaliar os dados utilizados no domínio da pesquisa, bem como decidir quais conjuntos serão

utilizados como teste. Deste modo é importante observar as recomendações elencadas no quadro abaixo:

Quadro 14. Recomendações – dados

RECOMENDAÇÕES - DADOS
1. Apurar se existem dados acionáveis por máquina no domínio da pesquisa. Caso os dados que serão analisados estejam em formato analógico, será necessário elaborar um plano de digitalização das coleções.
2. Analisar se os dados de treinamento disponíveis possuem cobertura que abrange todos os aspectos da pesquisa.
3. Caso o projeto não conte com dados de treinamento disponíveis para o domínio a ser modelado será necessário buscar dados semelhantes que possam ser utilizados para pré-treinar um modelo de ML.
4. Auditar os dados de treinamento para garantir que sejam representativos e mitigar os inevitáveis vieses de disponibilidade, seleção e especialização.
5. Verificar se existem comunidades de especialistas ou pesquisadores que possam envolver significativamente na anotação de dados de treinamento, avaliação de resultados ou interfaces e visualizações piloto.

Fonte: Adaptado de Cordell (2020)

Por fim, na categoria resultados deverão ser observados a comunicação dos resultados do projeto o que inclui dados de treinamento, algoritmos e códigos e a descrição de ferramentas utilizadas em todas as etapas, conforme recomendações do quadro abaixo:

Quadro 15. Recomendações – resultados

RECOMENDAÇÕES - RESULTADOS
1. Existe um *pipeline* claro estabelecido para passar da concepção do projeto à implementação sustentável dos resultados de ML?
2. Verificar se o projeto permite que interfaces ou visualizações que modelem ML explicável comuniquem os resultados de nosso trabalho para a comunidade?
3. Verificar se existe a possibilidade de publicar todos os dados de treinamento, código e dados anotados por ML para beneficiar outros pesquisadores de ML.

Fonte: Adaptado de Cordell (2020)

Em suma, todos esses aspectos destacados nesta seção corroboram para o fortalecimento do papel das bibliotecas que continuará centrado na gestão, tratamento e descoberta de coleções, porém desenvolvidas sob uma nova perspectiva que inclui a inteligência artificial e o *machine learning*.

6 CONCLUSÃO

Esta pesquisa se propôs a mapear o estado da arte da inteligência artificial com ênfase no aprendizado de máquina em bibliotecas sob a perspectiva da Ciência da Informação. À vista disso, foram discutidos e explorados as aplicações, benefícios e impactos gerados em produtos e serviços; identificados as novas competências do profissional da informação, às questões relativas aos valores éticos, bem como foi produzido um conjunto de recomendações para a aplicação de técnicas de Ml em bibliotecas baseado nas tendências apresentadas.

Neste cenário, foram analisados 42 documentos em cinco fontes de informação selecionados a partir de 3 critérios de inclusão e 11 critérios de exclusão, cuja natureza das publicações revelou tratar-se de publicações em língua inglesa concentradas nos Estados Unidos e que se dedicavam a discutir as aplicações práticas do desenvolvimento de produtos e serviços para bibliotecas por meio de técnicas de Ml. Neste sentido, as aplicações práticas se concentravam na alfabetização de dados, desenvolvimento de coleções e circulação, processos de recuperação da informação desenvolvidas em conjunto de dados textuais e imagéticos. Além disso, verificou-se uma profusão ferramentas, metodologias e técnicas por meio do emprego de linguagens de programação, algoritmos, ferramentas e métodos citados ao longo dos trabalhos que compõem a RSL.

Já em relação as competências do profissional da informação destacaram-se contribuições no domínio do pensamento crítico complementado por elementos éticos e consciência algorítmica, o fomento a habilidades técnicas com foco em alfabetização de dados e curadoria digital e o pensamento analítico baseados em raciocínio lógico e análise de fluxos de dados. Já sob a perspectiva ética as pesquisas evidenciaram a responsabilidade algorítmica dos profissionais da informação e o respeito aos princípios da liberdade intelectual, da privacidade e do acesso à informação.

A análise segundo os critérios listados no relatório da Europeana revelou uma preocupação com desenvolvimento de coleções e com a descoberta de conhecimento e a pesquisa em detrimento de projetos que apliquem

crowdsourcing e *Human in the Loop*. Por fim, foram identificadas seis tendências promissoras em bibliotecas, a saber: geração automática de metadados, classificação e *clustering* (ou agrupamento), vinculação de coleções, anotação de dados, descoberta de conhecimento e alfabetização de dados. Nesta etapa foram desenvolvidos um conjunto de recomendações que abarcam questões relativas aos valores, infraestrutura, resultados, recursos humanos e dados.

A literatura aponta que o desenvolvimento de produtos e serviços por meio de técnicas de IA/ML é uma realidade em instituições internacionais, haja vista que o cenário exponencial de dados torna desafiador o tratamento manual de grandes conjuntos de dados, devido à limitação da tríada tempo, trabalho e pessoal.

Além disso, observa-se que a maioria dos experimentos apresentados e discutidos na literatura cientifica indicam uma tendência para aplicações em atividades meio, que se concentram em técnicas de seleção, descarte, tratamento e descoberta de conhecimento, constatadas por meio do conjunto de palavras-chaves utilizadas na descrição das publicações, na análise das aplicações mais utilizadas na literatura e nas técnicas empregas no tratamento dos conjuntos de dados. Além disso, a aplicação dessas técnicas possibilitou a criação de novos *insights* e conexões inesperadas entre as coleções.

Em suma, as bibliotecas precisam tornar-se agentes ativos neste cenário em meio à abundância de informações, pois a literatura aponta que AI/ML será cada vez mais incorporada tanto aos fluxos de trabalho, como na transformação da experiência do usuário.

Por fim, como sugestão de pesquisas futuras:

- Análise acerca das aplicações de *Crowdsource* e *Human in the loop* no contexto das bibliotecas brasileiras;
- Estudo acerca dos impactos da catalogação e indexação automática ou semiautomática por meio da aplicação de técnicas de ML;
- Desenvolvimento de projetos automação de aquisição e descarte de obras;

- Criação de modelos para a geração automática de metadados em bibliotecas digitais.

REFERÊNCIAS BIBLIOGRÁFICAS

AKHTAT, S. M. F. *Big Data Architect's Hand-book*. Birmingham: Packt Publishing, 2018. *E-book*. Disponível em: https://pt.scribd.com/document/458827908/BIG-DATA-ARCHITECTS-HANDBOOK-pdf. Acesso em: 3 set. 2021.

ALGORITHM. *In*: Encylopaedia Britannica. [Inglaterra: Encylopaedia Britannica Inc. 2022]. Disponivel em: https://www.britannica.com/science/algorithm. Acesso em: 27 ago. 2022.

ARLITSCH, K.; NEWELL, B. Thriving in the age of accelerations: a brief look at the societal effects of artificial intelligence and the opportunities for libraries. *Journal of Library Administration*, [*S. l*], v. 54, n. 7, 2017. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01930826.2017.1362912?tab=permissions&scroll=top. Acesso em: 27 ago. 2022.

AYRE, L.; CRANER, J. Algorithms: avoiding the implementation of institutional biases. *Public Library Quarterly*, [*S. l*], v. 37, n. 3, 2018. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01616846.2018.1512811?journalCode=wplq20. Acesso em: 27 ago. 2022.

AVERKAMP, S. et al. *Humans-in-the-loop*: recommendations report. Nova York: Library of Congress, 29 nov. 2021. Disponivel em: https://labs.loc.gov/static/labs/work/reports/LC-Labs-Humans-in-the-Loop-Recommendations-Report-final.pdf. Acesso em: 27 ago. 2022.

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.

BARDIN, L. *Análise de conteúdo*. 4. ed. Edições 70: Lisboa, 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.

BOMAN, C. An Exploration of Machine Learning in Libraries. *In*: GRIFFEY, J. *Artificial Intelligence and Machine Learning in Libraries*. [*S. l*]: ALA; Library Technology Reports, v. 55, n. 1, p. 21-25, 2019. Disponível em: https://journals.ala.org/index.php/ltr/article/view/6911/9303. Acesso em: 27 ago. 2022.

BOMHOLD, C. R. Educational use of smart phone technology. *Program:* electronic library and information systems, [*S. l*.], v. 47, n. 4, p. 424-436, 2013. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PROG-01-2013-0003/full/html. Acesso em: 27 ago. 2022.

BRAGA, K. S. Aspectos relevantes para a seleção de metodologia adequada à pesquisa social em Ciência da Informação. *In*: MUELLER, S. (org.). *Métodos para a pesquisa em Ciência da Informação*. Brasília: Thesaurus, 2007. p. 17-38.

BORTOLETTI, M. *Como desenvolver o pensamento analítico com 9 dicas*. Goiânia: PUC Goiás, 20 dez. 2021. Disponível em: https://ead.pucgoias.edu.br/blog/pensamento-analitico. Acesso em: 27 ago. 2022.

BROUGHTON, V. The Respective Roles of Intellectual Creativity and Automation in Representing Diversity: Human and Machine Generated Bias. *Knowledge Organization*, [*S. l*], v. 46, n. 8, 2019. p. 586-606. Disponível em: https://www.researchgate.net/publication/339175818_The_Respective_Roles_of_Intellectual_Creativity_and_Automation_in_Representing_Diversity_Human_and_Machine_Generated_Bias. Acesso em: 27 ago. 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.

BUSH, V. As we may think. *Atlantic Monthly*, [*S. l*.], v. 176, 1, p. 101-108, 1945. Disponível em: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm. Acesso em: 28 set. 2020.

CARDOSO, M. *Um pouco sobre a linguagem de programação C*. Campina Grande: UFCG, 2 jul. 2020. Disponível: https://edu.ieee.org/br-ufcgras/um-pouco-sobre-a-linguagem-de-programacao-c/#:~:text=A%20linguagem%20C%2B%2B,das%20limita%C3%A7%C3%B5es%20da%20linguagem%20C. Acesso em: 27 ago. 2022.

CASTRO, L. N.; FERRARI, D. G. *Introdução à mineração de dados*: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CORDELL, R. *Machine learning + Libraries*: a report on the state of the field. [*S. l*], Library of Congress, 14 jul. 2020. Disponível em: https://blogs.loc.gov/thesignal/2020/07/machine-learning-libraries-a-report-on-the-state-of-the-field/. Acesso em: 1 set. 2022.

CORTEZ, E.; SILVA, A. S. da. Unsupervised strategies for information extraction by text segmentation. *In*: Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research. New York, NY, USA: ACM, 2010. p. 49–54. Disponível em: https://dl.acm.org/doi/abs/10.1145/1811136.1811145. Acesso em: 1 set. 2022.

COX, A. M.; PINFIELD, S.; RUTTER, S. The intelligent library Thought leaders' views on the likely impact of artificial intelligence on academic libraries. *Library Hi Tech*, [*S. l*], v. 37, n. 3, 2019. Disponível em:

https://www.emerald.com/insight/content/doi/10.1108/LHT-08-2018-0105/full/html. Acesso em: 27 ago. 2022.

CUNNINGHAM, S. J.; LITTIN, J.; WITTEN, I. H. Applications of machine learning in information retrieval. *Working Paper*, Nova Zelândia, v. 6, n. 97, 1997. Disponível em: https://researchcommons.waikato.ac.nz/handle/10289/1069. Acesso em: 1 set. 2022.

DARÁNYI, S.; WITTEK, P.; MCPHERSON, M. P. D. Using wavelet analysis for text categorization in digital libraries: a first experiment with Strathprints. *International Journal on Digital Libraries*, [*S. l*], v. 3, n. 12, p. 3-12, 2012. Disponível em: https://www.researchgate.net/publication/226911903_Using_wavelet_analysis_for_text_categorization_in_digital_libraries_A_first_experiment_with_Strathprints. Acesso em: 27 ago. 2022.

EBSCO INFORMATION SERVICES. Information Science & Technology Abstracts (ISTA). Massachusetts: EBSCO Information Services, 2021.

EBSCO INFORMATION SERVICES. Library, Information Science & Technology Thesaurus. *In*: EBSCO INFORMATION SERVICES. *Library, Information Science & Technology*. Massachusetts: EBSCO Information Services, 2021.

EBSCO INFORMATION SERVICES. *Library, Information Science & Technology (LISA)*. Massachusetts: EBSCO Information Services, 2021.

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.

EUROPEANA. *Interim analysis of EuropeanaTech AI in Relation to GLAMs survey*. [Amsterdam]: EUROPEANA, 2020. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/Final_Interim_Report_AI_in_GLAMs_TF.pdf. Acesso em: 27 ago. 2022.

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

FACELI et al. *Inteligência artificial*: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2022. *E-book*.

FELIZARDO, K. R. et al. Seleção e avaliação de estudos. *In*: FELIZARDO, K. M. et al. *Revisão sistemática da literatura em engenharia de software*: teoria e prática. 1. ed. Rio de Janeiro: Elsevier, 2017. p. 51-70.

FABBRI, S. C. P. F.; OCTAVIANO, F. R.; HERNANDES, E. C. M. Protocolo da revisão sistemática. *In*: FELIZARDO, K. M. et al. *Revisão sistemática da*

literatura em engenharia de software: teoria e prática. 1. ed. Rio de Janeiro: Elsevier, 2017.

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.

FEDERAÇÃO INTERNACIONAL DE ASSOCIAÇÃO DE BIBLIOTECAS E INSTITUIÇÕES (IFLA). *IFLA statement on libraries and artificial intelligence*. IFLA: Haia, 17 set. 2020. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/faife/ifla_statement_on_libraries_and_artificial_intelligence.pdf. Acesso em: 23 set. 2021.

FEDERAÇÃO INTERNACIONAL DE ASSOCIAÇÃO DE BIBLIOTECAS E INSTITUIÇÕES (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.

FEIGENBAUM, E. A. Toward the library of the future. *Long Range Planning*, [*S. l*], v. 22, n. 1, feb., 1989, p. 118-123. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/0024630189900599. Acesso em: 1 set. 2022.

FELIZARDO, K. M. *Revisão sistemática de literatura em engenharia de software*: teoria e prática. 1. ed. Rio de Janeiro: Elsevier, 2017.

FERREIRA, A. A. et al. Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science & Technology*, [*S. l.*], v. 6, n. 65, p. 1257-1276, 2014. Disponível em: https://dl.acm.org/doi/10.1002/asi.22992. Acesso em: 27 ago. 2022.

FINLEY, T. The Democratization of Artificial Intelligence: One Library's Approach. *Information Technology & Libraries*, [*S. l.*], v. 38, n. 1, 2019. Disponível em: https://ejournals.bc.edu/index.php/ital/article/view/10974. Acesso em: 27 ago. 2022.

FOX, R. Digital libraries: the systems analysis perspective machine erudition. *Digital Library Perspectives*, [*S. l.*], v. 32, n. 2, 2016. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/DLP-02-2016-0006/full/html. Acesso em: 27 ago. 2022.

GALVÃO, M. C. B.; RICARTE, I. L. M. Revisão sistemática da literatura: conceituação, produção e publicação. *Logeion*: filosofia da informação, Rio de Janeiro, v. 6, n. 1, p. 57-73, 2020. Disponível em: https://revista.ibict.br/fiinf/article/view/4835. Acesso em: 27 ago. 2022.

GARCIA-FEBO, L. Exploring AI: How libraries are starting to apply artificial intelligence in their work. *American Libraries*, mar./apr. 2019. Disponível em:

https://americanlibrariesmagazine.org/2019/03/01/exploring-ai/. Acesso em: 27 ago. 2022.

GERHARDT, T. E.; SILVEIRA, D. T. *Métodos de pesquisa*. Porto Alegre: Editora da UFRGS, 2009.

GIBLIN, R. et al. What can 100,000 books tell us about the international public library e-lending landscape*?. Information Research-An International Electronic Journal*, [*S. l.*], v. 24, n. 3, set. 2019. Disponível em: http://informationr.net/ir/24-3/paper838.html. Acesso em: 27 ago. 2022.

GIL, A. C. *Métodos e técnicas de pesquisa social*. 7. ed. São Paulo: Atlas, 2019. *E-book*.

GIL-LEIVA, I. SISA – Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules. *Knowledge Organization*, [*S. l.*], v. 3, n. 44, 2014. P. 139-162. Disponível em: https://www.researchgate.net/publication/318222896_SISA-Automatic_Indexing_System_for_Scientific_Articles_Experiments_with_Location_Heuristics_Rules_Versus_TF-IDF_Rules. Acesso em: 1 set. 2022.
44: 139-62.

GODBY, C. J.; REIGHART, R. R. The WordSmith indexing system. *Journal of Library Administration*, [*S. l.*], v. 34, n. 3-4, 2008. Disponível em: https://www.tandfonline.com/doi/abs/10.1300/J111v34n03_18. Acesso em: 1 set. 2022.

GOOGLE Inc. *Google Ngram Viewer*: Artificial Intelligence; Machine Learning. [*S. l.: s. n.*], 2022a. Disponível em: https://books.google.com/ngrams/graph?content=Artificial+Intelligence%2CMachine+Learning&year_start=1800&year_end=2019&corpus=26&smoothing=3&direct_url=t1%3B%2CArtificial%20Intelligence%3B%2Cc0%3B.t1%3B%2CMachine%20Learning%3B%2Cc0#t1%3B%2CArtificial%20Intelligence%3B%2Cc0%3B.t1%3B%2CMachine%20Learning%3B%2Cc0. Acesso em: 1 set. 2022.

GOOGLE Inc. *TensorFlow*: por que usar o TensorFlow. [*S. l.: s. n.*], 2022a. Disponível em: https://www.tensorflow.org/about?hl=pt-br. Acesso em: 27 ago. 2022.

GOOGLE Inc. *About keras*. [*S. l.: s. n.*], 2022b. Disponível em: https://keras.io/about/. Acesso em: 27 ago. 2022.

GOLUB, K. Automatic Subject Indexing of Text. *Knowledge Organization*, [*S. l.*], v. 2, n. 46, 2019, p. 104-121. Disponível em: https://www.researchgate.net/publication/333085307_Automatic_Subject_Indexing_of_Text. Acesso em: 27 ago. 2022.

GOLUB, K.; HAGELBÄCK, J.; ARDÖ, A. Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *Journal of Data and Information Science*, [*S. l.*], v. 5, n. 1, fev. 2020. Disponível

em: https://www.sciendo.com/article/10.2478/jdis-2020-0003. Acesso em: 27 ago. 2022.

GOMEZ, J. et al. Experimenting with a Machine Generated Annotations Pipeline, *Code4Lib Journal*, [*S. l.*], v. 48, 2020. Disponível em: https://journal.code4lib.org/articles/15209. Acesso em: 27 ago. 2022.

GRASER, M.; BUREL, M. Metadata Automation: The Current Landscape and Future Developments. *VRA Bulletin*, [*S. l*], v. 45, n. 2, p. 1-14, 2018. Disponivel em: https://online.vraweb.org/index.php/vrab/article/view/34. Acesso em: 27 ago. 2022.

GRIFFEY, J. AI and Machine Learning: The challenges of artificial intelligence in libraries. *American Libraries Magazine*, [*S. l*], mar./apr., p. 47, 2019. Disponível em: https://americanlibrariesmagazine.org/2019/03/01/ai-machine-learning-libraries/. Acesso em: 27 ago. 2022.

HAHN, J.; MCDONALD, C. Account-based recommenders in open discovery environments, *Digital Library Perspectives*, [*S. l.*], v. 34, n. 1, 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/DLP-07-2017-0022/full/html. Acesso em: 27 ago. 2022.

HOLLEY, R. Crowdsourcing: how and why should libraries do it?. *D-Lib Magazine*, [*S. l*], v. 16, n. 3-4, mar./apr. 2010. Disponível em: http://www.dlib.org/dlib/march10/holley/03holley.html. Acesso em: 27 ago. 2022.

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.

HARPER, C. A. Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). *Code4Lib Journal*, [*S. l.*], v. 33, 2016. Disponivel em: https://journal.code4lib.org/articles/11752. Acesso em: 27 ago. 2022.

HARPER, C. Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords. *Code4Lib Journal*, [*S. l.*], n. 41, 2018. Disponivel em: https://journal.code4lib.org/articles/13671. Acesso em: 27 ago. 2022.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Brasília: IBICT, 2021.

INSTITUTO FEDERAL DE EDUCAÇÃO (Santa Catarina) (IFSC). *Linguagem de programação*. Florianópolis: IFSC, 27 out. 2020. Disponível em: https://wiki.sj.ifsc.edu.br/index.php/MCO018703_2020_2_AULA02. Acesso em: 27 ago. 2022.

JOHNSON, B. Libraries in the Age of Artificial Intelligence. *Computers in Libraries*, [*S. l.*], v. 38, n. 1, jan./feb., 2018, p. 14-16. Disponível em: https://www.infotoday.com/cilmag/jan18/Johnson--Libraries-in-the-Age-of-Artificial-Intelligence.shtml. Acesso em: 27 ago. 2022.

JOORABCHI, A.; MAHDI, A. E. Classification of scientific publications according to library controlled vocabularies A new concept matching-based approach. *Library Hi Tech*, v. 31, n.4, 2013. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LHT-03-2013-0030/full/html. Acesso em: 27 ago. 2022.

JOORABCHI, A.; MAHDI, A. E. Improving the visibility of library resources via mapping library subject headings to Wikipedia articles. *Library Hi Tech*, [S. l.], v. 36, n. 1, 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LHT-04-2017-0066/full/html. Acesso em: 27 ago. 2022.

JOORABCHI, A.; MAHDI, A. E. Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts. *Journal of Information Science*, [*S. l.*], 2014. Disponivel em: https://journals.sagepub.com/doi/10.1177/0165551513514932. Acesso em: 27 ago. 2022.

KENNEDY, M. L. What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries? *Research Library Issues*, [*S. l.*], v. 299, p. 3-13, 2019. Disponível em: https://publications.arl.org/18nm1db/. Acesso em: 27 ago. 2022.

KIAM, L. C.; CHINNASAMY, B. Harnessing Apache Mahout to Link Content, *Code4Lib Journal*, [*S. l*], n. 22, 2013. Disponível em: https://journal.code4lib.org/articles/8912. Acesso em: 27 ago. 2022.

KIM, B. AI and Creating the First Multidisciplinary AI Lab. *In*: GRIFFEY, J, Artificial Intelligence and Machine Learning in Libraries: ALA; Library Technology Reports, [*S. l.*], v. 55, n. 1, 2019, p.16-20. Disponível em: https://digitalcommons.uri.edu/lib_ts_pubs/115/. Acesso em: 27 ago. 2022.

KIM, B. Machine learning for libraries and archives. *Online Searcher*, [*S. l.*], v. 1, n. 45, p. 39-41, 2021.

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.

KITCHENHAM, B. Procedures for performing systematic reviews. *Joint Technical Report SE Grou*p: Keele, Keele University, jul. 2004. Disponivel em: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf. Acesso em: 27 ago. 2022.

KOVAČEVIĆ, A. et al. Automatic extraction of metadata from scientific publications for CRIS systems. *Program-Electronic Library And Information Systems*, [*S. l.*], v. 45, n. 4, 2011, p. 376-396. Disponível em: https://www.researchgate.net/publication/216592386_Automatic_extraction_of_metadata_from_scientific_publications_for_CRIS_systems. Acesso em: 27 ago. 2022.

KRAGELJ, M.; BORŠTNAR, M. K. Automatic classification of older electronic texts into the Universal Decimal Classification-UDC. *Journal of Documentation*, [*S. l.*], v. 77, n. 3, 2020. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-06-2020-0092/full/html. Acesso em: 27 ago. 2022.

KRAUS-FRIEDBERG, C. Interactive Web Column: Machine Learning Algorithms, In and Out of Libraries. *Journal of Electronic Resources in Medical Libraries*, [*S. l.*], v. 16, n. 3-4, p. 111-115, 2019. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/15424065.2019.1700864?journalCode=werm20. Acesso em: 27 agp. 2022.

LAMBA, M.; MADHUSUDHAN, M. Mapping of ETDs in ProQuest Dissertations and Theses (PQDT) Global database (2014-2018). *Cadernos BAD*, Lisboa, n. 1, p. 169-185, 2019. Disponível em: https://brapci.inf.br/index.php/res/v/134567. Acesso em: 19 set. 2021.

LANEY, D. *3D Data management*: controlling data volume, velocity, and variety. *Gartner blog network*, [*S. l*] 2001. Disponível em: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Acesso em 3 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.

LITSEY, R.; MAULDIN, W. Knowing What the Patron Wants: Using Predictive Analytics to Transform Library Decision Making. *Journal of Academic Librarianship*, [*S. l.*], n. 44, p. 140-144, 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0099133317301866. Acesso em: 27 ago. 2022.

LORANG, E. et al. *Digital libraries, intelligent data analytics, and augmented description*: a demonstrations project: final report. University of Nebraska-Lincoln: Lincoln, 15 jun. 2020. Disponível em: https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf. Acesso em: 23 set. 2021.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada* – RITA, Porto Alegre, 2007, v. 14, n. 2. p. 43-67. Disponível em:

https://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67. Acesso em: 22 ago. 2022.

MARTINS, L. G. A. *Apostila de linguagem C*: (conceitos básicos). Uberlândia: Universidade Federal de Uberlândia, 2011. Disponível em: https://www.facom.ufu.br/~gustavo/ED1/Apostila_Linguagem_C. Acesso em: 27 ago. 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.

MICROSOFT INC. *Treine os modelos Keras em escala com o Azure Machine Learning*. [*S. l.*: *s. n.*], 27. ago. 2022a. Disponível em: https://docs.microsoft.com/pt-br/azure/machine-learning/how-to-train-keras. Acesso em: 27 ago. 2022.

MICROSOFT INC. *Tradução automática*. [*S. l.*: *s. n.*], 2022b. Disponível em: https://www.microsoft.com/pt-br/translator/business/machine-translation/. Acesso em: 27 ago. 2022.

MYSQL AB. *Manual de referência do MySQL 4.1*. [*S. l.*]: Oracle, 14 mar. 2010. Disponível em: https://downloads.mysql.com/docs/refman-4.1-pt.a4.pdf. Acesso em: 27 ago. 2022.

NEVES, B. C. As perspectivas e aplicações da computação cognitiva em unidades de informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 20 ENANCIB, 2019. **Anais**.... Disponível em: http://hdl.handle.net/20.500.11959/brapci/123114. Acesso em: 19 set. 2021.

NEVES, B. C. Inteligência artificial e computação cognitiva em unidades de informação: conceitos e experiências. *Logeion*: filosofia da informação, v. 7, p. 186-205, 2021. Disponível em: https://brapci.inf.br/index.php/res/v/147573. Acesso em: 19 set. 2021.

NLTK PROJECT. *NLTK*: documentation. [*S. l: s. n*.], 25 mar. 2022. Disponível em: https://www.nltk.org/. Acesso em: 27 ago. 2022.

OGURI, P. *Aprendizado de máquina para o problema de sentiment classification*. 2007. Dissertação (Mestrado em Informática) – Departamento de Informática, Pontifícia Universidade Católica, Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/9947/9947_5.PDF. Acesso em: 27 ago. 2022.

OLIVEIRA, H. C. R. de et al. Desenvolvimento de uma plataforma de software para o reconhecimento óptico de caracteres. *In*: CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 35., 2014, Natal. *Anais*... Natal: Sociedade Brasileira de Matemática Aplicada e Computacional, 2014, p. 1-2. Disponível em: https://www.researchgate.net/publication/266317024_Desenvolvimento_de_uma_Plataforma_de_Software_para_o_Reconhecimento_Optico_de_Caracteres. Acesso em: 27 ago. 2022.

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.

ORACLE. O *que é tecnologia Java e por que preciso dela*?. [*S. l.*]: Oracle, 2022a. Disponível em: https://www.java.com/pt-BR/download/help/whatis_java.html. Acesso em: 27 ago. 2022.

ORACLE. *MySQL Enterprise Edition*: guia do produto. [*S. l.*]: Oracle, 2022b. Disponível em: https://www.mysql.com/why-mysql/white-papers/mysql-enterprise-edition-guia-do-produto/. Acesso em: 27 ago. 2022.

ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS. *Información basíca sobre la COVID-19*. [*S. l.*]: OMS, 12 oct. 2020 Disponível em: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19. Acesso em 19 set. 2021.

PADILLA, T. *Responsible operations*: data science, machine learning, and AI in libraries. Dublin: OCLC, 2019. Disponível em: https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html. Acesso: 1 set. 2022.

PEDREGOSA, F. et al. Scikit-learn: machine learning in python. [*S. l*], *Journal of Machine Learning Research*, 2011, n. 12, p. 2825-2830. Disponível em: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 27 ago. 2022.

PERES, F. R. *O estudo da inteligência artificial no âmbito da Ciência da Informação*. 2017. 64 p. Dissertação (Mestrado em Ciência da Informação) – Universidade de Londrina, Londrina, 2017.

PEZZINI, A. Mineração de textos: conceito, processo e aplicações. *Revista Eletrônica do Alto Vale do Itajaí*, [*S. l*.], v. 5, n. 8, dez. p. 58-61. Disponível em: https://www.revistas.udesc.br/index.php/reavi/article/view/6750. Acesso em: 27 ago. 2022.

PYTHON SOFTWARE FOUDATION. *Python 2.0.1 license*. [*S. l.*]: Python Foudation, 2022a. Disponível em:

https://www.python.org/download/releases/2.0.1/license/. Acesso em: 25 ago. 2022.

PYTHON SOFTWARE FOUDATION. *History and license: history of the software*. [*S. l.*]: Python Foudation, 2022b. Disponível em: https://docs.python.org/3/license.html. Acesso em: 25 ago. 2022.

PINHEIRO, L. V. R.; FERREZ, H. D. *Tesauro brasileiro de Ciência da Informação*. Rio de Janeiro: IBCT, 2004. Disponível em: http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1. Acesso em: 22 ago. 2022.

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.

R FOUNDATION. *What is R?.* [*S. l.: s. n.*], 2022. Disponível em: https://www.r-project.org/about.html. Acesso em: 23 set. 2021.

RIDLEY, M. Explainable Artificial Intelligence. *Research Library Issues*, [*S. l.*], v. 299, p. 39-46, 2019. Disponível em: https://publications.arl.org/18nm1df/. Acesso em: 27 ago. 2022.

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.

SANTANA, A. F. et. Al. Combining Domain-Specific Heuristics for Author Name Disambiguation. *In*: IEEE/ACM JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL), 2014. Disponível em: https://ieeexplore.ieee.org/document/6970165. Acesso em: 27 ago. 2022.

SARACEVIC, T. Ciência da informação: origens, evolução e relações. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em: 28 set. 2020.

SCIKIT-LEARN INC. *About us*. [*S. l*]: Scikit-learn, 2022. Disponível em: https://scikit-learn.org/stable/about.html. Acesso em: 27 ago. 2022.

SCHOENENBERGER, H. *Lithium-ion batteries*: a machine-generated summary of current research: about this book. Suíça: Springer Cham, 2019. Disponível em: https://link.springer.com/book/10.1007/978-3-030-16800-1. Acesso em: 27 ago. 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, N.; NATHANSOHN, B. Análise da produção científica em Inteligência Artificial na área da Ciência da Informação no Brasil. *In*: ENANCIB, 19., 2018, Marília. *Anais* [...]. Marília: ANCIP, 2018. p. 111-216.

SALLES, T. et al. A Quantitative Analysis of the Temporal Effects on Automatic Text Classification. *Journal of the Association for Information Science and Technology*, [S. l]., v. 67, n. 7. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23452. Acesso em: 1 set. 2022.

SMIRAGLIA, R. P.; XIN, C. Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization. *Knowledge Organization*, [*S. l.*], v. 44, n. 3, p. 215-2333, 2017. Disponível em: https://www.researchgate.net/publication/318214976_Tracking_the_Evolution_of_Clustering_Machine_Learning_Automatic_Indexing_and_Automatic_Classification_in_Knowledge_Organization. Acesso em: 27 ago. 2022.

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.

RANGANATHAN, S. R. *As cinco leis da biblioteconomia*. Briquet de Lemos: Brasília, 2009.

SIDDIQUI, A.; MISHRA, N.; VERMA, J. S. A survey on automatic image annotation and retrieval. *International Journal of Computer Applications*, [*S. l.*], v. 118, n. 20, may, 2015. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.2043&rep=rep1&type=pdf. Acesso em: 1 set. 2022.

TABLEAU SOFTWARE. *Alfabetização de dados*: com uma força de trabalho alfabetizada em dados, sua organização e sua cultura prosperarão. [*S. l*]: Salesforce, 2022.

TANIGUCHI, S. Duplicate bibliographic record detection with an OCR-converted source of information. *Journal of Information Science*, [*S. l.*], 2012. Disponível em: https://journals.sagepub.com/doi/10.1177/0165551512459923. Acesso em: 22 ago. 2022.

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

TAURION, C. *Big data*. Rio de Janeiro: Brasport, 2013. *E-book*.

TECNOLOGIA. *In*: Michaelis Dicionário Brasileiro da Língua Portuguesa. [*S. l*]: Melhoramentos, [2021]. Disponível em: https://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=tecnologia. Acesso em: 28 set. 2021.

THOMÉ, A. G. *Fundamentos sobre processamento de imagens*. Rio de Janeiro: UFRJ, 2004. 76 slides. Disponível em: http://hpc.ct.utfpr.edu.br/~charlie/docs/PID/PID_AULA_01.pdf. Acesso em: 1 set. 2022.

THOMPSON, R.; SAFER, K. VIZINE-GETZ, D. Evaluating Dewey concepts as a knowledge base for automatic subject assignment. In: Proceeding of the second ACM INTERNACIONAL CONFERENCE ON DIGITAL LIBRARIES, jul. 1997. Disponivel em: https://dl.acm.org/doi/10.1145/263690.263790. Acesso em: 1 set. 2022.

THOMSON REUTERS. *Web of Science* (WoS). [*S. l*]: Thomson Reuters, 2021.

TOEPFER, M.; SEIFERT, C. Fusion architectures for automatic subject indexing under concept drift Analysis and empirical results on short texts. *International Journal on Digital Libraries*, [*S. l*.], n. 21, 2018, p. 169-189. Disponível em: https://link.springer.com/article/10.1007/s00799-018-0240-3. Acesso em: 27 ago. 2022.

TURING, A. M. Computing machinery and intelligence. *Mind*, Inglaterra, v. 59, n. 236, oct. 1950, p. 433-460. Disponível em: https://academic.oup.com/mind/article/LIX/236/433/986238. Acesso em: 27 ago. 2022.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. *Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação* (Brapci). Porto Alegre: UFRGS; UFPR, 2021.

UNIVERSIDADE WAIKATO. *Weka 3*: machine learning software in Java. [New Zealand]: Universidade Wikato, 2022. Disponível em: https://www.cs.waikato.ac.nz/~ml/weka/index.html. Acesso em: 27 ago. 2022.

VELOSO, A. et al. Cost-effective on-demand associative author name disambiguation. *Information Processing & Management*, [*S. l*.], v. 48, n. 4, 2012, p. 680-697. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457311000847. Acesso em: 27 ago. 2022.

VIEIRA, R.; LOPES, L. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. *Linguagens especializadas em corpora*: modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPURS, 2010, p. 183-201. Disponível em: https://editora.pucrs.br/edipucrs/acessolivre//livros/linguagensespecializadasemcorpora.pdf. Acesso em: 27 ago. 2022.

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.

WAGSTAFF, K. L.; LIU, G. Z. Automated Classification to Improve the Efficiency of Weeding Library Collections. *Journal of Academic Librarianship*, [*S. l.*], v. 44, n. 2, mar. 2018, 238-247. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0099133317304160. Acesso em: 27 ago. 2022.

WALKER, K.; JIANG, Z. Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. *Journal of Academic Librarianship*, [*S. l.*], v. 45, n. 3, mar. 2019. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0099133319300436. Acesso em: 27 ago. 2022.

YELTON, A. HAMLET: Neural-Net-Powered Prototypes for Library Discovery. In: GRIFFEY, J. *Artificial Intelligence and Machine Learning in Libraries*: ALA; Library Technology Reports, [*S. l.*], v. 55, n. 1, p. 10-15, 2019. Disponível em: https://journals.ala.org/index.php/ltr/article/view/6909/9301. Acesso em: 27 ago. 2022.