8º EBBC
Encontro Brasileiro de Bibliometria e Cientometria
MACEIÓ2022
20, 21 E 22 • JULHO

REFERÊNCIAS A ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA CIÊNCIA DA INFORMAÇÃO:
uma análise descritiva a partir da *Web of Science*

Dalton Lopes Martins[1]

Resumo: O artigo estuda as referências de algoritmos de aprendizagem de máquina em documentos indexados na área da Ciência da Informação na base *Web of Science*. Avalia 3111 documentos identificados e conclui que a abordagem supervisionada por meio de técnicas de classificação é a mais utilizada no campo, com evidência para os algoritmos de *Support Vector Machine, Decision Tree, Random Forest.*

Palavras-Chave: Algoritmos. Aprendizagem de Máquina. KNIME. Ciência da Informação. Web of Science.

1 INTRODUÇÃO

A pesquisa e o desenvolvimento de aplicações de aprendizagem de máquina (AM) têm se tornado um fenômeno contemporâneo com impactos nas mais diversas áreas do conhecimento, não sendo diferente no campo da Ciência da Informação (CI). A pesquisa do tema no Brasil na CI ainda é bastante incipiente. A base BRAPCI apresenta apenas 6 documentos quando se realiza a pesquisa com o termo "aprendizagem de máquina"[2] em todos os campos de busca. No entanto, importantes iniciativas têm surgido estimulando a experimentação e desenvolvimento de novos serviços de informação mundo afora. Chama atenção algumas iniciativas internacionais, tais como o relatório encomendado pela *Library of Congress* (LoC) (CORDELL, 2020) intitulado "*Machine learning+Libraries: a report on the state of the field*" que procura apresentar as tendências de pesquisa e aplicações de AM que podem impactar os produtos e serviços ofertados pela biblioteca. Como desdobramento desse primeiro estudo, a LoC realizou um projeto demonstrativo (LORANG *et al.,* 2020) de aplicações de diversas técnicas e algoritmos de AM em diferentes bases de dados da biblioteca demonstrando resultados de alto potencial para o campo. Outra ação que demonstra

[1] Universidade de Brasília (UnB)
[2] Link da pesquisa: https://brapci.inf.br/?q=%22aprendizagem+de+m%C3%A1quina%22&type=1&year_s=1972&year_e=2022&order=0. Acesso em: 30 jan. 2022.

interesse crescente na área é a declaração da IFLA (2020) denominada "*Statement on Libraries and Artificial Intelligence*" que evidencia o potencial das aplicações que utilizam AM para as bibliotecas bem como ressalta preocupações éticas e políticas que necessitam ser levadas em consideração. Para finalizar os exemplos das iniciativas na área da CI, a rede denominada "*The Museums + AI Network*" (MURPHY; VILLAESPESA, 2020) apresenta um conjunto técnico de orientações para museus de como lidar com seus dados para a realização de projetos que podem se beneficiar dos algoritmos de AM para a geração de novas possibilidades de uso dos dados de suas coleções. Há diversos outros exemplos que poderiam aqui ser citados.

O presente artigo é parte de um esforço maior de pesquisa envolvendo vários estudos em andamento visando compreender melhor como a área da CI tem aplicado as técnicas de AM em seus problemas característicos, que questões têm sido priorizadas, que resultados têm sido obtidos, os cuidados a serem tomadas, as ferramentas e técnicas utilizadas. No presente trabalho, o objetivo consiste em analisar quais são os algoritmos de AM mais mencionados nos artigos científicos em nível internacional da área da CI. Entende-se que conhecer os algoritmos mais mencionados, como estão sendo usados e suas aplicações é uma etapa importante para compreender o estado da arte do tema na CI.

2 METODOLOGIA

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

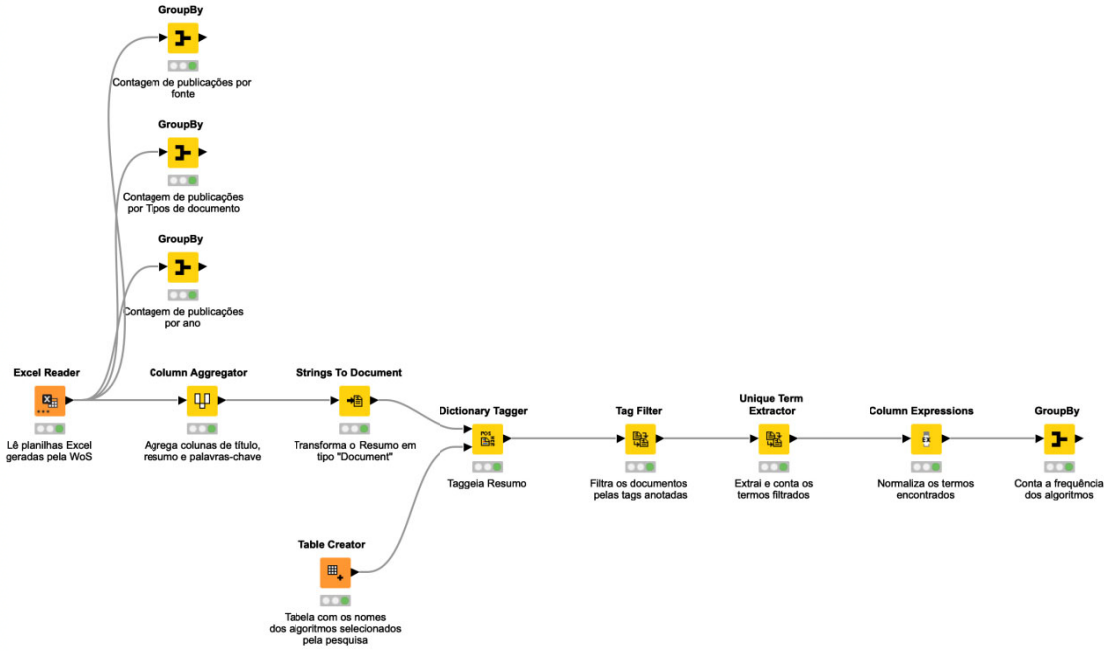
A WoS permite exportar os arquivos com os resultados de busca em diferentes formatos. Os dados foram baixados em formato Excel e foram tratados utilizando o aplicativo KNIME[3]. O KNIME é um aplicativo para a construção visual de fluxos de processos de ciência de dados, disponibilizando recursos para as etapas de coleta, pré-processamento e modelagem de dados

[3] Disponível em: https://www.knime.com/. Acesso em: 30 jan. 2022.

utilizando técnicas de AM. O fluxo criado para a presente pesquisa é apresentado na figura 1 e disponível para acesso no Github[4].

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

Figura 1 - Fluxo de processos de tratamento de dados implementados no KNIME



Fonte: dos autores.

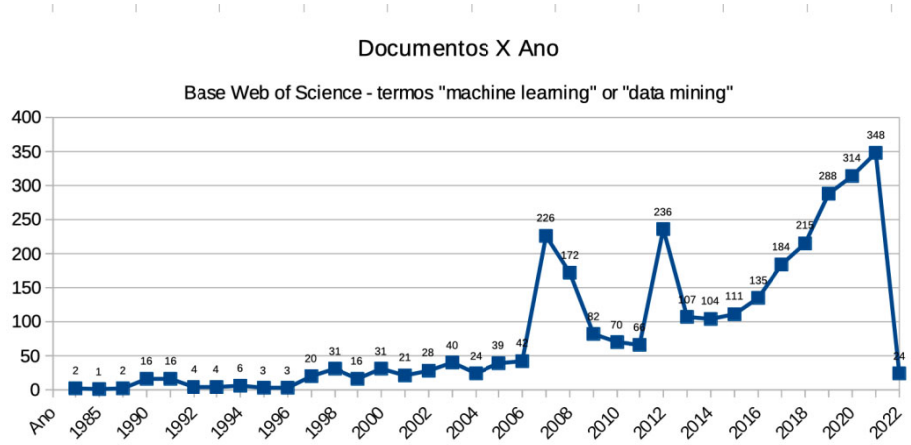
[4] Disponível em: https://github.com/daltonmartins/aprendizagemdemaquina_aplicada_na_cienciadainformacao/blob/main/EBBC%202022.knar.knwf. Acesso em: 30 jan. 2022.

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

3 RESULTADOS

Foram identificados 3111 documentos a partir dos critérios da pesquisa mencionados na seção Metodologia. A distribuição do número de documentos por ano pode ser vista na figura 2. Pode-se notar que as primeiras menções aparecem no ano de 1984, permanecendo em número de pequeno volume até o ano de 1997. A partir de 1998 ocorre um primeiro salto na quantidade de documentos que se mantém constante até aproximadamente 2006. A partir desse período, observa-se dois picos importantes nos anos de 2007 e 2013 e um crescimento linear e contínuo do ano de 2014 em diante. Cabe ressaltar que o ano de 2022 apresenta baixa quantidade de documentos pelo fato da pesquisa ter sido realizada no primeiro mês do ano. Optou-se por incluir esses documentos para não perder as tendências mais recentes de menção dos algoritmos de AM.

Figura 2 - Distribuição dos documentos X Ano



Fonte: dos autores.

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

que mais apresenta menções, indicando potencialmente que o campo da cientometria pode ser uma das áreas da CI que mais tem se beneficiado dos algoritmos de AM.

Tabela 1 - As 10 principais fontes de documentos

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

Fonte: dos autores.

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

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

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

Fonte: dos autores.

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

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

4 CONCLUSÃO

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

REFERÊNCIAS

CORDELL, Ryan. **Machine learning + libraries:** a report on the state of the field. Washington: Biblioteca do Congresso Americano, 2020. 97 p. Disponível em: http://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf. Acesso em: 30 jan. 2022.

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

LORANG, Elizabeth; LEEN-KIAT, Soh; YI, Liu; CHULWOO, Pack. **Digital libraries, intelligent data analytics, and augmented description:** a demonstration project. Nevada: University of Nevada, 2020. 47 p. Disponível em: https://digitalcommons.unl.edu/libraryscience/396/. Acesso em: 30 jan. 2022.

MURPHY, Oonagh; VILLAESPESA, Elena. **AI:** a museum planning toolkit. Londres: University of London, 2020. 15 p. Disponível em: https://research.gold.ac.uk/id/eprint/28201/. Acesso em: 30 jan. 2022.

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

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