Tendências da Pesquisa Brasileira em Ciência da Informação

USO DO APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO AUTOMÁTICA DE DOCUMENTOS DE ARQUIVO: EXPERIMENTO INICIAL EM UMA ORGANIZAÇÃO PÚBLICA[1]

USE OF MACHINE LEARNING FOR THE AUTOMATIC CLASSIFICATION OF ARCHIVE DOCUMENTS: INITIAL EXPERIMENT IN A PUBLIC ORGANIZATION

Eduardo Watanabe[2]
Renato Tarcíso Barbosa de Sousa[3]

Resumo: A evolução recente das tecnologias leva à seguinte pergunta de pesquisa: o aprendizado de máquina pode contribuir com a classificação automática de documentos de arquivo de uma organização pública? Os procedimentos metodológicos consistem na revisão de literatura e nas tarefas propostas pelo modelo CRISP-DM em um experimento com 4.800 documentos, divididos em 24 classes. Foram desenvolvidos 20 (vinte) modelos de aprendizagem supervisionada aplicados a três vocabulários criados (nomes de pessoas, lugares e tempo). O melhor resultado foi o *F1 score* de 0,870. É proposto um subprocesso específico para trabalhar o espaço de aperfeiçoamento do modelo de classificação com base na Ciência da Informação e Arquivologia.

Palavras-Chave: Aprendizado de Máquina. Processamento de Linguagem Natural. Classificação automática de documentos. Gestão de documentos.

Abstract: *The recent evolution of technologies leads to the following research question: can machine learning contribute to the automatic classification of records in a public organization? The methodological procedures consist of a literature review and tasks proposed by the CRISP-DM model in an experiment with 4,800 documents, divided into 24 classes. Twenty supervised learning models were developed and applied to three created vocabularies (names of people, places and time). The best result was the F1 score of 0.870. A specific subprocess is proposed to work on the space for improving the classification model based on Information Science and Archival Science.*

Keywords: *Machine Learning. Natural Language Processing. Automatic classification os records. Records management.*

[1] Trabalho submetido, avaliado, aprovado e apresentado no XXII Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação - ENANCIB.

[2] Doutorando em Ciência da Informação (UnB). Advogado da Advocacia Geral da União (AGU). E-mail: edwatanabe@protonmail.com. ORCID: https://orcid.org/0000-0002-7576-2793.

[3] Doutor em História Social (USP). Docente da Universidade de Brasília (UnB). E-mail: renasou@unb.br. ORCID: https://orcid.org/0000-0002-5647-7903.

1 INTRODUÇÃO

Adrian Cunninghan (2021) fez uma revisão dos 30 anos de experiência dos profissionais de arquivo na Austrália frente aos desafios da transformação digital. Na sua avaliação, a gestão de documentos digitais continua a deteriorar-se não obstante algum progresso possa ser reconhecido. O otimismo vislumbrado com a tecnologia no início da década de 1990 não se concretizou, ele não considera realistas as soluções rápidas e fáceis ou do tipo "bala de prata".

Trace (2022) alerta para o longo decurso de tempo até que os documentos de arquivo sejam disponibilizados para pesquisa. O acúmulo de tarefas de tratamento documental é designado *backlog*, que consiste na interrupção da distribuição e consumo final do processo de pesquisa, e pode ser considerado então um problema de infraestrutura de conhecimento.

Nesta pesquisa, destacamos dentre as sete funções arquivísticas a de classificação de documentos tendo em vista que a sua automação pode ser muito útil para reduzir o tempo de tratamento de documentos, ainda mais em uma realidade de aumento dos acervos e limitação de recursos humanos (Lee, 2018).

A transformação digital tem reconfigurado os arquivos no sentido deles passarem a ser considerados como conjuntos de dados a serem minerados (Moss; Thomas; Gollins, 2018). Com isso, abrem-se espaços para a automação em grande escala com o uso da Inteligência Artificial (IA) tanto de atividades tradicionais de gestão de documentos como de experimentos inovadores para capturar, organizar e acessar documentos (Colavizza *et al.*, 2021).

O Aprendizado de Máquina é um subcampo da IA que remonta à expressão em inglês *Machine Learning*, criada por Arthur Samuel em 1959. O conceito mais contemporâneo de Aprendizado de Máquina consiste em um conjunto de técnicas que se utilizam da indução, uma forma de inferência lógica que busca obter conclusões

genéricas sobre um conjunto particular de exemplos (Monard; Baranauskas, 2003).

Não obstante todo o potencial do Aprendizado de Máquina a partir dos avanços tecnológicos de processamento e armazenamento de dados nos últimos anos, ainda há poucas pesquisas no Brasil sobre a sua aplicação em unidades de informação (Monterei; Lopes, 2021) ou sobre Inteligência Artificial (Pinheiro; Oliveira, 2022). Na mesma situação temos o Processamento da Linguagem Natural, considerada um elo da Ciência da Computação com a Ciência da Informação, mas que também tem sido pouco utilizada na pesquisa de CI (Falcão; Lopes; Souza, 2022).

Um dos desafios consiste em como a CI pode identificar o seu "lugar" nos estudos de IA (Silva; Nathansohn, 2018). Nesse contexto é que formulamos a seguinte pergunta de pesquisa: em que medida o aprendizado de máquina pode contribuir com a classificação automática de documentos de arquivo?

O objetivo geral do trabalho consiste em identificar as contribuições de aplicações práticas de Aprendizagem de Máquina em arquivos na literatura e, em seguida, realizar um experimento com documentos de arquivo de uma organização pública, no caso a Advocacia-Geral da União (AGU).

2 DESENVOLVIMENTO

A presente pesquisa é de natureza mista (quantitativa e qualitativa), com abordagem exploratória e aplicada, feita em ambiente de estudo natural e com horizonte de tempo transversal. Os métodos da pesquisa são a revisão de literatura o estudo de caso, a análise documental e o *Design Science*.

Como primeiro objetivo específico de pesquisa será feita a revisão da literatura. O segundo objetivo específico consiste em aplicar algoritmos de Aprendizagem de Máquina supervisionados que façam a classificação automatizada de documentos. A

pesquisa, coleta e análise dos dados abrangeu o período de abril a maio de 2022 e de março a abril de 2023.

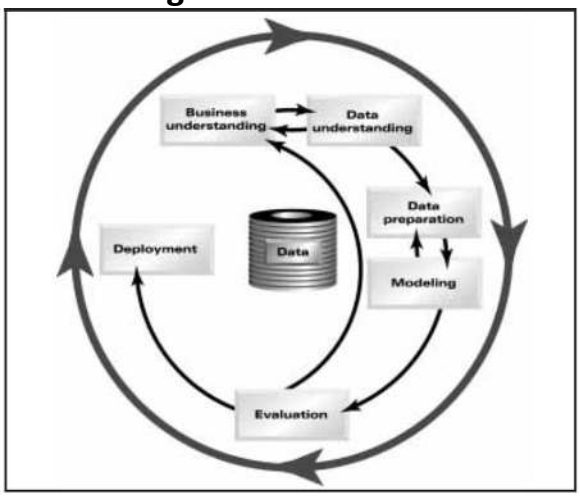
O levantamento bibliográfico incluiu as bases de dados da *Library & Information Science Abstracts* (LISA) (de 1969 até 24.03.2023 exclusivo com itens revisados por especialistas) e a Base de Dados em Ciência da Informação (BRAPCI) (de 1972 até 24.03.2023). Os operadores de pesquisa utilizados foram recordkeeping AND ("artificial intelligence" OR "machine learning"); "records management" AND ("artificial intelligence" OR "machine learning"); "gestão de documentos" AND ("inteligência artificial" OR "aprendizado de máquina"); "archival science" AND ("artificial intelligence" OR "machine learning"); arquivologia AND ("inteligência artificial" OR "aprendizado de máquina") e "computational archival science" OR "arquivologia computacional".

Foram localizados n = 215 itens, dos quais 200 foram excluídos por não serem específicos, restando ao final 15. Os 200 itens não específicos abrangem a menção incidental de um termo de pesquisa ou de todos, podemos exemplificar com artigos com o tema principal da tecnologia *blockchain* em que são feitas menções incidentais sobre gestão de documentos e IA, sem que tratem de pesquisas relacionando os conceitos entre si como é o objeto deste trabalho. Com a leitura dos artigos foi possível acessar outras 19 referências específicas que somaram 34 no total.

Os procedimentos metodológicos consistem nas tarefas propostas pelo *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que consiste em metodologia e modelo de processo de mineração de dados criado em 2000 por um consórcio de empresas (Shearer, 2000). A revisão sistemática da literatura feita por Schröer, Kruse e Gómez (2021) sobre o CRISP-DM demonstrou que ele continua como o modelo padrão de mineração de dados ainda hoje.

O CRISP-DM é composto por seis fases que se desenvolvem como um ciclo em que as lições aprendidas durante o processo de mineração de dados podem gerar questões de negócio ainda mais focadas.

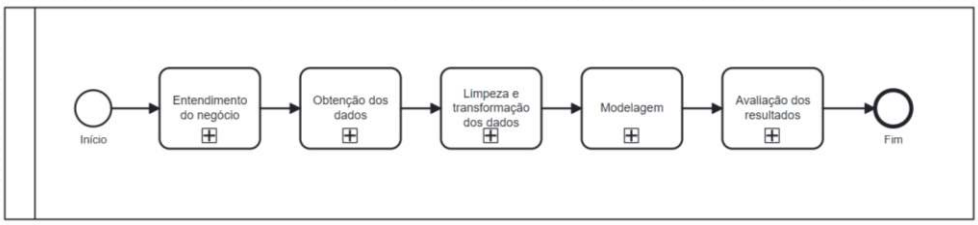
Figura 1: CRISP-DM



Fonte: Shearer (2000).

Como este trabalho se limita à pesquisa dos modelos, serão utilizadas somente as cinco primeiras fases do CRISP-DM, que são representadas na Figura 2.

Figura 2: Diagrama simplificado com subprocessos do experimento



Fonte: elaborada pelos autores.

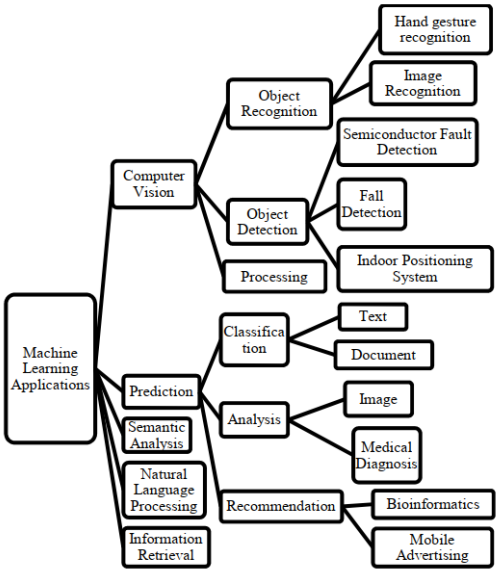
3 APRENDIZADO DE MÁQUINA NA ARQUIVOLOGIA E CLASSIFICAÇÃO DE DOCUMENTOS

De acordo com a Associação Brasileira de Normas Técnicas (ABNT), NBR ISO 15489-1/2018, a classificação de documentos consiste na "[...] identificação

sistemática e/ou configuração de atividades de negócio e/ou documentos de arquivo em categorias de acordo com convenções, métodos e regras estruturadas logicamente".

Sobre as aplicações de Aprendizado de Máquina, Shinde e Shah (2018) propuseram uma tipologia a partir da revisão da literatura.

Figura 3: Aplicações de Aprendizado de Máquina



Fonte: Shinde e Shah (2018).

Para Lee (2018), o Aprendizado de Máquina confere aos *softwares* o aperfeiçoamento progressivo de desempenho sem a necessidade de uma programação explícita, uma vez que o aprendizado constrói e refina um modelo estatístico baseado em dados de treinamento. Já o Processamento de Linguagem Natural (NLP na sigla da expressão em inglês) é muito útil para identificar e organizar informações do contexto relacionadas ao documento (Lee, 2018).

Para Amozurrutia e outros (2022) o uso das ferramentas automatizadas não deve ter a pretensão de substituir a atividade humana, mas sim apoiar esse trabalho. Eles então propõem cinco linhas relevantes para a elaboração de diretrizes de uso de

ferramentas de aprendizagem automatizada para auxiliar os arquivistas na avaliação de documentos e identificação de dados pessoais. São elas a seleção de métricas apropriadas de avaliação dos resultados obtidos, a curadoria dos documentos processados de modo a evitar que a anonimização dos dados pessoais venha a prejudicar futuras análises, utilizar ferramentas com flexibilidade que permitam a inspeção manual do resultado dos documentos processados, desenvolver ferramenta que possa incorporar a retroalimentação feita por especialistas de forma interativa e prover adaptabilidade da ferramenta para que possa ser utilizada em outros documentos da mesma instituição.

A diretrizes apresentadas servem para fazer frente a fatores relacionados à aplicação da IA, tais como a ausência de normas aplicáveis para o seu uso, opacidade no desenvolvimento e funcionamento dos algoritmos utilizados; falta de informação para que leigos possam compreender o funcionamento dos algoritmos, falta de modelos e aplicações de IA no idioma espanhol e a revisão dos modelos de IA e dos vieses humanos de IA (Amozurrutia *et al.*, 2022).

O experimento da presente pesquisa aplicará algoritmos de Aprendizado de Máquina de "Predição-Classificação de Documentos" combinada com o Processamento da Linguagem Natural na classificação de documentos de arquivo.

Em um cenário de digitização, *big data* e avanços tecnológicos, Nathaniel Payne (2018) propôs a criação de um novo campo de estudos transdisciplinar, a *Computational Archival Science* (CAS), fundada na Arquivologia, Ciência da Informação e Ciência da Computação. A CAS consiste na aplicação de métodos e recursos computacionais, padrões de design, constructos técnico-sociais e interação homem-máquina aplicada ao processamento de documentos e arquivos em grande escala (*big data*), análise, armazenamento, preservação de longo prazo e problemas de acesso. Os objetivos da CAS são aperfeiçoar e otimizar a eficiência, autenticidade, veracidade,

procedência, produtividade, computação, estrutura e design informacional, precisão e interação homem-máquina em apoio à aquisição, avaliação, arranjo e descrição, preservação, comunicação, transmissão, análise e decisões de acesso.

Payne (2018) destaca que os pesquisadores da CAS até então centraram esforços em áreas como análise de arquivos com *text-mining* e *data-mining* aplicada em serviços de avaliação, arranjo e descrição. E como tendências futuras estariam o uso do Aprendizado de Máquina, incluindo *deep learning*, pesquisas para compreensão da linguagem natural com o uso de análise de textos com IA.

Colavizza e outros (2021) revisaram na literatura a intersecção entre a IA e a Arquivologia sob a ótica do modelo *Records Continuum*, no que identificaram como temas as considerações teóricas e profissionais, a automação de processos de gestão de documentos, a organização e acesso aos arquivos, e os formatos inovadores de arquivos digitais. Os autores concluem como tendências emergentes e direções para trabalhos futuros a aplicação dos princípios de gestão de documentos aos próprios dados e processos com o uso da IA, bem como de uma integração estrutural da IA com os sistemas de gestão de documentos e a sua prática.

A partir da revisão da literatura, identificamos experimentos que apresentaram resultados quantitativos agora organizamos no Quadro 1 e na Tabela 1, todos eles são de aprendizagem supervisionada (classificação). Alguns experimentos efetuaram testes combinando diferentes parâmetros com a geração de muitos resultados; como o objetivo da tabela não é trazer detalhes muito específicos de todas as combinações utilizadas, optamos por selecionar o melhor resultado alcançado apenas para efeito ilustrativo.

Quadro 1: Pesquisas quantitativas relatadas na literatura: dados básicos

Referência	Instituição envolvida	Descrição	Ano
Marcus, 2002; Shinkle, 2017	*National Archives and Records Administration (NARA)*	Classificar automaticamente documentos e arquivá-los no sistema de acordo com a classificação atribuída.	2001
Warland; Mokhtar, 2013	*Universidade de Richmond (Virginia)*	Classificar automaticamente e-mails	2011
Vellino e Alberts, 2016	Pesquisa acadêmica com voluntários consultores de gestão de informação	Classificar automaticamente e-mails no processo de avaliação se eles possuem ou não valor para o negócio	2016
Rolan *et al.*, 2019	*New South Wales State Archives*	Automatizar a avaliação de documentos acordo com tabela de temporalidade da instituição.	2017
Rolan *et al.*, 2019; Public Record Office Victoria, 2018	*Public Record Office Victoria*	Identificar o formato do arquivo de e-mails para redução do volume de documentos a serem avaliados.	2018
Hutchinson, 2018	*University of Saskatchewan Associate VicePresident for Information and Communications Technology*	Classificar automaticamente documentos de recursos humanos que contenham informações pessoais individualizadas.	2020
Wang *et al.*, 2021	*Archives of Liaoning Province*	Classificação de itens do catálogo de dados conforme uma das 11 categorias previstas na Lei Chinesa de Classificação de Arquivos	2021
Tkachenko; Denisova, 2022	*Siberian State Automobile and Highway University*	Classificar automaticamente os documentos de uma universidade em quatro classes.	2022

Fonte: elaborado pelos autores.

Tabela 1: Pesquisas quantitativas relatadas na literatura: resultados

Referências	Dataset	Algoritmos utilizados	Acurácia	F1 score
Marcus, 2002; Shinkle, 2017	n/d	n/d: utilizou aplicação proprietária *AutoRecords* da TrueArc	**96,0%**	**n/d**
Warland; Mokhtar, 2013	n/d	n/d: menciona que utilizou aplicação de e-Discovery com revocação de 76,7% e precisão de 84,7%	**n/d**	**n/d**
Vellino e Alberts, 2016	1.023 e-mails	*Support Vector Machine* (SVM)	**98,0%***	**0,98***
Rolan *et al.*, 2019	8.784 documentos	*Multinomial Naïve Bayes* e *Multi-Layer Perceptron* (MLP)	**84,0%**	**0,835**

Rolan *et al.*, 2019; Public Record Office Victoria, 2018	4,6 milhões de e-mails	n/d: utilizou aplicação proprietário da Nuix	**98% a 100%**	**n/d**
Hutchinson, 2018	1.784 documentos	*Multinomial Naïve Bayes*	**90,4%***	**0,983***
Wang *et al.*, 2021	96.680 itens do catálogo	*Support Vector Machine* (SVM) e *Network Analysis*	**n/d**	**0,716**
Tkachenko; Denisova, 2022	1.778 documentos	Híbrido de *Support Vector Machine* (SVM) e *k-nearest neighbor* (kNN)	**n/d**	**0,983***

Fonte: elaborada pelos autores.

4 ENTENDIMENTO DO NEGÓCIO E DOS DADOS

A Advocacia-Geral da União (AGU) é órgão da administração pública federal constituída como função essencial à Justiça pela Constituição Federal de 1988. Desde 2014 a AGU utiliza o Sistema AGU de Inteligência Jurídica (SAPIENS), que é um Sistema de Gestão Arquivística de Documentos (SIGAD), responsável por gerenciar todos os processos em meio físico, digital e híbrido.

No SAPIENS são utilizados de forma obrigatória o Código de classificação e tabela de temporalidade e destinação de documentos relativos às atividades-meio do Poder Executivo Federal e o Código de Classificação e a Tabela de Temporalidade e Destinação dos Documentos de Arquivo relativos às atividades-fim da AGU. O SAPIENS possui um módulo Arquivista, no qual é feita a gestão das transições arquivísticas em painel próprio com listagem dos processos com informações do código de classificação e do prazo de guarda previsto.

Os metadados dos processos do SAPIENS são armazenados de forma estruturada em tabelas em Banco de Dados *Oracle*, o que possibilitou a extração de dados por meio de script em linguagem *Structured Query Language* (SQL). Devido à existência de muitas tabelas, foram selecionados só os atributos necessários para compor o conjunto de dados de pesquisa: identificador do processo, código de classificação e respectivo nome fase e número identificador de cada documento que compõe o processo.

Foram também utilizados os seguintes filtros para que o conjunto de dados atenda aos objetivos da pesquisa:

- Limitação aos processos da espécie Administrativo e dos códigos de classificação das atividades-meio da AGU processos das atividades-meio, uma vez que os documentos das atividades finalísticas incluiriam um universo de documentos mais heterogêneo e com maior complexidade para um estudo inicial por terem uma característica de maior homogeneidade em comparação com os documentos das atividades-fim;
- Limitação aos processos na fase de Arquivo Intermediário por já terem sido encerrado, que já foram classificados por um profissional do setor de arquivo e que não possuem mais documentos a serem juntados;
- Limitação aos documentos produzidos por usuário da própria AGU e que utilizaram o editor de textos do SAPIENS, que salva o documento em formato padronizado HTML. O objetivo é reduzir a complexidade, tendo em vista que se presume maior padronização nos documentos produzidos pela AGU em comparação com os documentos produzidos por pessoas externas. Outra vantagem consiste na maior qualidade de dados do formato HTML em comparação com a extração de texto em arquivos em formato PDF ou imagem.
- Limitação aos processos com código de classificação que somaram 200 ou mais documentos, tendo em vista a necessidade de utilizar um *dataset* equilibrado no que se refere à composição das classes.

Quadro 2: Códigos de classificação selecionados para o experimento

Código	Descritor do código
020.5	Assentamentos individuais. Cadastro
022.111	Propostas, Estudos, Editais, Programas, Relatórios finais, Exemplares únicos de exercícios, Relação de participantes, Avaliação e controle de expedição de certificados
023.03	Reestruturações e alterações salariais, Ascensão e progressão funcional, Avaliação de desempenho, Enquadramento, Equiparação, Reajuste e reposição salarial, Promoções
023.11	Admissão. Aproveitamento. Contratação. Nomeação. Readmissão. Readaptação. Recondução. Reintegração. Reversão
023.13	Lotação. Remoção. Transferência. Permuta
023.15	Requisição. Cessão
024.123	[ADICIONAIS] Noturno
024.2	Férias
026.13	Aposentadoria
029.6	Ações trabalhistas. Reclamações trabalhistas
033.2	Material de consumo
033.21	Compra
040	Patrimônio. Normas, regulamentações, diretrizes, procedimentos, estudos e/ou decisões de caráter geral
041.3	Desapropriação. Reintegração de Posse. Reivindicação de domínio. Tombamento
042.5	Acidentes. Infrações. Multas
050	Orçamento e finanças. Normas, regulamentações, diretrizes, procedimentos, estudos e/ou decisões de caráter geral
052.1	Programação financeira de desembolso
052.22	Despesa
059.1	Tributos (impostos e taxas)
063.2	Protocolo: recepção, tramitação e expedição de documentos
063.61	Análise. Avaliação. Seleção
070	Comunicações. Normas, regulamentações, diretrizes, procedimentos, estudos e/ou decisões de caráter geral
992	Comunicados e informes
995	Controle de frequência. Livros. Cartões. Folhas de ponto. Abono de faltas. Cumprimento de horas extras

Fonte: elaborado pelos autores.

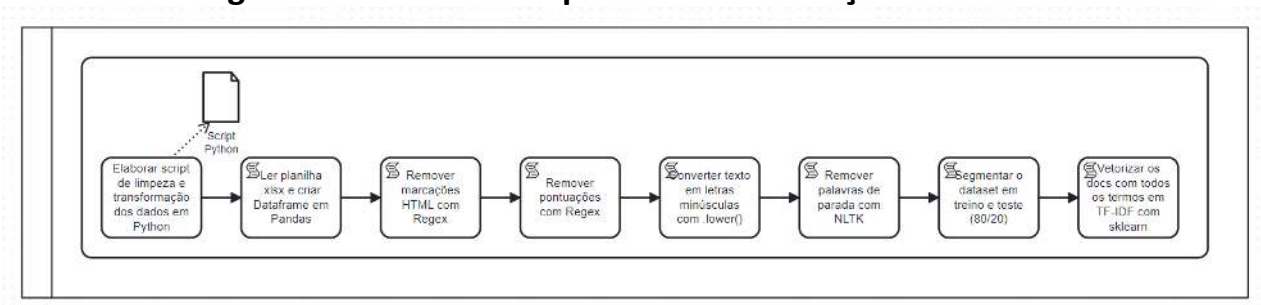
Já o conteúdo dos documentos digitais produzidos pela AGU é indexado em instância do *Elasticsearch*, o que permitiu a obtenção dos dados a partir de *script* em linguagem *Python* com o uso da biblioteca *Elasticsearch*. Ao final os dados foram consolidados em planilha em formato Excel com os dados de 4.800 documentos digitais no total, separados em 24 códigos de classificação com 200 documentos cada.

5 PREPARAÇÃO DOS DADOS

Devido às extrações de dados terem sido bem direcionadas para os objetivos da pesquisa, não foi necessária a execução de tarefas relacionadas a eliminação manual de atributos, integração de dados e transformação de dados.

No que se refere às tarefas de limpeza e transformações de dados necessárias para melhorar a aplicação futura do Aprendizado de Máquina, foram utilizados algoritmos para limpeza e tratamento de dados em linguagem *Python*, com o uso das bibliotecas *Pandas*, *Natural Language Toolkit* (NLTK) e *scikit-learn* (*sklearn*), no *software PyCharm Community Edition*. Foi adotada a vetorização dos termos dos documentos com TF-IDF, abreviação da expressão em inglês para frequência do termo–inverso da frequência nos documentos, que foi utilizado nos experimentos pesquisados como por Tkachenko e Denisova (2022). A vetorização utilizou termos simples (*ngram*=1) para otimizar a execução dos treinamentos com os diferentes modelos e tornar menos complexa a análise qualitativa de vocabulário feita em etapas posteriores.

Figura 4: Tarefas de limpeza e transformação dos dados



Fonte: elaborada pelos autores.

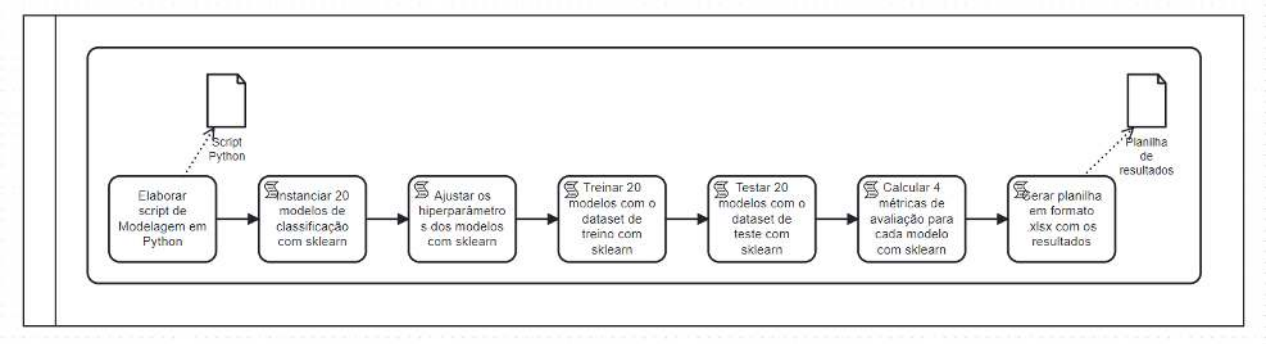
6 MODELAGEM E RESULTADOS

Na presente pesquisa buscamos a partir unicamente do texto de documentos de um processo predizer por algoritmos de Aprendizado de Máquina a classificação do

documento de arquivo. Tais algoritmos são modelos baseados em cálculos estatísticos, que buscam identificar padrões nos dados fornecidos e então mapear os resultados desejados (Rolan *et al.*, 2019).

Foram utilizados os modelos de classificação disponibilizados na biblioteca *sklearn* para o treinamento e teste conforme os procedimentos ilustrados na Figura 5.

Figura 5: Procedimentos de Modelagem



Fonte: elaborada pelos autores.

Um dos principais pontos fortes do CRISP-DM, adotado nessa pesquisa, é ser um processo iterativo de refinamento contínuo em que a execução das fases não fica engessada em uma sequência pré-determinada. Desde os primeiros resultados gerados foi possível identificar espaços para aperfeiçoamento do uso dos modelos.

A partir do uso de ferramentas de IA pelo governo da França para organizar as informações das mais 1,5 milhões de contribuições feitas no "Grande debate nacional" de 2019, Chabin (2020) demonstra como a utilização de conhecimentos arquivísticos e análise diplomática podem enriquecer o *corpus* de documentos de modo a melhorar o desempenho dos modelos de IA.

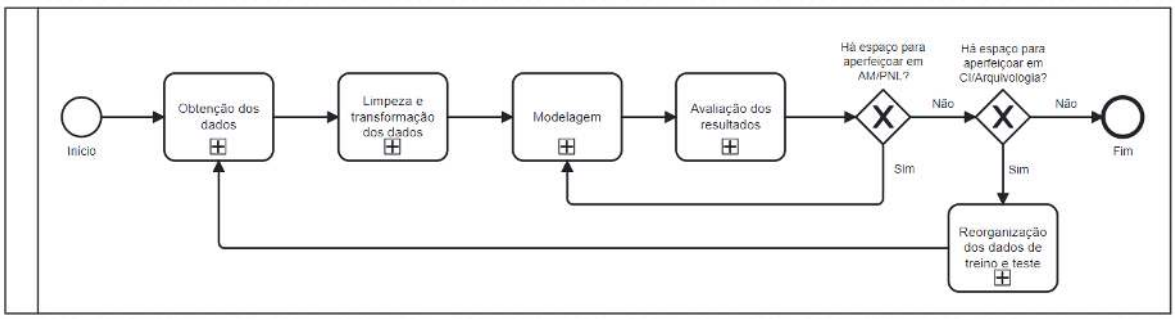
Shabou e outros (2020) desenvolveram um método automatizado para avaliar a relevância de dados registrados em diferentes formatos e conteúdo. A metodologia utilizada por eles adotou a combinação de dois eixos, o primeiro é um modelo de

Arquivologia e o segundo a Mineração de Dados, em que cada um deles possui métricas próprias para compor ao final o modelo de avaliação.

Meireles, Cendón e Almeida (2016) realizaram dois experimentos para categorizar de forma automática artigos científicos de um domínio de conhecimento restrito. No primeiro a categorização foi feita a partir de palavras-chave, ao passo que no segundo foram geradas categorias com o uso de Redes Neurais Artificiais a partir da citação dos artigos. Os autores concluíram que a segunda abordagem foi mais eficiente para a categorização por inserir a variável relacionada ao acoplamento bibliográfico com uma forte relação semântica entre artigos do mesmo grupo.

As três pesquisas mencionadas contribuem no experimento ao enfatizar a importância de que os métodos automatizados na área comportem um espaço dedicado à reflexão e análise a partir dos conhecimentos em Arquivologia. Nesse sentido, elaboramos o diagrama da Figura 6, que segmenta e destaca os dois principais espaços de aperfeiçoamento do modelo.

Figura 6: Diagrama do processo de *Data mining* com destaque para os espaços de aperfeiçoamento



Fonte: elaborada pelos autores.

São exemplos de espaços de aperfeiçoamento em Aprendizagem de Máquina e Processamento de Linguagem Natural que foram aproveitados nessa pesquisa o ajuste dos hiperparâmetros dos modelos a partir dos resultados obtidos com a ferramenta

GridSearchCV do *sklearn* e a utilização do modelo híbrido com uso do *VotingClassifier* do *sklearn* composto pelos três modelos que apresentaram melhores resultados.

O texto de partida utilizado nos modelos é o *dataset* "preproc mínimo" que só faz a retirada das anotações dos documentos próprias do formato HTML. O segundo *dataset* é o "preproc complet" com a retirada de pontuação, converter todos os caracteres em letras minúsculas e remover as *stopwords*. Foi a partir desse segundo *dataset* que iniciamos a pesquisa dos espaços de aperfeiçoamento relacionados à Ciência da Informação e Arquivologia, que no caso foram concentrados no vocabulário dos documentos utilizados para treinamento dos modelos.

Duff e Johnson (2001) pesquisaram como os usuário de arquivos buscam informações, no que identificaram como principais termos utilizados os nomes próprios, datas, lugares, assunto, formato e, ocasionalmente, eventos. Após análise do vocabulário do *dataset* "preproc complet", elaboramos quatro listas com vocabulários selecionados de nomes próprios (pessoas físicas e jurídicas), lugares (cidades, bairros), setores/unidades (nomes de setores ou órgãos públicos) e tempo (mês, ano, período) por entendermos serem viáveis de obter mesmo considerando que o vocabulário base é composto exclusivamente por termos simples.

O vocabulário de setores/unidades somou 2.258 termos, composto em geral por siglas, porém infelizmente não foi possível utilizá-lo para treinar os modelos, uma vez nem todos os documentos continham ao menos um desses termos. Assim, para treinar os modelos de classificação foram utilizados os três vocabulários (nomes, lugares e tempo) mais um quarto vocabulário que reúne os três anteriores ("vocab juntos"). O procedimento de uso dos vocabulários consistiu em manter nos documentos treinados e testados exclusivamente os termos que compõem cada vocabulário, de forma que o conteúdo dos documentos passou a ser representado exclusivamente pelos termos constantes no vocabulário.

A métrica de avaliação dos resultados de algoritmos para avaliação de documentos deve levar em conta que um erro de falso positivo para conservar o documento é menos grave do que um falso negativo, em que o documento é eliminado de forma indevida, de modo que a medida de acurácia não se mostra muito representativa da qualidade da solução buscada (Rolan *et al.*, 2019). Assim, na linha do que propõem os autores, utilizamos o F1 Score na Tabela 2, que é a média harmônica das medidas de precisão e revocação, e que dessa forma consegue expressar melhor a avaliação do resultado pretendido os resultados estão organizados.

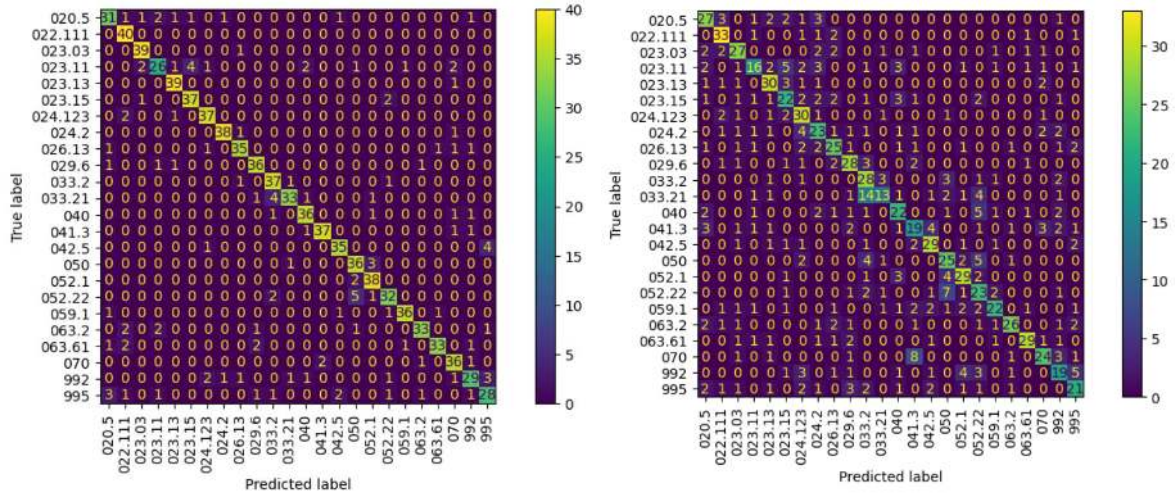
Tabela 2: Resultados obtidos pelos modelos para cada *dataset* com *F1 score*

Modelos classificadores	preproc mínimo n=30.683	preproc complet n=21.459	vocab nomes n=2.820	vocab lugares n=486	vocab tempo n=45	vocab juntos n=3.351
KNeighborsClassifier	**0,700**	**0,719**	**0,573**	**0,056**	**0,206**	**0,597**
LogisticRegression	**0,752**	**0,757**	**0,558**	**0,054**	**0,164**	**0,580**
StochasticGradientDescent	**0,826**	**0,831**	**0,602**	**0,022**	**0,084**	**0,641**
Perceptron	**0,598**	**0,772**	**0,460**	**0,012**	**0,034**	**0,559**
MultiLayerPerceptron	**0,830**	**0,861**	**0,592**	**0,073**	**0,220**	**0,610**
PassiveAggressive	**0,832**	**0,859**	**0,601**	**0,018**	**0,067**	**0,646**
LinearDiscriminantAnalysis	**0,359**	**0,093**	**0,578**	**0,050**	**0,158**	**0,583**
QuadraticDiscrimAnalysis	**0,085**	**0,129**	**0,177**	**0,003**	**0,060**	**0,109**
Ridge	**0,842**	**0,870**	**0,612**	**0,042**	**0,128**	**0,649**
SupportVectorMachine	**0,790**	**0,802**	**0,592**	**0,061**	**0,168**	**0,630**
MultinomialNaiveBayes	**0,745**	**0,789**	**0,586**	**0,043**	**0,144**	**0,605**
ComplementNaiveBayes	**0,776**	**0,851**	**0,557**	**0,042**	**0,122**	**0,573**
BernoulliNaiveBayes	**0,712**	**0,743**	**0,568**	**0,048**	**0,166**	**0,587**
RandomForestClassifier	**0,726**	**0,726**	**0,564**	**0,068**	**0,218**	**0,574**
DecisionTreeClassifier	**0,594**	**0,629**	**0,358**	**0,064**	**0,186**	**0,428**
GaussianNaiveBayes	**0,757**	**0,795**	**0,504**	**0,037**	**0,050**	**0,538**
GradientBoosting	**0,032**	**0,653**	**0,318**	**0,034**	**0,119**	**0,316**
AdaBoostClassifier	**0,460**	**0,534**	**0,240**	**0,040**	**0,121**	**0,204**
HistGradientBoosting	**0,008**	**0,023**	**0,003**	**0,009**	**0,003**	**0,018**
VotingClassifier	**0,838**	**0,865**	**0,622**	**0,055**	**0,196**	**0,648**
Mediana	**0,736**	**0,764**	**0,566**	**0,042**	**0,136**	**0,581**
Melhor resultado	**0,842**	**0,870**	**0,622**	**0,073**	**0,220**	**0,649**

Fonte: elaborada pelos autores.

O melhor resultado obtido foi o do modelo Ridge com o uso do *dataset* "preproc completo" com 21.459 termos, com *F1 score* de 0,870, acurácia de 0,871, precisão de 0,872 e revocação de 0,871, com as predições ilustradas na matriz de confusão da Figura 7 com o uso da biblioteca *Matplotlib*.

Figura 7: Matrizes de confusão dos resultados do modelo Ridge com *datasets* "preproc completo" e "vocab nomes"



Fonte: elaborada pelos autores.

7 AVALIAÇÃO DOS RESULTADOS

O melhor resultado obtido, de *F1 score* de 0,870 e acurácia de 0,871 fica em uma posição intermediária diante das pesquisas relatadas na literatura na Tabela 1. A partir da presente pesquisa, consideramos que a aferição dos resultados quantitativos é importante nesse tipo de investigação no sentido de mensurar os reflexos das decisões adotadas, porém nos parece que a relevância para a Ciência da Informação e a Arquivologia reside mais nos pontos de articulação dessas áreas com o uso do Aprendizado de Máquina e do Processamento da Linguagem Natural.

Rolan e outros (2019) apontam que, mesmo na classificação binária "simples" no processo de avaliação de manter ou descartar, está subjacente a complexidade de automatizar a compreensão do contexto. As limitações estão inclusive na portabilidade de modelos mesmo entre conjuntos de dados que sejam em aparência semelhantes, o que requer atenção a suposições introduzidas, preconceitos ou erros com os desvios de escopo ou mudanças no contexto ainda que sutis.

Entendemos que o diferencial da presente pesquisa em relação à literatura empírica pesquisada (Quadro 1) reside na utilização de conhecimentos arquivísticos para balizar as decisões de aperfeiçoamento dos modelos de classificação de modo a captar e tratar com maior precisão as complexidades apontadas por Rolan *et al.* (2019). Portanto, a abordagem adotada segue as abordagens de Meireles, Cendón e Almeida (2016) e Chabin (2020). Por isso a análise dos resultados quantitativos será mais no sentido da comparação entre os diferentes vocabulários utilizados, ou seja, como as diferentes formas de organizar os *datasets* de treino impactam os resultados dos modelos.

A utilização de vocabulários parciais do *dataset* "preproc completo" permitiu identificar que os termos relacionados a nomes próprios influenciam bastante de forma isolada os modelos de classificação (mediana de 0,622), mais do que lugares (mediana de 0,042) e tempo (mediana de 0,136). Nessa avaliação temos que considerar o fato dos vocabulários utilizados possuírem somente termos simples, o que provavelmente reduziu o potencial a ser alcançado se tivéssemos termos compostos de modo a confirmar com maior precisão as evidências relatadas por Duff e Johnson (2001). Outro ponto a ser observado é que o tamanho dos vocabulários não é proporcional aos resultados obtidos (a mediana para o "vocab tempo" foi mais de três vezes superior à do "vocab lugares", embora o tamanho do *dataset* seja apenas 10% deste), o que evidencia a importância dos critérios de seleção de termos.

Em relação ao resultado do *dataset* "vocab juntos" percebemos que a mediana do resultado dos modelos é superior aos três *datasets* parciais criados. Somente para quatro modelos a reunião dos vocabulários significou resultados piores (*Quadratic Discrimant Analysis, Gradient Boosting, Ada Boost Classifier* e *Hist Gradient Boosting*), sendo que todos eles possuem resultados bem abaixo da mediana.

As matrizes de confusão (Figura 7) demonstra que o melhor modelo produziu resultados diferentes para cada código de classificação. A variação ocorreu tanto no *dataset* "preproc completo" como no "vocab nomes", o que indica o potencial de pesquisa para alocar determinados atributos para códigos de classificação específicos em detrimento de outros.

O fato de utilizarmos os códigos de classificação previamente estabelecidos evitou o trabalho de definição do esquema de classificação, como ocorreu com o projeto *AutoRecords* em que se consumiu muito tempo para definir sobre a granularidade das classes de assuntos (Marcus, 2002). Por outro lado, como trabalhamos com um esquema de classificação fixo e definido previamente, não tivemos a possibilidade de aperfeiçoá-lo, o que poderia ser feito a partir da análise das matrizes de confusão (Figura 7) para os códigos que concentram muitos erros em determinada classe. Pajares, Tornero e Martin (2022) vão além ao apontar que os modelos de IA podem fornecer *datasets* que contribuam para modificar o modelo documental do Sistema de Gestão de Documentos da organização por meio da incorporação dos dados na sua gestão.

Os resultados da pesquisa dos arquivos nos parecem reforçam a necessidade de criação não só de novas ferramentas, como também de metodologias e abordagens de uso e análise dos arquivos como dados (Moss; Thomas; Gollins, 2018). O diagrama da Figura 6 nos parece ser uma importante contribuição desta pesquisa ao evidenciar em um fluxo o subprocesso para trabalhar o espaço de aperfeiçoamento do modelo de

classificação com base na Ciência da Informação e Arquivologia, o que fazemos a partir dos trabalhos de Chabin (2020), Shabou e outros (2020) e Meireles, Cendón e Almeida (2016), bem como das conclusões do *The National Archives UK* (2016) de que a contribuição humana nesse processo é aprimorada pela tecnologia.

O diagrama proposto vai ao encontro de Sousa (2022), para quem o avanço na classificação automática de documentos requer a junção dos conhecimentos em tecnologia da informação com os de arquivologia. Além disso, o diagrama destaca a importância da organização de *datasets* de treinamento que podem chegar a dezenas de milhares de documentos anotados manualmente por humanos como instrumento essencial para aperfeiçoar os modelos de classificação automática (Lee, 2018).

Por fim, os resultados positivos do *Multi-layer Perceptron Classifier*, único dos modelos testados que faz uso de algoritmos de aprendizagem profunda, confirmam o seu potencial para esse tipo de tarefa como adiantado por Payne e Baron (2017), Ribeiro e Assis (2018) e Sousa (2022).

8 CONCLUSÕES

Podemos concluir que é positiva a resposta à pergunta de pesquisa se o uso do Aprendizado de Máquina pode contribuir com a classificação automática de documentos de arquivo. Diante do acervo cada vez maior de documentos digitais, muitos deles que vão se acumulando nos arquivos intermediários sem uma perspectiva de tratamento arquivístico adequado, o Aprendizado de Máquina pode enriquecer os dados de processos e documentos além de agregar informações úteis para as decisões a cargo dos gestores e dos arquivistas.

Como destacado por Guercio (2017), a reflexão teórica sobre os documentos digitais não possui a mesma presença após o período de 1995 a 2005, o que tem

deixado nas mãos de fornecedores de *software* a tarefa de propor as soluções operacionais de gestão de documentos. Ocorre que tais ferramentas não contemplam inteligência arquivística nas suas funcionalidades, o que é objeto da proposta do diagrama da Figura 6 no qual propomos um subprocesso específico para trabalhar o espaço de aperfeiçoamento do modelo de classificação com base na Ciência da Informação e Arquivologia.

Como pesquisas futuras indicamos desenvolver novas formas de categorizar os termos dos documentos de modo a testar e avaliar seus potenciais e limites para criação de dados estruturados para treinamento dos algoritmos de classificação, incorporar a análise diplomática, de tipos e séries documentais na elaboração de *datasets* de treinamento de modelos de classificação automática, acrescentar a utilização de termos compostos por duas ou mais palavras (*ngrams*) de modo a ampliar a apropriação de significados do texto dos documentos, e utilizar a combinação de técnicas modernas de colaboração humana no processo de aperfeiçoamento do modelo de aprendizado de máquina, como o *Active Learning*, o que exige a participação de profissionais da informação capacitados (Lee, 2018).

REFERÊNCIAS

AMOZURRUTIA, Alicia Barnard; ASTORGA, Yaminel Bernal; HIDALGO, Rodrigo Cuéllar; VELÁZQUEZ, Claudia Alin Escoto; GARCÍA-VELÁZQUEZ, Luis Miguel. Inteligencia artificial en los archivos. **Tábula**, *[s. l.]*, n. 25, p. 41-59, 2022. Disponível em: https://publicaciones.acal.es/tabula/article/view/936. Acesso em: 8 jun. 2022.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 15489-1**: Informação e Documentação: gestão de documentos de arquivo: Parte 1: Conceitos e princípios. Rio de Janeiro: ABNT, 2018.

CHABIN, Marie-Anne. The potential for collaboration between AI and archival science in processing data from the French great national debate. **Records Management**

Journal, *[s. l.]*, v. 30, n. 2, p. 241-252, 2020. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/RMJ-08-2019-0042/full/html. Acesso em: 8 jun. 2022.

COLAVIZZA, Giovanni; BLANKE, Tobias; JEURGENS, Charles; NOORDEGRAAF, Julia. Archives and AI: an overview of current debates and future perspectives. **ACM Journal on Computing and Cultural Heritage**, *[s. l.]*, v. 15, n. 1, p. 1-15, 2021. Disponível em: https://dl.acm.org/doi/full/10.1145/3479010. Acesso em: 8 jun. 2022.

CUNNINGHAM, Adrian. ¿Como se lleno está el vaso? Cambios e desafios para los profesionales de los documentos frente a la transformación digital em la era de los datos. **Tabula**, *[s. l.]*, n. 24, p. 25-40, 2021. Disponível em: https://www.bne.es/es/blog/biblioteconomia/2022/03/02/como-de-lleno-esta-el-vaso-cambios-y-desafios-para-los-profesionales-de-los-documentos-frente-a-la-transformacion-digital-en-la-era-de-los-datos. Acesso em: 8 jun. 2022.

DUFF, Wendy M.; JOHNSON, Catherine A. A virtual expression of need: An analysis of e-mail reference questions. **The American Archivist**, *[s. l.]*, v. 64, n. 1, p. 43-60, 2001. Disponível em: https://www.jstor.org/stable/40294158. Acesso em: 9 ago. 2022.

FALCÃO, Luander Cipriano de Jesus; LOPES, Brenner; SOUZA, Renato Rocha. Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. **Em Questão**, Porto Alegre, v. 28, n. 1, p. 13-34, 2022. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/111323. Acesso em: 9 jun. 2022.

GUERCIO, Maria. La classificazione nell'organizzazione dei sistemi documentari digitali: criticità e nuove prospettive. **Italian Journal of Library, Archives and Information Science**, *[s. l.]*, v. 8, n. 2, p. 4-17, 2017. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=6085170. Acesso em: 10 ago. 2022.

HUTCHINSON, Tim. Protecting privacy in the archives: Supervised machine learning and born-digital records. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2018, Seatle, USA. **Proceedings** [...]. Seatle: IEEE, 2018. p. 2696-2701.

LEE, Christopher A. Computer-assisted appraisal and selection of archival materials. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2018, Seatle, USA. **Proceedings** [...]. Seatle: IEEE, 2018. p. 2721-2724.

MARCUS, Richard W. NARA: a sneak preview. **Information Management Journal**, *[s. l.]*, v. 36, n. 2, p. 56-58, 2002. Disponível em: https://www.proquest.com/docview/227759197. Acesso em: 8 jun. 2022.

MEIRELES, Magali Rezende Gouvêa; CENDÓN, Beatriz Valadares; ALMEIDA, Paulo Eduardo Maciel de. Comparação do processo de categorização de documentos utilizando palavras-chave e citações em um domínio de conhecimento restrito. **Transinformação**, Campinas, v. 28, n. 1, p. 87-96, 2016. Disponível em: https://www.scielo.br/j/tinf/a/k6xqTtLLCZ6TxrYWzbPsxLw/abstract/?lang=pt#. Acesso em: 8 jun. 2023.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. *In:* WAINER, Jacques; RHEINGANTZ, Paulo (org.). **Sistemas Inteligentes**: fundamentos e aplicações. Barueri: Manole Ltda., 2003. p. 89-114.

MONTEREI, Rafaella Carine; LOPES, Dalton Martins. Perspectivas do uso do Aprendizado de Máquina em Bibliotecas: reflexões iniciais de uma pesquisa em andamento. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 21., 2021, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2021.

MOSS, Michael; THOMAS, David; GOLLINS, Tim. The reconfiguration of the archive as data to be mined. **Archivaria**, *[s. l.]*, v. 86, p. 118-151, 2018. Disponível em: https://archivaria.ca/index.php/archivaria/article/view/13646. Acesso em: 9 jun. 2022.

PAJARES, Pepita Raventós; TORNERO, Celio Hernández; MARTIN, Meritxell Simon. IA y archivo: La tecnología de reconocimiento de textos manuscritos en fondos patrimoniales: un ejemplo de 10 diarios en la formación de maestras y maestros en el año 1932. **Tabula**, *[s. l.]*, n. 25, p. 83-100, 2022.

PAYNE, Nathaniel. Stirring the cauldron: redefining computational archival science (CAS) for the Big Data domain. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2018, Seatle, USA. **Proceedings** [...]. Seatle: IEEE, 2018. p. 2743-2752.

PAYNE, Nathaniel; BARON, Jason R. Auto-categorization methods for digital archives. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2017, Boston, USA. **Proceedings** [...]. Boston: IEEE, 2017. p. 2288-2298.

PINHEIRO, Mayara; OLIVEIRA, Hamilton. Inteligência Artificial: estudos e usos na Ciência da Informação no Brasil. **Revista Ibero-Americana de Ciência da Informação**, Brasília, DF, v. 15, n. 3, p. 950-968, 2022. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/42767. Acesso em: 19 nov. 2022.

PUBLIC RECORD OFFICE VICTORIA. **Email Machine Assisted Appraisal**: proof of concept. Disponível em: https://prov.vic.gov.au/sites/default/files/files/Blog/Government%20recordkeeping/Victoria%20Government%20Email%20Machine%20Assisted%20Appraisal%20Final.pdf. Acesso em: 12 jun. 2022.

RIBEIRO, Patrick Dourado; ASSIS, João Marcus Figueiredo. Deep learning e suas potencialidades para a classificação arquivística. *In*: CONGRESSO INTERNACIONAL EM HUMANIDADES DIGITAIS, 1., 2018, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2018.

ROLAN, Gregory; HUMPHRIES, Glen; JEFFREY, Lisa; SAMARAS, Evanthia; ANTSOUPOVA, Tatiana; STUART, Katharine. More human than human? Artificial intelligence in the archive. **Archives and Manuscripts**, *[s. l.]*, v. 47, n. 2, p. 179-203, 2019. Disponível em: https://www.tandfonline.com/doi/full/10.1080/01576895.2018.1502088. Acesso em: 9 jun. 2022.

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. A systematic literature review on applying CRISP-DM process model. **Procedia Computer Science**, *[s. l.]*, v. 181, p. 526-534, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050921002416. Acesso em: 9 jun. 2023.

SHABOU, Basma Makhlouf; TIÈCHE, Julien; KNAFOU, Julien; GAUDINAT, Arnaud. Algorithmic methods to explore the automation of the appraisal of structured and unstructured digital data. **Records Management Journal**, *[s. l.]*, v. 30, n. 2, p. 175-200,

2020. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0049/full/html. Acesso em: 9 jun. 2023.

SHEARER, Colin. The CRISP-DM Model: the new blueprint for data mining. **Journal of Data Warehousing**, *[s. l.]*, v. 5, n. 4, p. 13-22, 2000.

SHINDE, Pramila P.; SHAH, Seema. A review of machine learning and deep learning applications. *In*: INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION CONTROL AND AUTOMATION, 14., 2018. **Proceedings** [...]. Pune: IEEE, 2018. p. 1-6.

SHINKLE, Tim. Automated electronic records management: are we there yet? **IQ**: the RIM Quarterly, *[s. l.]*, v. 33, n. 4, p. 36-40, 2017.

SILVA, Narjara; NATHANHSON, Bruno Macedo. Análise da produção científica em Inteligência Artificial na área da Ciência da Informação no Brasil. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina. **Anais** [...]. Londrina: Universidade Estadual de Londrina, 2018.

SOUSA, Renato Tarciso Barbosa de. A classificação automática de documentos de arquivo é uma solução para os problemas que os usuários encontram com a classificação funcional? Algumas reflexões e caminhos a percorrer. *In:* BARROS, Thiago Henrique Bragato; LAIPELT, Rita do Carmo Ferreira (org.). **Organização e representação do conhecimento em múltiplas abordagens**. São Paulo: Pimenta Cultural, 2022. p. 221-251.

THE NATIONAL ARCHIVES UK. **The application of technology-assisted review of born-digital records transfer, inquiries and beyond**. London: Crown, 2016. Disponível em: https://cdn.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf. Acesso em: 28 maio 2022.

TKACHENKO, A. L.; DENISOVA, L. A. Designing an information system for the electronic document management of a university: automatic classification of documents. **Journal of Physics**: conference series, *[s. l.]*, p. 1-10, 2022.

TRACE, Ciaran B. Archival infrastructure and the information backlog. **Archival Science**, Dordrecht, v. 22, n. 1, p. 75-93, 2022. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34776768. Acesso em: 9 jun. 2022.

VELLINO, André; ALBERTS, Inge. Assisting the appraisal of e-mail records with automatic classification. **Records Management Journal**, *[s. l.]*, v. 26, n. 3, 2016, p. 293-313, 2016.

WANG, Zhiyu; WU, Jingyu; YU, Guang; SONG, Zhiping. Text Analysis and Visualization Research on the Hetu Dangse During the Qing Dynasty of China. **Information Technology and Libraries**, *[s. l.]*, v. 40, n. 3, p. 1-23, 2021.

WARLAND, Andrew, MOKHTAR, Umi Asma. Can technology classify records better than a human? **IRMS Bulletin**, *[s. l.]*, n. 171, p. 16-19, 2013.

Copyright: Esta obra está licenciada com uma Licença Creative Commons Atribuição 4.0 Internacional.

tpbci@ancib.org @anciboficial @ancib_brasil