XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-8 – Informação e Tecnologia

CONTRIBUIÇÕES DO USO DA APRENDIZAGEM DE MÁQUINA PARA A AVALIAÇÃO DE DOCUMENTOS DE ARQUIVO

CONTRIBUTIONS OF THE USE OF MACHINE LEARNING FOR THE APPRAISAL OF RECORDS

Eduardo Watanabe. UNB.

Renato Tarciso Barbosa de Sousa. UNB.

Modalidade: Trabalho Completo

Resumo: A transformação digital nos últimos 30 anos não resultou em avanços significativos na qualidade da gestão de documentos nas organizações, mas diante da evolução das tecnologias e métodos é que se pergunta: o uso de algoritmos de aprendizado de máquina pode contribuir com a avaliação de documentos de arquivo por meio da sugestão do código de classificação a ser atribuído a um documento de uma organização pública? Os procedimentos metodológicos consistem na revisão de literatura e nas tarefas propostas pelo modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) em um experimento com 1.768 documentos de 598 processos da Advocacia-Geral da União. Foram desenvolvidos modelos de aprendizagem supervisionada com o uso de algoritmos especializados para fazer a atribuição automatizada do código de classificação do documento de forma a apoiar o processo de avaliação.

Palavras-Chave: Aprendizagem de Máquina. Gestão de Documentos. Avaliação de Documentos.

Abstract: The digital transformation in the last 30 years has not resulted in significant advances in the quality of records management in organizations, but given the evolution of technologies and methods, the question arises: can the use of machine learning algorithms contribute to the records appraisal by suggesting the classification code to be assigned to a record of a public organization? The methodological procedures consist of a literature review and the tasks proposed by the CRISP-DM (Cross-Industry Standard Process for Data Mining) model in an experiment with 1,768 records from 598 processes of the Attorney General's Office in Brazil. Supervised learning models were developed using specialized algorithms to perform the automated assignment of the document classification code in order to support the evaluation process.

Keywords: Machine Learning. Records Management. Records Appraisal.

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".

No Brasil e de forma mais específica na administração pública, a situação dos arquivos carrega um histórico de precariedade de massas documentais acumuladas por falta de política estabelecida, ausência de metodologia consolidada e nível insuficiente de qualificação de pessoal (SOUSA, 1997). Nesta pesquisa, destacamos dentre as sete funções arquivísticas a de avaliação de documentos como um dos principais desafios enfrentados pelo arquivista, por envolver a "vida" ou a "morte" de um documento (CHAGAS, 2020).

Em outra perspectiva, a transformação digital tem reconfigurado os arquivos no sentido deles passarem a ser considerados como conjuntos de dados a serem minerados (MOSS et. al., 2018). Com isso, abrem-se espaços para a automação em grande escala com o uso da Inteligência Artificial 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 Inteligência Artificial 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).

Nesse contexto é que formulamos a seguinte pergunta de pesquisa: o uso de algoritmos de aprendizado de máquina pode contribuir com a avaliação de documentos de arquivo por meio da sugestão do código de classificação a ser atribuído a um documento de uma organização pública?

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). Como primeiro objetivo específico de pesquisa será feita a revisão da

literatura. O segundo objetivo específico é a aplicar algoritmos de Aprendizagem de Máquina supervisionados que façam a atribuição automatizada de um código de classificação para documentos reais.

2 PROCEDIMENTOS METODOLÓGICOS

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

O levantamento bibliográfico incluiu as bases de dados da *Library & Information Science Abstracts - LISA* (de 1969 até 24.05.2022) e a Base de Dados em Ciência da Informação - BRAPCI (de 1972 até 24.05.2022). 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".

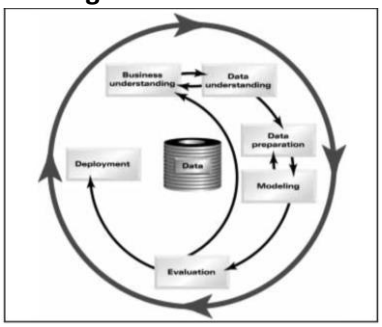
Foram localizados n = 1.201 itens, dos quais 254 foram excluídos por serem repetidos e outros 919 por não serem específicos, restando ao final 26. Os 919 itens não específicos abrangem a menção incidental de um termo de pesquisa ou de todos, podemos exemplificar com artigos com o tema principal da tecnologia *blockchain* em que são feitas menções incidentais sobre gestão de documentos e inteligência artificial, sem que tratem de pesquisas relacionando os conceitos entre si como é o objeto deste trabalho.

Importante ressaltar que durante a leitura das referências bibliográficas em inglês a regra foi associar os termos "*records*", "*recordkeeping*" e "*records management*" com "gestão de documentos nas fases corrente e intermediária", e os termos "*archives*" e "*archival*" com "gestão de documentos na fase de guarda permanente". A exceção foi o artigo de Colavizza e outros (2021), que, embora utilize as expressões "*archives*" e "*archival*", consideramos que elas se referem à gestão de documentos em todas as fases (corrente, intermediária e permanente), tendo em vista que os autores partem da ótica do modelo *Records Continuum*.

Os procedimentos metodológicos consistem nas tarefas propostas pelo CRISP-DM (*Cross-Industry Standard Process for Data Mining*), 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 e outros (2011) 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, p. 14).

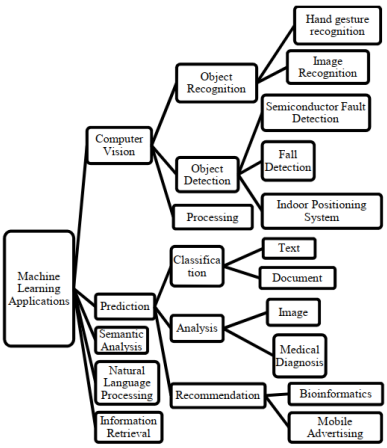
Como este trabalho é limitado à pesquisa do modelo, serão utilizadas somente as cinco primeiras fases do CRISP-DM, que são (1) Entendimento do negócio, (2) Entendimento dos dados, (3) Preparação dos dados, (4) Modelagem e (5) Avaliação do modelo. Um dos principais pontos fortes do CRISP-DM é ser um processo iterativo de refinamento contínuo em que a execução das fases não fica engessada em uma sequência pré-determinada. Será utilizado o *software* KNIME para a análise dos metadados do processo e conteúdo dos documentos digitais.

3 APRENDIZADO DE MÁQUINA NA ARQUIVOLOGIA E AVALIAÇÃO DE DOCUMENTOS

De acordo com a Associação Brasileira de Normas Técnicas (ABNT) ISO/TR 21946:2020, a avaliação de documentos consiste no processo recorrente de avaliar as atividades de negócio para então determinar quais documentos de arquivo precisam ser produzidos e capturados, bem como por quanto tempo precisam ser mantidos. Os seus principais benefícios são a conformidade aos requisitos legais e regulatórios para documentos, a satisfação das necessidades de negócio com a gestão dos documentos, a provisão da tempestiva destinação, dentre outros.

Sobre as aplicações de Aprendizado de Máquina, Shinde e Shah (2018) fizeram a revisão de pesquisas na literatura, no que propõem uma classificação conforme a Figura 1.

Figura 1 - Aplicações de Aprendizado de Máquina.



Fonte: SHINDE e SHAH (2018).

O experimento do presente trabalho aplicará na avaliação de documentos de arquivo por meio da sugestão do tempo em que eles devem ser mantidos por algoritmos de Aprendizado de Máquina relativos a "Predição-Classificação de Documentos" combinada com o Processamento da Linguagem Natural (NLP na sigla da expressão em inglês).

Em um cenário de digitização, *big data* e avanços tecnológicos, Nathaniel Payne (2018) propõe a criação de um novo campo de estudos transdiciplinar, 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 Inteligência Artificial.

Colavizza e outros (2021) revisaram na literatura a intersecção entre a Inteligência Artificial 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 Inteligência Artificial, 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). No capítulo de Discussão dos resultados, a Tabela 1 será comparada com os resultados do experimento da presente pesquisa.

Quadro 1 - Pesquisas quantitativas realizadas ou 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
VELLINO et al., 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; 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 e DENISOVA, 2022	Siberian State Automobile and Highway University	Classificar automaticamente os documentos de uma universidade em quatro classes.	2022

Fonte: Elaborado pelos autores (2022).

Tabela 1 - Pesquisas quantitativas realizadas ou 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
VELLINO et al., 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; 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 e DENISOVA, 2022	1.778 documentos	Híbrido de Support Vector Machine (SVM) e k-nearest neighbor (kNN)	n/d	0,983*

* Os experimentos efetuaram testes que combinaram diferentes de 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.

Fonte: Elaborado pelos autores (2022).

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 filtros para que o conjunto de dados atenda aos objetivos da pesquisa: processos das atividades-meio por terem uma característica de maior

homogeneidade em comparação com os documentos das atividades-fim; processos na fase de Arquivo Intermediário por já terem sido encerrado; só os documentos produzidos por usuário da própria AGU para 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.

Quadro 2 - Códigos de classificação selecionados para o experimento.

Código	Descritor do código
023.11	Admissão. Aproveitamento. Contratação. Nomeação. Readmissão. Readaptação. Recondução. Reintegração. Reversão
023.13	Lotação. Remoção. Transferência. Permuta
023.15	Requisição. Cessão
029.11	Controle de frequência. Livros. Cartões. Folhas de ponto. Abono de faltas. Cumprimento de horas extras

Fonte: Elaborado pelos autores (2022).

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 598 processos administrativos com 1.768 documentos digitais no total.

5 PREPARAÇÃO DOS DADOS

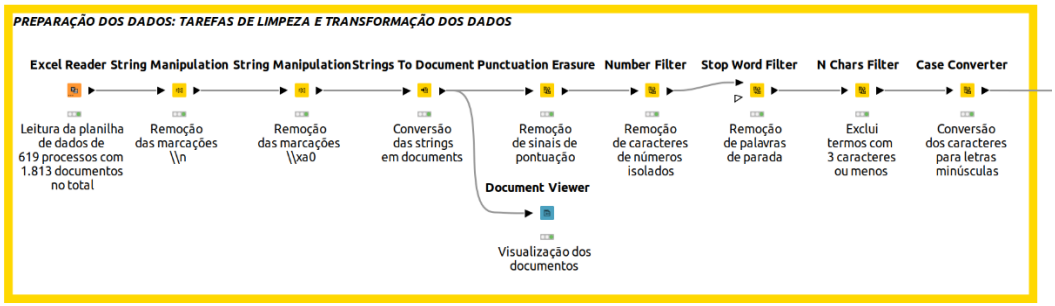
A partir da fase de preparação dos dados foi utilizada a ferramenta *KNIME Analytics Platform*, acrônimo de *Konstanz Information Miner*, que é *open source* para mineração de dados em linguagem Java, criada em 2017 por uma equipe de desenvolvedores da *University of Konstanz* na Alemanha. O *KNIME* tem se destacado como uma das ferramentas de mineração de dados mais promissoras conforme avaliações feitas em estudos na área (ALTALHI et al., 2017; HORA et al., 2018).

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. O planejamento de mais tarefas da fase de preparação dos dados teve que levar em consideração a gestão tempo para que a pesquisa concluísse com sucesso todas as cinco fases no prazo estabelecido. Nesse sentido, foi necessário simplificar alguns passos na preparação dos dados, de modo que não foram

executadas as tarefas de amostragem e balanceamento dos dados a partir de critérios técnicos.

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.

Figura 2 - Tarefas de limpeza e transformação de dados.



Fonte: Elaborado pelos autores (2022).

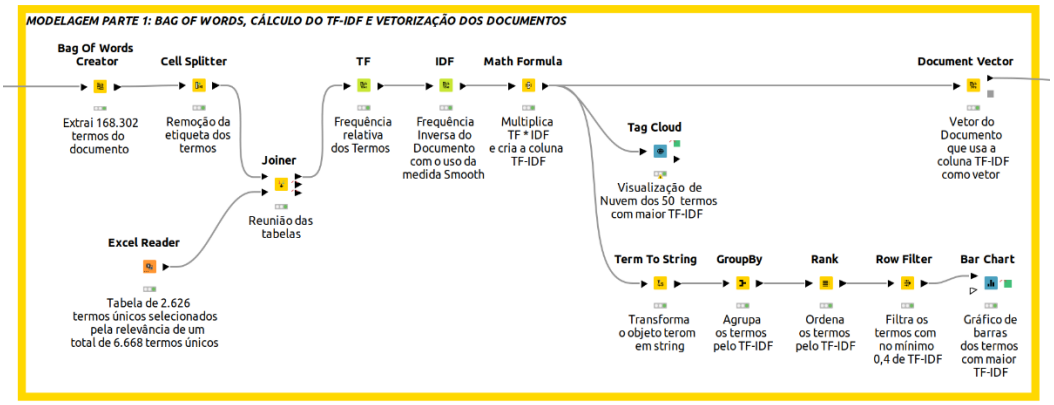
6 MODELAGEM

Na mesma linha de Cunninghan (2021), consideramos que a obtenção de resultados satisfatórios pela gestão de documentos requer a combinação de bons padrões, instrumentos e modelos. Na presente pesquisa buscamos a partir unicamente do texto de documentos de um processo inferir por algoritmos de Aprendizado de Máquina informações que possam ser úteis para o arquivista definir qual é a classificação desse processo para fins de avaliação.

Por isso a opção utilizada foi utilizar algoritmos de aprendizagem não-supervisionada, de forma mais específica algoritmos que busquem agrupar (clusterizar) documentos similares de modo a ser possível inferir qual a classificação do processo para fins de avaliação. 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).

Não obstante a execução de todas as tarefas de limpeza e tratamento de dados, nessa parte verificamos a partir da análise de nuvem de termos a presença de muitos nomes próprios, o que mostrou a limitação do uso do algoritmo para sua remoção considerando ter sido feito para a língua espanhola, destacando que não há ainda algoritmo próprio no *KNIME* para a língua portuguesa. Com isso foi necessária a seleção manual dos termos mais relevantes a serem utilizados, que é descrita na figura a seguir.

Figura 3 - Algoritmos do KNIME para a Modelagem - Parte 1.



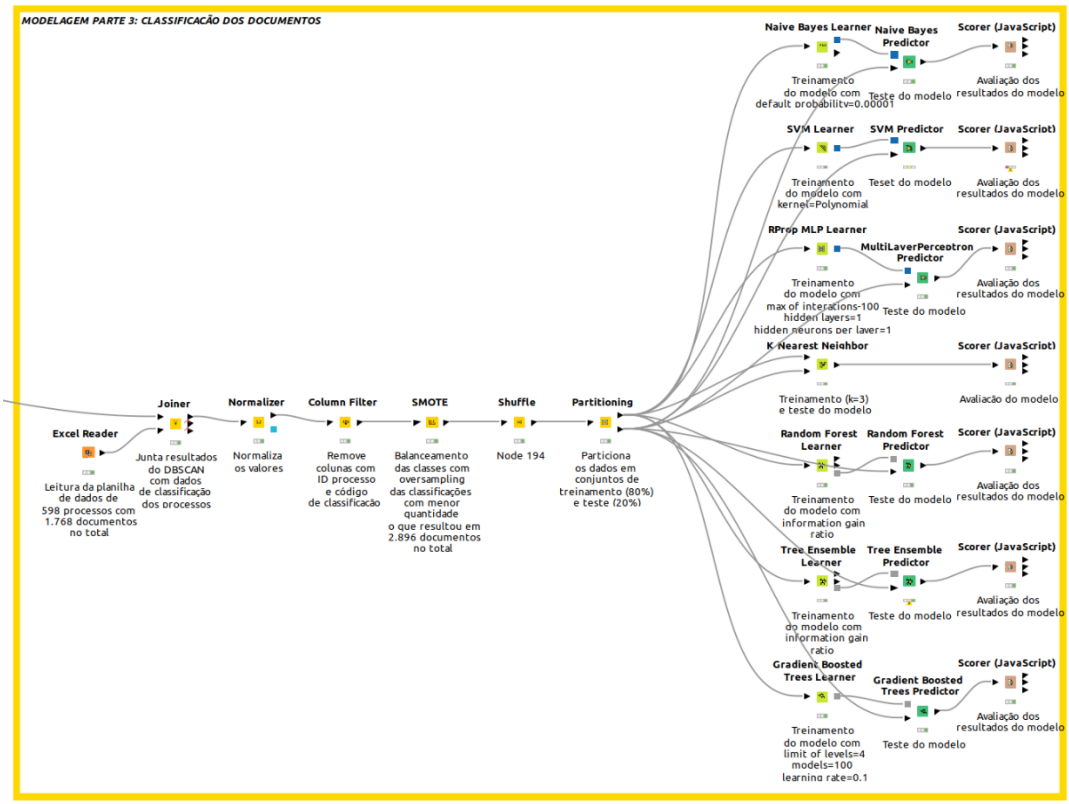
Fonte: Elaborado pelos autores (2022).

A partir dos vetores de documentos, foram utilizados na parte 2 da modelagem diferentes algoritmos de classificação. Os quatro primeiros algoritmos utilizados foram os identificados na revisão de literatura como disponíveis por não serem proprietários: *Support Vector Machine* (SVM), *Multinomial Naïve Bayes*, *Multi-Layer Perceptron* (MLP) e *k-Nearest Neighbor* (kNN). Acrescentamos no experimento os algoritmos *Random Forest*, *Tree Ensemble Learner* e *Gradient Boosted Trees*, que são utilizados com bons resultados em pesquisas de NLP (ANWAR et al., 2021), conforme ilustrado na Figura 4.

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 Rolan e outros (2019), 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.

Dois algoritmos não geraram resultados parcialmente (Naive Bayes) ou nenhum (SVM) no *Knime*, ao que indica pelo fato da quantidade de atributos do vetor de documentos (termos expressos nas colunas) superar a quantidade de linhas. Essa situação impediu também de executar o algoritmos de Regressão Logística. Preferimos manter os dois primeiros algoritmos na tabela pelo fato de terem sido utilizados em pesquisas anteriores.

Figura 4 - Algoritmos do KNIME para a Modelagem - Parte 2.



Fonte: Elaborado pelos autores (2022).

Tabela 2 - Resultados obtidos pelos algoritmos para cada classe com *F1 score*.

Algoritmos	Código 023.11	Código 029.11	Código 023.13	Código 023.13
Naive Bayes	0,624	0,446	n/d	n/d
SVM	n/d	n/d	n/d	n/d
Multi Layer Perceptron	0,982	0,949	0,911	0,821
K Nearest Neighbor	0,851	0,914	0,785	0,577
Random Forest	0,960	0,944	0,883	0,837
Tree Ensemble	0,935	0,939	0,879	0,810
Gradient Boosted Trees	0,943	0,942	0,872	0,845

Fonte: Elaborado pelos autores (2022).

7 AVALIAÇÃO DOS RESULTADOS

Os melhores resultados apresentados na Tabela 2 são em geral próximos dos encontrados na literatura sobre gestão de documentos (Tabela 1), o que indica uma possível agenda de pesquisa sobre o potencial de generalização de uso dos algoritmos em acervos diversos com as devidas cautelas.

Destacamos também que os algoritmos produziram resultados em diferentes códigos de classificação, o que sugere que a criação de modelos híbridos que combinam diferentes

algoritmos em linha, como feito por Tkachenko e Denisova (2022), pode ser uma linha promissora para melhorar em pesquisas futuras os resultados alcançados.

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).

No decurso da pesquisa nos deparamos com uma informação que pode ser útil para o planejamento da gestão de documentos, que é a lista de 6.668 termos acompanhados do respectivo *TF-IDF*. A partir da análise criteriosa dessa lista poderá ser dado início à elaboração de instrumentos para instituir uma linguagem documentária na organização, ou mesmo servir de insumo para a atualização de instrumentos já existentes.

Essa situação evidenciou no curso do experimento a importância das etapas de preparação dos dados e da análise e das linguagens documentárias para a aperfeiçoar os resultados obtidos. Nesse ponto é que o CRISP-DM atua para que então se retorne na tarefa de limpeza de dados para que os resultados possam ser então melhorados.

Os resultados da pesquisa dos arquivos 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 et. al., 2018). Um desses aspectos nos parece ser a publicação dos resultados de pesquisa.

No presente trabalho publicamos os resultados do experimento na Tabela 2 dos diferentes algoritmos e não apenas do melhor resultado obtido como feito nas pesquisas quantitativas no Quadro 1. Essa forma de divulgação dos resultados nos parece ser útil para que futuras pesquisas possam reutilizar aprendizados obtidos em relação à aplicabilidade dos algoritmos em diferentes contextos.

8 CONCLUSÕES

Podemos concluir que é positiva a resposta à pergunta de pesquisa se o uso do Aprendizado de Máquina pode contribuir com a avaliação de documentos de arquivo por meio da sugestão do código de classificação a ser atribuído a um documento de uma organização pública. 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.

Os resultados parecem animadores, mas ainda demandam mais pesquisas para compreender e melhorar a utilização da Aprendizagem de Máquina, isso no sentido de que possam fornecer um entendimento mais rico dos documentos com as dimensões de conceitos, palavras-chave, entidades, relacionamentos, sentimentos, autoria e mais (SHINKLE, 2017). Na revisão da literatura das pesquisas aplicadas, somente Wang e outros (2021) avançaram na análise estatística de produção documental para compreensão do contexto em que foram produzidos, o que evidencia um enorme campo de pesquisas ainda inexplorado.

Concluímos no mesmo sentido que *The National Archives UK* (2016), de que novas tecnologias como a *technology-assisted review* permitem a extração de significado de coleções volumosas de documentos digitais ainda que haja pouca ou nenhuma estrutura ou conhecimento institucional sobre a informação. E de que a contribuição humana nesse processo é aprimorada pela tecnologia, no que acrescentamos que a própria evolução dos resultados obtidos com as tecnologias envolvidas requer o direcionamento contínuo por especialistas na organização e recuperação da informação.

Como pesquisas futuras indicamos a utilização de maior quantidade de documentos e processos analisados, a inclusão de profissionais da AGU na análise das informações obtidas, a realização de limpeza de dados dos nomes de pessoas por não agregar valor à pesquisa, a inclusão de procedimentos de conferência dos documentos que geraram resultados muito abaixo da média, a elaboração de propostas de elaboração de tesauros e definição de tipos e séries documentais e a análise estatística sobre as atividades de produção documental.

REFERÊNCIAS

ALTALHI, Abdulrahman H.; LUNA, José María; VALLEJO, Mangel; VENTURA, Sebastián. Evaluation and comparison of open source software suites for data mining and knowledge discovery. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 7, n. 3, p. e1204, 2017.

ANWAR, Muchamad Taufiq; PRATIWI, Anggy Eka; UDHAYANA, Khadijah Febriana Rukhmanti. Automatic Complaints Categorization Using Random Forest and Gradient Boosting. **Advance Sustainable Science, Engineering and Technology** (ASSET), v. 3, n. 1, p. 0210106, 2021.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ISO/TR 21946: Informação e Documentação - Avaliação para gestão de documentos de arquivo**. Rio de Janeiro, 2018.

BAILEY, Steve. **Managing the Crowd: Rethinking records management for the Web 2.0 world**. Facet Publishing, 2008.

CHAGAS, Cintia Aparecida. Avaliação de documentos arquivísticos: teoria e metodologia. *In*: **Ágora: Arquivologia em Debate**, v. 30, n. 61, p. 478-498, 2020.

COLAVIZZA, Giovanni; BLANKE, Tobias; JEURGENS, Charles; NOORDEGRAAF, Julia. Archives and AI: an overview of current debates and future perspectives. *In*: **ACM Journal on Computing and Cultural Heritage** (JOCCH), v. 15, n. 1, p. 1-15, 2021.

COX, Richard J. Appraisal and the Future of Archives in the Digital Era. **The Future of Archives and Recordkeeping**: A Reader. J. Hill, p. 213-237, 2011.

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. *In*: **Tabula**, n. 24, p. 21-43.

HORA, Gleidison Santos; SANTOS JÚNIOR, Gilson Pereira; MENEZES, Jislane Silva Santos; REHEM NETO, Almerindo Nascimento. Avaliação de ferramentas de mineração de dados: uma abordagem com o modelo tam. *In*: **Interfaces Científicas-Exatas e Tecnológicas**, v. 2, n. 3, p. 109-121, 2018.

HUTCHINSON, Tim. Protecting privacy in the archives: Supervised machine learning and born-digital records. *In*: **2018 IEEE International Conference on Big Data (Big Data)**. IEEE, p. 2696-2701, 2018.

MARCUS, Richard W. NARA: a sneak preview. *In*: **Information Management Journal**, v. 36, n. 2, p. 56-58, 2002.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. *In*: **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 Biblioteas: reflexões iniciais de uma pesquisa em andamento. *In*: XXI Encontro Nacional de Pesquisa em Ciência da Informação, 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. *In*: **Archivaria**, v. 86, n. 86, p. 118-151, 2018.

PAYNE, Nathaniel. Stirring the cauldron: redefining computational archival science (CAS) for the Big Data domain. *In*: **2018 IEEE International Conference on Big Data (Big Data)**. IEEE, 2018. p. 2743-2752.

ROLAN, Gregory; HUMPHRIES, Glen; JEFFREY, Lisa; SAMARAS, Evanthia; ANTSOUPOVA, Tatiana; STUART, Katharine. More human than human? Artificial intelligence in the archive. *In*: **Archives and Manuscripts**, v. 47, n. 2, p. 179-203, 2019.

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. A systematic literature review on applying CRISP-DM process model. *In*: **Procedia Computer Science**, v. 181, p. 526-534, 2021.

SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. *In*: **Journal of data warehousing**, v. 5, n. 4, p. 13-22, 2000.

SHINDE, Pramila P.; SHAH, Seema. A review of machine learning and deep learning applications. *In*: **2018 Fourth international conference on computing communication control and automation** (ICCUBEA). IEEE, 2018. p. 1-6.

SHINKLE, Tim. Automated electronic records management: Are we there yet?. *In*: **IQ: The RIM Quarterly,** v. 33, n. 4, p. 36-40, 2017.

SOUSA, Renato Tarciso Barbosa de. Os arquivos montados nos setores de trabalho e as massas documentais acumuladas na administração pública brasileira: uma tentativa de explicação. *In*: **Revista de Biblioteconomia de Brasília**, v. 21, n. 1, 1997, p. 31-50.

THE NATIONAL ARCHIVES UK. **The Application of Technology-Assisted Review of Born-Digital Records Transfer, Inquiries and Beyond**. London: Crown, 2016. 28 p. Disponível em: https://cdn.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf. Acesso em: 28 mai 2022.

TKACHENKO, A. L.; DENISOVA, L. A. Designing an information system for the electronic document management of a university: automatic classification of documents. *In*: **Journal of Physics: Conference Series**. IOP Publishing, 2022. p. 012035.

VELLINO, André; ALBERTS, Inge. Assisting the appraisal of e-mail records with automatic classification. *In*: **Records Management Journal**, vol. 26, no. 3, 2016, pp. 293–313.

VICTORIA, Public Record Office. **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.

WANG, Zhiyu; WU, Jingyu; YU, Guang; SONG, Zhiping. Text Analysis and Visualization Research on the Hetu Dangse During the Qing Dynasty of China. *In*: **Information Technology and Libraries**, v. 40, n. 3, 2021.