PRODUÇÃO ONLINE
REVISTA CIENTÍFICA ELETRÔNICA DE ENGENHARIA DE PRODUÇÃO
ISSN 1676-1901

MELHORIA DA CATEGORIZAÇÃO DE PRODUTOS A PARTIR DO USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA E MEDIDAS DE SIMILARIDADE

IMPROVEMENT IN PRODUCT CATEGORIZATION FROM MACHINE LEARNING ALGORITHMS AND SIMILARITY COEFFICIENTS

Maicom Sergio Brandão* E-mail: maicom.brandao@usp.br
Moacir Godinho-Filho* E-mail: moacir@dep.ufscar.br
Walther Azzolini Junior** E-mail: wazzolini@sc.usp.br
Bruna Christina Battissacco** E-mail: brunacb@usp.br
Josadak Astorino Marçola*** E-mail: josadak@gestareconsultoria.com.br
*Universidade Federal de São Carlos (UFScar), São Carlos, SP
** Universidade de São Paulo, São Carlos, SP
***Universidade Paulista, Araraquara, SP

Resumo: O cadastro de produtos é uma atividade primária e essencial de qualquer negócio, mas pode estar cercada por várias armadilhas quando é feita exclusivamente de forma manual, pois inconsistências nos cadastros podem gerar análises incorretas sobre o negócio, resultando em decisões equivocadas. Nesse sentido, o uso de técnicas de aprendizado de máquina pode contribuir para melhorar esse processo. O presente estudo avaliou o uso de diferentes algoritmos e estratégias de aprendizado de máquina em uma atividade de categorização de produtos a partir de suas descrições em uma empresa com alta frequência de criação de novos produtos. Um novo processo foi sugerido a partir da escolha do melhor algoritmo, que apresentou potencial para a redução de erros e revisou o tipo de processo de totalmente manual para semiautomatizado. Além do ganho específico para o caso analisado, o artigo também apresenta o caminho de construção, validação e escolha de modelos de aprendizado, o que contribui para a reprodutibilidade em outros contextos.

Palavras-chave: Cadastro de produtos. Aprendizado de máquina. Árvore de decisão. Redes Neurais. Naive Bayes.

Abstract: Product categorization is an ordinary task in every business, but it involves some pitfalls when it is made by people’s judgment only. Inconsistences in product’s attributes can lead to wrong analysis, and wrong business decisions at the end. Thus, the use of machine learning techniques can contribute to improve this process. The present study evaluated the use of different machine learning algorithms and problem-solving strategies in a product categorization activity based on their descriptions taking into account a company with high speed of creation of new products, and therefore more susceptible to errors when this task is made manually and proposed a new process for this activity that integrates technology as a support. A new process was proposed from the best algorithm, converting the process from manual to semiautomatic. Besides the specific benefits to the company, this study also contributes to practice in unveiling the processes of building, validating and choosing machine learning models.

Keywords: Product categorization. Machine Learning. Decision-Tree. Neural Network. Naive Bayes.

1 INTRODUÇÃO

Aprendizado de máquina é um subcampo da inteligência artificial, que inclui conhecimentos das áreas de engenharia, da ciência da computação, da matemática, da estatística e da ciência de dados para a resolução de diversos problemas da sociedade. Recentemente, tem sido observado um crescente interesse pelo tema, motivado por fatores como o desenvolvimento computacional, que tem gerado baixo custo computacional, e pelo desenvolvimento de novos algoritmos apoiados por grandes bancos de dados (JORDAN; MITCHELL, 2015).

Com inúmeras oportunidades, o aprendizado de máquina tem se mostrado um meio relevante de apoio ao processo decisório dentro das organizações. Por exemplo, utilizar algoritmos de predição de demanda contribui para reduzir o risco de estimativas equivocadas, que em última instância prejudicam tanto o atingimento de níveis de serviço elevados quanto podem resultar em aumento de estoques, prejudicando em termos financeiros toda a organização (VARIAN, 2018).

Por meio da experiência advinda de problemas de aprendizagem, como destacam Jordan e Mitchell (2015), é possível obter melhor desempenho para os algoritmos de aprendizado de máquina, dentre os quais destacam-se desde árvores de decisão até redes neurais, que podem ser classificados de diversas formas, sendo uma comum entre aprendizado supervisionado ou não-supervisionado.

De forma geral, o aprendizado supervisionado ocorre quando as observações da base de dados utilizadas no treinamento de um algoritmo de aprendizado de máquina são rotuladas, ou seja, possuem uma classificação pré-determinada. Por sua vez, em um aprendizado não-supervisionado, a base de dados utilizada não é rotulada, deixando para que o algoritmo retorne suas próprias classificações (MOORE *et al.,* 2019).

Além disso, uma outra característica do aprendizado de máquina é sua ampla possibilidade de aplicação. Além de poder ser utilizado em tarefas de predição dos volumes de vendas, que geralmente utiliza dados quantitativos e históricos (CHERIYAN *et. al*, 2018; PAVLYSHENKO, 2019), também pode envolver o uso de imagens para a predição e reconhecimento de padrões (WANG *et al*., 2018; MARTINEZ-MARTIN, 2019). Há também casos de uso de textos como bases de entrada dos algoritmos, realizando o processamento de linguagem natural, e gerando aplicações para o campo do reconhecimento de fala, análise de sentimentos e da classificação de textos (HASAN *et al*., 2018; MIAO *et al*., 2018; NASSIF *et al*., 2019).

O advento do "*Big Data*" permitiu as organizações o acesso e a capacidade de armazenamento de grande quantidade de dados, inclusive dados textuais, gerando novas possibilidades de análise e de conhecimento organizacional, o que possibilita melhorar a competitividade destas organizações. Corroboram com esta afirmação Harris e Davenport (2017) que, em seu estudo publicado pela *Harvard Business Review*, citam que as empresas que utilizarem essas novas ferramentas para a criação de oportunidades e melhoria dos processos internos estarão em melhores posições no cenário competitivo futuro.

Ainda que grande parte dessas oportunidades esteja focada em grandes avanços, como monitoramento real do processamento de pedidos e a otimização de rotas de entrega (HARRIS; DAVENPORT, 2017), outras oportunidades podem ser exploradas. Esse estudo se debruça sobre o tema da gestão da informação organizacional de classificação de produtos, tradicionalmente conhecido como cadastro de produtos.

O cadastro de produtos é um tema central dentro das atividades organizacionais, sendo muitas vezes lembrado somente a partir dos impactos negativos que pode causar em uma operação. Afinal, caso a base de cadastros de uma empresa tenha inúmeras inconsistências, todas as análises e decisões decorrentes estarão comprometidas. Apesar disso, não é incomum que inconsistências em cadastros sejam encontradas de forma recorrente em empresas. Para alguns autores, esse problema pode se desenvolver até gerar um "caos" nas informações da empresa (FARIA, 2004; IMAM, 2021).

Problemas com o cadastro de produtos podem impactar clientes organizacionais externos e internos. Por exemplo, num *e-commerce*, ao procurar um produto por meio de suas possíveis características, um cliente pode não o encontrar, caso esteja cadastrado erroneamente. Da mesma forma, dentro da organização, pode ocorrer problemas fiscais, ocasionando multas, caso a classificação fiscal do produto esteja incorreta. Ainda, considerando o foco em operações, no qual se contextualiza esse estudo, um cadastro incorreto de uma característica pode gerar equívocos na avaliação da cadeia de suprimentos. Por exemplo, o crescimento ou a redução de alguma categoria de produto pode estar mascarado por problemas de inexatidão no cadastro dos produtos, levando os tomadores de decisão a crer que determinada categoria é lucrativa quando pode estar ocorrendo o contrário, e vice-versa.

Essa situação se agrava em ambiente onde o nível de inovação é elevado. Isso ocorre porque num contexto de elevada inovação o número de lançamentos de produtos aumenta e, por consequência, a frequência de cadastro cresce, assim como a probabilidade de erros. Somado a esse fato, existe ainda a complexidade operacional, quer advinda de um portfólio com alto número de categorias/famílias de produto ou por um desenho e projeto de toda cadeia de abastecimento que engloba diversas tecnologias de fabricação.

Dessa maneira, o objetivo desse artigo é avaliar o desempenho de diferentes técnicas de aprendizado de máquina e de medidas de similaridade na capacidade de classificação correta de produtos conforme o texto de suas descrições. Em paralelo, apresentar uma proposta de inserção desse tipo de técnica em um processo real. De forma complementar aos artigos de natureza mais técnica, com foco na análise pura do desempenho de determinado algoritmo (geralmente com o suporte de bases de dados compartilhadas e construídas para desenvolvedores), o presente estudo destaca-se pela aplicação comparativa das técnicas de aprendizado de máquina em um estudo de caso, facilitando a aplicabilidade em outros ambientais organizacionais. E também possibilita o apontamento de futuras contribuições no processo decisório no setor de cadastro de produtos.

O presente artigo está organizado da seguinte forma: a seção 2 apresenta a fundamentação teórica; a seção 3, o método; na seção 4 há a apresentação dos resultados e; a seção 5 apresenta as conclusões dessa pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA - TÉCNICAS DE APRENDIZADO DE MÁQUINA E MEDIDAS DE SIMILARIDADE

A categorização de textos por meio de algoritmos tem sido uma atividade amplamente estudada ao longo dos anos (JOHNSON *et al*., 2002; SEBASTIANI, 2005; HARRAG; EL-QAWASMEH; PICHAPPAN, 2009; SHI *et al.,* 2010; SABUNA; SETYOHADI, 2017). De forma geral, envolve a determinação de alguma categoria pré-estabelecida para um texto a partir de uma base de referência previamente rotulada (JOHNSON *et al*., 2002). Existem diversas técnicas de automatização do processo de categorização de texto que podem ocorrem por diversos métodos ou técnicas. Sebastiani (2002) classifica-os da seguinte forma: probabilísticos – com *Naïve Bayes* sendo o mais tradicional; de árvore de decisão; regras de decisão; métodos de regressão; métodos online e de batelada; método Rocchio; redes neurais; baseados em exemplos; entre outros. Nesse estudo, dado a natureza do problema e a ferramenta utilizada para a sua resolução, foram estabelecidas as seguintes técnicas: árvore de decisão; *Naïve Bayes*, redes neurais e, complementarmente, o uso de medidas de similaridade como a similaridade do cosseno, a distância euclidiana e de Manhanttan.

2.1 Árvore de decisão

Algoritmos de árvore de decisão são um dos mais encontrados na literatura para classificação (CHANDRA *et al*., 2002). Em geral, esse algoritmo tem sido utilizado para classificações de dados quantitativos, mas existem estudos que utilizaram entradas qualitativas, como atributos simbólicos do texto (SEBASTIANI, 2002; LAKSHMI *et al*., 2013). A classificação por meio de árvore de decisão representa um processo no qual se estabelece uma árvore a partir de nós internos, dos quais partem ramificações que são utilizadas para determinar, com base um critério de ponderação a partir do texto de entrada, em qual categoria – ou folha da árvore – existe a classificação mais adequada para esse texto (SEBASTIANI, 2002).

Nesse sentido, algumas estratégias podem ser utilizadas, por exemplo, converter as diversas categorias em números (*label encoding*), o que em geral é visto como problemático ao introduzir um possível viés de hierarquia sobre as classes; ou transformar os dados em vetores de ausências e presenças de cada uma das unidades analisadas (*one hot encoding*), o que elimina o viés da hierarquização, mas pode resultar no problema da multidimensionalidade, gerando ruído nos resultados obtidos dependendo do tipo de problema abordado.

Os tipos mais tradicionais de algoritmos de aprendizado de árvore de decisão são: ID3, C4.5 e CART. O ID3 é um dos algoritmos mais conhecidos, sendo caracterizado por ser uma abordagem *top-down* com estratégia de busca “gulosa” (*greedy*, do inglês), que escolhe aquele que melhor classifica baseado em ganho de entropia e ganho de informação após a varredura de cada um dos nós (JIN; DE-LIN, FEN-XIANG, 2009; PENG; CHEN; ZHOU, 2009). Por sua vez, o C4.5 pode ser visto como uma evolução do ID3, ao qual, por exemplo, se endereçaram limitantes como: a possibilidade de uso de dados contínuos, o uso de informações desconhecidas (*missing values*), a capacidade de utilizar atributos com diferentes pesos e a poda após a criação da árvore. Por fim, há o CART (*Classification and Regression Trees*) que conta com algumas variações dos formatos mais tradicionais de árvore de decisão como o teste sendo sempre binário; o uso do índice de Gini para classificar os testes e o uso de métodos complexos de poda a partir de parâmetros estabelecidos por validação cruzada (HSSINA *et al*., 2014; SINGH; GUPTA, 2014).

2.2 *Naïve Bayes*

Naïve Bayes é um algoritmo de classificação probabilística que não considera a correlação entre os termos para realizar a classificação e, por isso, sua atribuição pode ser vista como "ingênua". A característica probabilística advém da forma de classificação, por exemplo, considerando a probabilidade de que um documento representado por um vetor de termos pertença a determinada categoria. Também é conhecido por utilizar o teorema de Bayes para cálculo de probabilidades (SEBASTIANI, 2002).

Apesar das premissas de independência entre as classes, essa técnica tem sido utilizada em estudos de classificação textual, principalmente para os casos em que a probabilidade resultante não é um parâmetro fundamental, como para determinar se determinado e-mail é ou não um *spam* (ZHANG; LI, 2007; QIANG, 2010), e tem sido utilizado com manipulações de *term-frequency* para melhoria dos resultados (KIM *et al*., 2006). Além disso, é vista como uma técnica de simples implementação (KIBRIYA *et al.,* 2004; QIANG, 2010).

2.3 *MLP (Multi-layer perceptron*) / Redes neurais

A classificação por redes neurais pode ser vista como uma rede de unidades, em que as unidades de entrada representam os termos e as de saída as categorias de interesse estabelecidas (SEBASTIANI, 2002). O *perceptron*, como é denominado o modelo desenvolvido para remeter o funcionamento do cérebro e abordar problemas de identificação de padrões, teve suas primeiras aparições na década de 1960, mas caiu em desuso logo em seguida pela limitação constatada de que somente poderiam ser utilizados em problemas linearmente separáveis (RAD; BEHJAT, 2019).

No entanto, no final da década de 1980, por meio da proposta de multicamadas e do treinamento por *backpropagation*, o *perceptron* voltou a ser utilizado com mais popularidade, visto que essa estratégia auxilia que classificações incorretas possam ser revistas e que novas classificações possam ser estipuladas a partir dos erros encontrados, resultando em um processo que minimiza a quantidade de inconsistências na classificação (PALMA NETO; NICOLETTI, 2010).

O MLP é caracterizado por ser uma rede com uma ou mais camadas ocultas com um determinado número de neurônios em cada camada, que lida essencialmente com o problema de linearidade do *perceptron*, possuindo três características básicas: o modelo de neurônio inclui uma função de ativação não linear; as camadas ocultas da rede possuem uma capacidade de aprender padrões complexos de entrada; a conectividade da rede presente no MLP é de alto nível (RAD & BEHJAT, 2019).

2.4 Medidas de similaridade

Medidas de similaridade (ou de distância) têm sido utilizadas no aprendizado de máquina como mecanismos intermediadores de algoritmos de aprendizagem, pois ao retornar índices, permitem agrupar elementos de um *dataset* de forma muito mais efetiva (CALVO-VALVERDE; MENA-ARIAS, 2020). Dentro do contexto em que o problema proposto se insere, as medidas de similaridade podem ser vistas como possíveis formas de classificação, ainda que não realizem as etapas de treinamento e teste como ocorre nas técnicas anteriormente apresentadas (VIJAYMEENA; KAVITHA, 2016).

A categorização acontece pela possibilidade de vetorizar um texto, submetê-lo a uma comparação com outro texto, por meio de métricas que expressem o quão próximo ou distante determinado vetor é de outro, ou seja, aplicando medidas de similaridade. Dentro do contexto de análise de conteúdo, essas são métricas utilizadas para a comparação de textos (DE LIMA *et al*., 2018; BRANDAO, GODINHO FILHO; SILVA, 2021). No entanto, também estão presentes nas tarefas de recomendação em outros contextos (BADRIYAH *et al*., 2017). Deste modo, foram consideradas as medidas de similaridade de distância euclidiana, do cosseno e de Manhanttan para a categorização.

A primeira forma de medida de similaridade é a distância euclidiana, que é calculada a partir da raiz quadrada da soma das diferenças quadradas entre os dois elementos de vetores (VIJAYMEENA; KAVITHA, 2016). Como a explicação de sua forma de cálculo sugere, ela é uma medida geométrica, amplamente utilizada em agrupamentos e categorizações, inclusive de texto (HUANG, 2008).

Por sua vez, o uso de vetores de termos no processamento de texto faz com que eles possam ser estudados pela perspectiva angular, na qual se encontra a similaridade do cosseno. Nesse caso, a correlação entre dois vetores – e como consequência, a similaridade – é quantificada pelo cosseno do ângulo formado por esses dois vetores. Além de ser uma das formas mais populares de análise de textos, uma vantagem desse formato é que, dado a sua avaliação angular, diferenças entre o módulo dos vetores são independentes do resultado gerado por essa medida (HUANG, 2008; GOMAA *et al*., 2013).

Por fim, a distância de Manhattan se configura como a medida resultante de distância entre dois pontos calculados a partir de um caminho traçado a partir de uma grade (VIJAYMEENA; KAVITHA, 2016)

2.5 Medidas adicionais

Como suporte às técnicas apresentadas, foram utilizadas outras medidas comuns no processamento de linguagem natural: a frequência do termo (em inglês, *term frequency* (TF)) e o inverso da frequência no corpus (em inglês, *inverse document frequency (*IDF)), combinados por meio da sua multiplicação que é conhecida como TF-IDF (RAD; BEHJAT, 2019). O TF-IDF combina o cálculo do peso de cada palavra com a frequência do termo no documento e ao número de documentos contendo esse termo em relação ao total de documentos do corpus, destacando-se os termos mais relevantes e que diferenciam os textos. (CALVO-VALVERDE; MENA-ARIAS, 2020).

3 MÉTODO

Nessa seção são apresentadas as etapas de investigação, análise e pré-tratamento dos dados, construção dos modelos e análise das técnicas de aprendizado de máquina para a resolução do problema proposto.

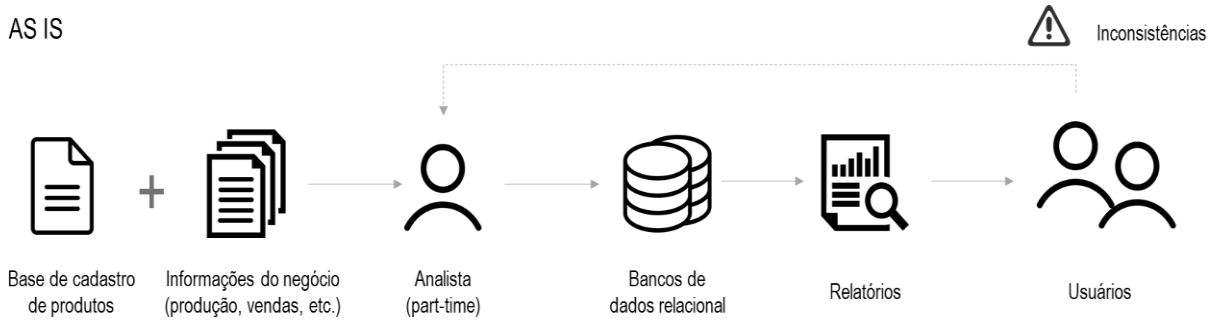
3.1 Caracterização geral da empresa, do processo e do tipo de dado

O estudo foi conduzido considerando uma empresa multinacional brasileira líder do setor de Higiene Pessoal, Perfumaria e Cosméticos (CFT: *Cosmetic, Fragrance and Toiletry*, do inglês) no Brasil e referência em inovação e sustentabilidade. Possui operações concentradas na América Latina, com destaque, além do Brasil, para Argentina, Chile, México, Colômbia e Peru. Por conta do alto potencial de inovação, possui um portfólio com alta rotatividade de produtos, os quais acabam tendo, em geral, ciclos de vida mais curtos. Possui uma operação complexa e dinâmica, na qual, o processo de cadastro de produtos figura como aspecto fundamental para a garantia da coordenação da informação entre as diversas áreas da empresa.

No processo de cadastro analisado, um analista dedicado parcialmente à atividade de cadastro de informações classifica determinado produto acabado em diversos níveis, juntamente com outros dados do negócio, como vendas e produção mensal e atualiza um banco de dados relacional da empresa.

Esse banco de dados utiliza as informações de cadastro para gerar relatórios utilizados como referências em análise de desempenho. No entanto, essa atividade é feita de forma manual e cada nível do cadastro de produto é feito com base na experiência do analista ou em alguma consulta de similaridade textual, na qual ele busca por produtos próximos com alguma parte da descrição, o que torna o processo moroso e sujeito a falhas, que regularmente são identificadas pelos usuários. Isto gera uma necessidade de retrabalho não somente ao analista em reclassificar a informação, como também nas análises geradas a partir dessa informação (Figura 1). Uma análise aleatória dos cadastros na base de dados revelou uma inconsistência (códigos cadastrados incorretamente ou incompletos) de 20% em média.

Figura 1 - Processo de cadastro de produtos



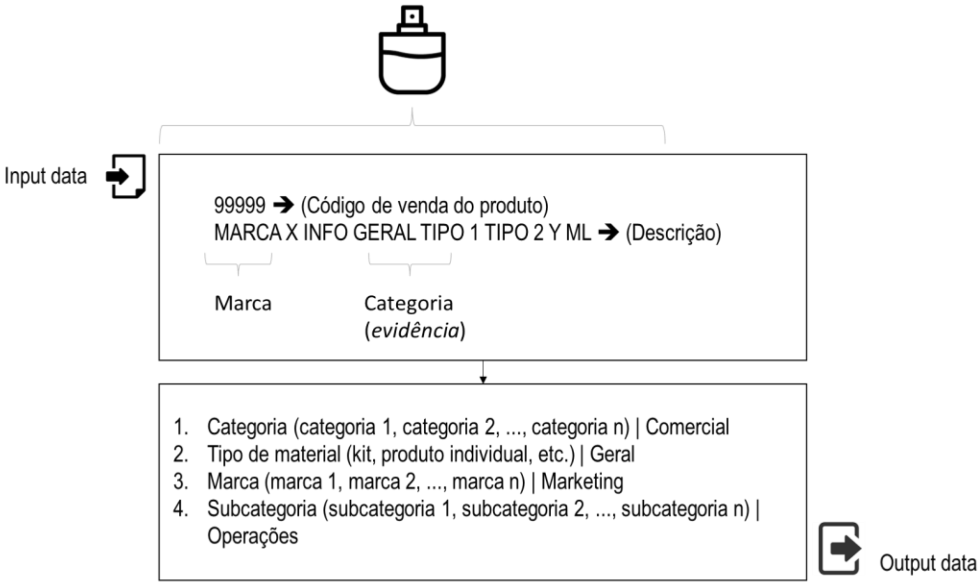
Fonte: os autores.

Os dados analisados são “*strings*”, ou seja, cadeias de caracteres que compõem a descrição dos produtos finais produzidos ou comprados pela empresa analisada, que estão na base de cadastro de produtos. Os dados analisados são classificados de acordo com suas características mercadológicas e operacionais, por grupos categóricos não ordinais.

Os dados utilizados para a construção dos modelos de aprendizado de máquina foram obtidos da base de cadastro da empresa analisada. A Figura 2 representa a forma como eles são estruturados, considerando quais atributos são de entrada e quais de saída. Na Figura 2, as informações de entrada no processo de cadastro – e também parcialmente utilizadas nesse estudo – são: código de venda e descrição do produto.

De forma geral, diferentemente do que pode ocorrer em outras empresas, o código de venda do produto não segue uma estratégia de atribuição de um padrão regular de caracterização, por exemplo, em que determinado número indica um tipo de informação, outro número, uma informação diferente, etc., o que consequentemente impõe um ambiente mais complexo de análise. No entanto, essa estratégia acaba sendo adotada porque, considerando o alto nível de inovação, se torna difícil mensurar efetivamente a quantidade de posições necessárias para alocar cada categoria, ficando saturada muito facilmente.

Figura 2 - Exemplo de classificação



Fonte: os autores.

Dessa maneira, a informação que realmente poderia ser utilizada nesse processo é a descrição do produto, que além da marca, permite encontrar evidências do tipo de produto e da categoria. Sendo tomado como premissa, torna-se importante destacar que a elaboração da descrição do produto também representa outra tarefa complexa dentro do contexto de desenvolvimento de produtos. Rueckel *et al*. (2005) destacam que a criação da descrição do produto envolve a coleta e análise de informações de diferentes áreas que precisam estar contempladas para que a descrição criada não seja incompleta. No entanto, a tarefa de criação da descrição está fora do escopo pela natureza do problema abordado nessa pesquisa.

Adicionalmente, ainda que o tipo de material esteja indicado como uma saída do processo, frequentemente este também é cadastrado no sistema ERP da empresa, junto com o código de venda e com a descrição e por isso pouco foco será dado nesse elemento na proposta preditiva do modelo de aprendizagem.

Como resultado do processo, um conjunto de classificações é gerado como apresentado na Tabela 1. Considerando a quantidade de possíveis distinções de cada um dos campos avaliados, tem-se: dois para o tipo – produto acabado, que seria basicamente o que classifica entre: item individual e kit com vários produtos; para categoria existem 21 tipos, dos quais 17 são úteis para o objetivo desse estudo

– o detalhamento sobre o processo de inclusão/exclusão está descrito na próxima seção; para subcategoria foram considerados elegíveis 73 das 90 possíveis; e, para marcas foram identificadas 70, mas utilizadas somente 56.

Tabela 1 - Caracterização do *dataset* utilizado

Nome do campo	Descrição	Tipo
Descrição do produto	Nome atribuído ao produto final da empresa, geralmente contém uma identificação da marca ao qual se refere, alguma indicação da categoria e do tipo de produto e a volumetria/ peso	Cadeia de caracteres
Tipo	Corresponde ao código do tipo de produto. Se é um item acabado individual, um agrupamento de produtos - estojo, ou uma matéria-prima	Categórico não ordinal
Categoria	Primeiro nível de classificação dos produtos, geralmente associado à perspectiva mercadológica, faz a distinção em grandes famílias de produtos	Categórico não ordinal
Subcategoria	Característica do produto que diferencia a tecnologia de produção utilizada. É uma denominação interna utilizada principalmente pela área de operações.	Categórico não ordinal
Marca	Atributo de marketing que diferencia as diversas comunicações e públicos da marca principal	Categórico não ordinal

Fonte: os autores.

Cabe destacar, porém, certos detalhes dos campos categorizáveis e suas distinções. Apesar de associado inicialmente ao contexto comercial, as categorias de produtos geralmente são mais gerais e amplas, por exemplo, no contexto de uma montadora, uma categoria de produtos poderia ser de carros e outra poderia ser de utilitários. No caso estudado, no cenário de itens de cosméticos e de bem-estar, uma categoria poderia ser perfumaria e outra maquiagem.

Por sua vez, manter unicamente uma divisão muito ampla acaba não sendo a realidade das empresas, que precisam alocar e avaliar os produtos com distintos níveis de detalhamento. Por isso, campos que dividem as categorias são comuns e geram as subcategorias. No presente caso, a subcategoria está atendendo às necessidades das áreas de operações, por isso o recorte está no tipo de tecnologia emprega – e que está associada às características dos produtos. Para o caso da indústria cosmética, uma subcategoria da maquiagem poderia ser batom e outra base, que em geral possuem processos produtivos distintos.

Por fim, o campo marca está associado ao marketing, pois lida basicamente com as marcas envolvidas no negócio e podem abranger distintas categorias e subcategorias, sendo, portando, um campo não dependente de categoria e subcategoria.

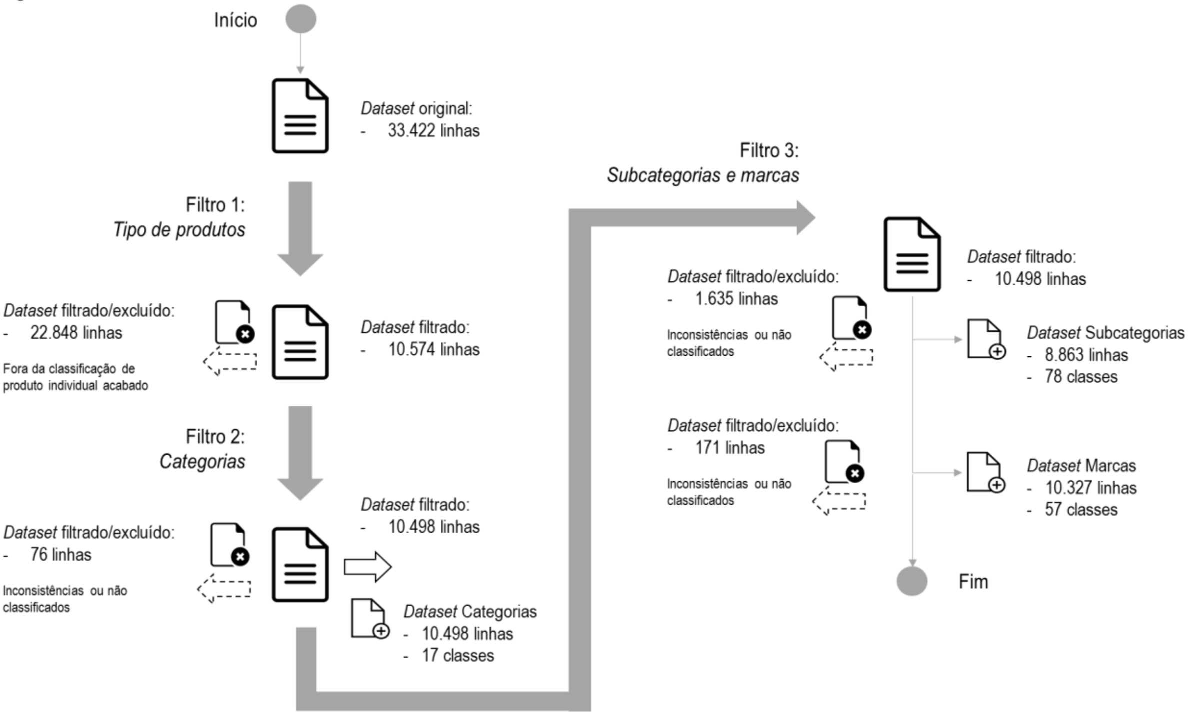
Em termos teóricos, todos esses níveis geram combinações que extrapolam o patamar de milhares. No entanto, considerando as restrições e configurações existentes, esse número cai para 894, que ainda é alto considerando que a atividade é feita de forma manual.

3.2 Tratamento da base de dados

O *dataset* original continha um total de 33.422 observações, das quais 16,52% eram de produtos que estão fora do escopo dessa análise por se tratar de materiais sub-processados ou de produtos acabados suplementares – para os quais não se aplicam as classificações de categoria, subcategoria ou marca. Adicionalmente, 49,63% de kits e o restante, 31,64%, de produtos acabados individuais. Considerando que foi identificado que a maior complexidade de classificações estava no tipo produto acabado individual, os demais foram excluídos no filtro 1.

Por exemplo, ainda que a quantidade itens classificados como tipo estojo seja representativa nesse *dataset*, a quantidade de classificações existentes é muito baixa, pois, em geral, eles são classificados como *kits*, por conta do tipo de produto. Além disso, mantê-los na base poderia gerar uma complexidade desnecessária aos algoritmos, pois muitos contêm informações similares aos itens individuais em suas descrições. Outro exemplo, é que em um kit que possui na descrição uma categoria *n* poderia, ao invés de ser classificado como *kits*, ser classificado como categoria *n*, o que não é desejável dentro do entendimento organizacional de cadastro. O tratamento de dados para a limpeza do *dataset* que foi realizado nesse estudo ocorreu conforme ilustrado na Figura 3:

Figura 3 - Processo de tratamento dos dados



Fonte: os autores.

Após a filtragem inicial do campo “Tipo”, realizou-se a segunda filtragem contemplando a coluna de categorias, apresentando o nível mais macro da classificação quando comparado à subcategoria e marca. Nesse momento, buscou-se na base observações inconsistentes, ou seja, itens do tipo: produto acabado individual, classificados como *kits* ou observações faltantes e incompletas.

Para esses casos, dado a natureza do dado, optou-se pela exclusão dos mesmos no *dataset*. O mesmo procedimento de busca por inconsistências aconteceu para os demais níveis. Após isso, cada arquivo final – indicado pelo ícone de *file* acrescido pelo símbolo de soma (+) na Figura 3, foi utilizado como entrada nos modelos de aprendizagem.

3.3 Modelos criados de aprendizado de máquina

Nessa seção são apresentados os modelos criados para a resolução de problemas de categorização, conforme proposto nesse estudo, e contempla as duas principais abordagens: a de classificação por meio de coeficientes de similaridade e de algoritmos de aprendizado de máquina.

3.3.1 Classificação por coeficientes de similaridade

A Figura 4 ilustra esquematicamente o modelo criado no KNIME para realizar a tarefa de classificação proposta. Antes da etapa 1, foi realizada a entrada dos dados, que ocorreu via arquivo no formato Excel (.xlsx).

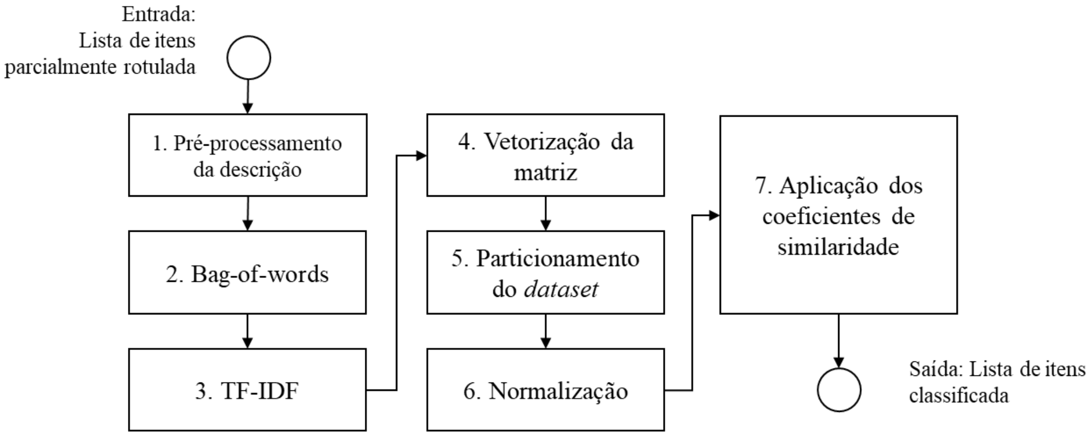
Em seguida, foi realizado um pré-processamento dos dados em que basicamente os dados foram transformados em caixa alta. A seguir, as *strings* de entrada foram transformadas em documento e ordenadas de forma ascendente. Em seguida, os dados foram particionados em *bag of words*.

Após isso, foi calculado o TF-IDF e também realizada a vetorização das descrições dos produtos de entrada, bem como a normalização do vetor. Assim os dados foram filtrados para manter somente uma matriz n (*strings*: *labels* pré-classificadas das descrições) x m (*double*: colunas com os termos vetorizados), que foram particionados (com 30% da base separada de forma aleatória) e serviram de entrada para os nós de cálculo de similaridade ("*Similarity Search"*).

Foram considerados os coeficientes de similaridade de cosseno, distância euclidiana e de Manhattan. Os parâmetros de distância foram configurados para retornar somente a observação correspondente à similaridade mais próxima e, por isso, a similaridade máxima possível foi estabelecida em 0,99999 (para evitar que se retorne a própria descrição), enquanto a mínima foi estabelecida em 0,8 para evitar o retorno de casos significativamente destoantes.

Cabe destacar que testes exploratórios foram realizados anteriormente, variando o número de vizinhos similares e também o intervalo de retorno, contudo a configuração apresentada à priori foi a que obteve os melhores resultados, e por essa razão são apresentadas na próxima seção.

Figura 4 - Modelo para classificação por similaridade



Fonte: os autores.

3.3.2 Classificação por algoritmos de aprendizado de máquina (árvore de decisão, *Naïve Bayes* e MLP

Da mesma forma que ocorreu no modelo apresentado na Figura 4, a Figura 5 contém uma etapa de entrada, pré-processamento de dados, cálculo do TF-IDF e vetorização das descrições dos produtos.

Os dados foram particionados, normalizados e serviram de entrada para os nós de treinamento (70% da base). O modelo treinado foi utilizado para predizer o conjunto de dados separados para teste, correspondendo a 30% da base total). Como parte de uma etapa exploratória, várias configurações foram testadas para cada uma das técnicas, realizando um tipo de análise de sensibilidade, na qual um parâmetro é alterado isoladamente para se observar o impacto no desempenho, que neste estudo são as medidas de qualidade do modelo.

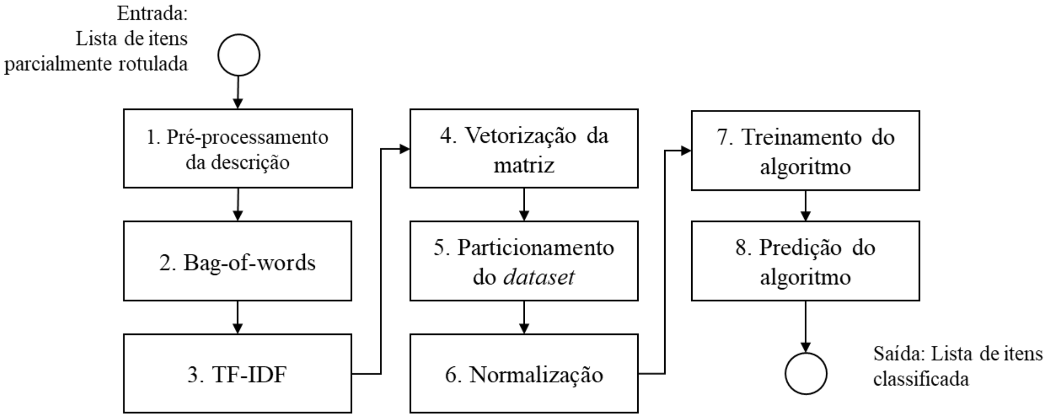
Considerando a técnica de árvore de decisão, no KNIME, o modelo existente se aproxima mais do C4.5 e, sendo uma extensão do ID3, utiliza o conceito de entropia para as decisões de análise e avança em termos de estratégias de poda e dados incompletos. Considera parâmetros como o da medida de qualidade (índice de Gini e de ganho), estratégia de poda e número mínimo de elementos por ramo.

Por sua vez, para *Naïve Bayes*, o KNIME considera a distribuição normal para atributos numéricos e possui como parâmetros de entrada a probabilidade padrão e o número máximo de valores únicos nominais por atributo.

Por fim, para MLP, o KNIME possui como parâmetros o número máximo de iterações, de camadas ocultas (máx. 100) e de neurônios por camada (máx. 100).

Ao final se obteve uma configuração mais adequada de acordo com a técnica aplicada. Por exemplo, para classificação probabilística de *Naïve Bayes*, os resultados foram melhores com probabilidade padrão de 0,5; para MLP com uma camada oculta e o número máximo possível de neurônios (i.e. n=100); e por fim para árvore de decisão a medida de qualidade como taxa de ganho, sem método de poda e mínimo de classificações por nó igual a um. E essas configurações estão apresentadas detalhadamente nas Tabelas que consolidam os resultados dos experimentos (Tabelas 2, 3 e 4).

Figura 5 - Modelo representativo para classificação por meio das técnicas de aprendizado de máquina de árvore de decisão, *Naïve Bayes* e MLP



Fonte: os autores.

Os problemas de classificação possuem medidas de qualidade dos modelos que são tradicionais na literatura para os algoritmos apresentados exceto para as medidas de similaridade, a saber: acurácia, recall, precisão, F-1, matriz de confusão (LEAL, 2017). Adicionalmente, para contemplar a avaliação das medidas de similaridade será considerado como critério de qualidade do modelo o tempo de execução e a quantidade de itens retornados.

4 RESULTADOS E DISCUSSÃO

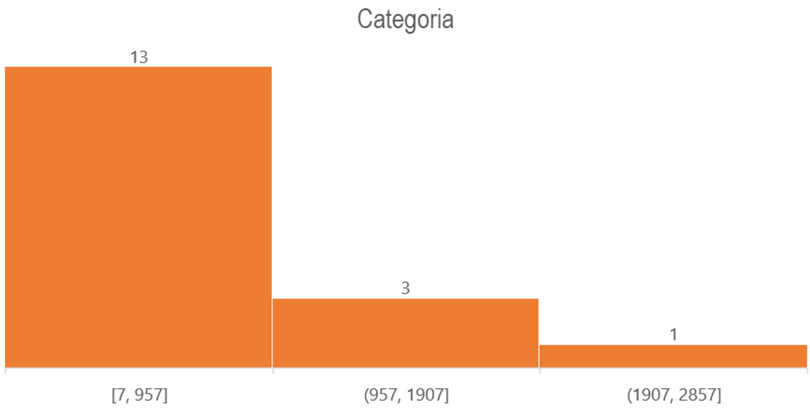
Essa seção apresenta e discute os resultados das avaliações realizadas e está dividida entre resultados descritivos e dos experimentos realizados.

4.1 Análise e Exploração dos Dados

Por conta da estrutura dos dados (“*strings*”), foi possível realizar a exploração dos dados por meio de estatísticas descritivas, como média e desvio-padrão, em conjunto com a distribuição de frequências de observações, que estão ilustradas por histogramas e são apresentadas nas Figura 6, 7 e 8.

Figura 6 - Análise descritiva dos dados do campo categoria

Categoria	
Média	618
Erro padrão	169
Mediana	306
Desvio padrão	697
Variância da amostra	486.114
Curtose	2
Assimetria	2
Intervalo	2.428
Mínimo	7
Máximo	2.435
Soma	10.498
Contagem	17
CV	1,129



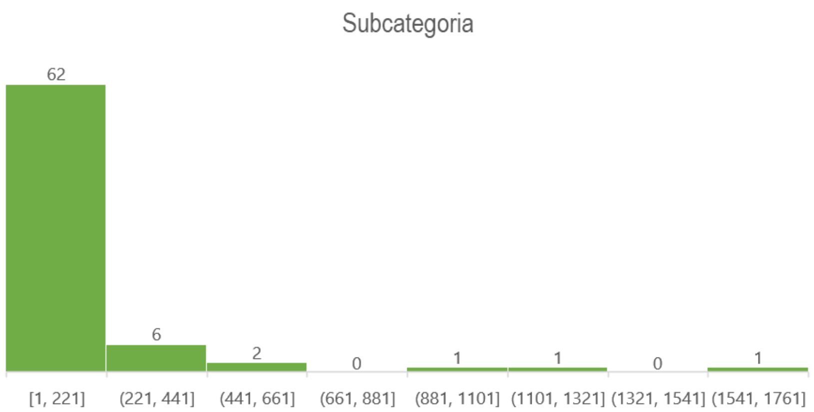
Fonte: os autores.

O primeiro campo é o de categoria, que possui a maior parte das classes com quantidade de observações no *dataset* e que varia entre 7 e 957, ou seja, uma das *n* categorias possui somente 7 produtos enquanto outra, na qual se observa a presença de muito mais itens, conta com 957 produtos.

Apesar do coeficiente de variação (CV) do campo “*categoria*” ser o menor observado quando comparada à “*subcategoria*” ou “*marca*”, ainda é maior do que 1, indicando uma alta dispersão dos dados. Pela a média estar posicionada no intervalo entre 7 e 957 do histograma e corresponder ao valor de 618, e a mediana apresentar o valor de 306, há indicação de uma maior assimetria à esquerda, revelando a predominância de classes com menores quantidades de produtos atrelados.

Figura 7 - Análise descritiva dos dados do campo subcategoria

Subcategoria	
Média	121
Erro padrão	31
Mediana	13
Modo	13
Desvio padrão	269
Variância da amostra	72.173
Curtose	17
Assimetria	4
Intervalo	1.622
Mínimo	1
Máximo	1.623
Soma	8.863
CV	2,213

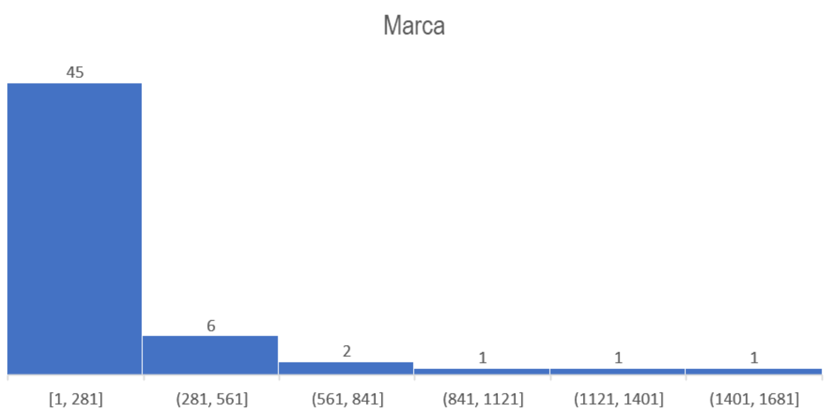


Fonte: os autores.

Um comportamento similar ocorre para o campo da subcategoria (Figura 7), no entanto, com um comportamento mais disperso e heterogêneo do que a categoria. Como pode ser observado, no primeiro intervalo do histograma está concentrado aproximadamente 85% das classes, e, assim como na categoria a mediana (=13) é menor do que a média (=121), há uma indicação de assimetria muito expressiva com uma grande quantidade de classes com poucos produtos, o que em termos de complexidade de classificação traz dificuldade para os modelos.

Figura 8 - Análise descritiva dos dados do campo marca

Marca	
Média	184
Erro padrão	40
Mediana	72
Modo	1
Desvio padrão	301
Variância da amostra	90.885
Curtose	7
Assimetria	3
Intervalo	1.428
Mínimo	1
Máximo	1.429
Soma	10.327
CV	1,635



Fonte: os autores.

Conforme identificado na Figura 8, o campo da marca se posiciona entre esses dois extremos (categoria/ subcategoria), também com a maior parte das classes dentro de um intervalo menor de valores e coeficientes de variação maiores do que um.

4.2 Resultados experimentais

Os experimentos realizados nesse estudo foram executados em um computador com processador *Intel Core i7-4700MQ 2,4GHz* com 8GB de memória RAM e sistema operacional *Windows 10*. Todo modelo foi projetado e implementado por meio do *KNIME Analytics Plataform v3.6.0*. Os experimentos foram divididos em dois blocos principais considerando o tamanho da amostra utilizada n=3500, correspondendo a aproximadamente 10% do *dataset* original. Os *datasets* tratados na etapa prévia de ingestão dos dados no KNIME são expostos na Figura 3. A subdivisão em classes de experimentos com base no número de amostra acorreu no momento da execução dos modelos. No entanto, ao verificar-se que para todos os campos os modelos baseados em similaridade não retornavam os resultados em um intervalo menor que 8 horas, chegou-se à conclusão de que existia um esforço computacional muito elevado para o tipo de tarefa executada, e que essa condição seria limitante para o contexto do processo analisado.

Dessa maneira, para os demais casos, a forma de seleção dessa subdivisão do *dataset* foi feita de forma aleatória por meio nó "*Partitioning*" no KNIME. Ainda que pontual, essa divisão também contribuiu para avaliar a sensibilidade do modelo em relação ao número de observações, podendo ser um parâmetro interessante de alteração.

De forma geral, os modelos que envolveram aprendizado de máquina obtiveram um resultado melhor para o campo de categoria, seja em termos de porcentagem de classificação quanto em tempo de execução, exceto para *Naïve Bayes*. Além disso, quando se aumenta para o *dataset* inteiro, verifica-se uma melhoria no desempenho dos modelos, implicando em uma decisão importante para obter um resultado melhor e maior tempo ou vice-versa.

Quadro 1 - Resultados dos experimentos para categoria (continua)

n	Modelo	% classif.	Acurácia	Precisão	Recall	F-means	t total execução (ms)	Observações
3.500	(1) TD-IDF/ Distância de Manhatan	1%	N/A	N/A	N/A	N/A	378.577	Intervalo de similaridade: 0,000001 - 0,999999/ Vizinhos: 1
	(2) TD-IDF/ Distância de Manhatan	1%	N/A	N/A	N/A	N/A	335.456	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	(1) TD-IDF/ Similaridade de Cosseno	99%	N/A	N/A	N/A	N/A	384.728	Intervalo de similaridade: 0,000001 - 0,999999/ Vizinhos: 1
	(2) TD-IDF/ Similaridade de Cosseno	9%	N/A	N/A	N/A	N/A	408.039	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	(1) TD-IDF/ Distância euclidiana	8%	N/A	N/A	N/A	N/A	361.032	Intervalo de similaridade: 0,000001 - 0,999999/ Vizinhos: 1
	(2) TD-IDF/ Distância euclidiana	5%	N/A	N/A	N/A	N/A	354.293	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	(1) One hot encoding/ Similaridade de Cosseno	99%	N/A	N/A	N/A	N/A	396.073	Intervalo de similaridade: 0,000001 - 0,999999/ Vizinhos: 1
	(2) One hot encoding/ Similaridade de Cosseno	12%	N/A	N/A	N/A	N/A	399.699	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	(1) TF-IDF/ Naive-Bayes	100%	0,677	0,737	0,488	0,605	15.938	Default KNIME: Probabilidade default: 0,0/ Número máximo de valores único nominais por atributo: 17
	(2) TF-IDF/ Naive-Bayes	100%	0,350	0,751	0,114	0,317	89.308	Sensibilidade prob: Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 17
	(1) TF-IDF/ Multilayer perceptron	100%	0,799	0,692	0,593	0,745	32.245	Default KNIME - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 10
	(2) TF-IDF/ Multilayer perceptron	**100%**	**0,845**	**0,682**	**0,653**	**0,745**	248.440	**Sensibilidade neurônios (máx=100) - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 100**
	(1) TF-IDF/ Decision Tree	100%	0,802	0,823	0,584	0,745	**8.212**	Default KNIME - Medida de qualidade: Gini/ Número mínimo de obs por nó: 2/ Método de poda: Sem poda
	(2) TF-IDF/ Decision Tree	100%	0,777	0,629	0,604	0,685	80.463	**Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda**

Quadro 1 - Resultados dos experimentos para categoria (conclusão)

n	Modelo	% classif.	Acurácia	Precisão	Recall	F-means	t total execução (ms)	Observações
10.498	TD-IDF/ Distância de Manhatan	N/A	N/A	N/A	N/A	N/A	N/A	Tempo de execução > 8h
	TD-IDF/ Similaridade de Cosseno	N/A	N/A	N/A	N/A	N/A	N/A	Tempo de execução > 8h
	One hot encoding/ Similaridade de Cosseno	N/A	N/A	N/A	N/A	N/A	N/A	Tempo de execução > 8h
	(1) TF-IDF/ Naive-Bayes	100%	0,685	0,860	0,530	0,701	366.137	Default KNIME: Probabilidade default: 0,0/ Número máximo de valores único nominais por atributo: 20
	(2) TF-IDF/ Naive-Bayes	100%	0,331	0,861	0,111	0,249	622.194	Sensibilidade prob: Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 20
	(1) TF-IDF/ Multilayer perceptron	100%	0,868	0,701	0,703	0,818	398.996	Default KNIME - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 10
	(2) TF-IDF/ Multilayer perceptron	**100%**	**0,906**	**0,809**	**0,788**	**0,839**	1.444.282	Sensibilidade neurônios (máx=100) - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 100
	(1) TF-IDF/ Decision Tree	**100%**	**0,895**	**0,889**	**0,769**	**0,845**	**443.093**	Default KNIME - Medida de qualidade: Gini/ Número mínimo de obs por nó: 2/ Método de poda: Sem poda
	(2) TF-IDF/ Decision Tree	**100%**	**0,918**	**0,893**	**0,844**	**0,861**	1.465.361	**Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda**

Fonte: os autores.

Considerando os resultados obtidos para subcategoria, verificou-se que este foi o campo no qual os modelos tiveram maior dificuldade de classificação, o que era esperado, considerando tanto a quantidade de classes existentes quanto o perfil de distribuição das observações (Figura 6, gráfico 2).

Em termos de desempenho, as classificações por meio de similaridade tiveram o pior resultado seja em termos de capacidade de classificação quanto em tempo de execução. Por sua vez, o modelo baseado na árvore de decisão teve o melhor resultado geral considerando os grupos de experimentos, com tamanho de amostra 3.500 e com a base completa.

O algoritmo com o *Naïve Bayes* teve melhor desempenho na medida de precisão, indicando uma melhor classificação positiva (verdadeiro e falso positivo), mas quando comparado com outras medidas, verificou-se uma discrepância importante em relação à árvore de decisão, e por isso, somente a precisão não seria suficiente para sustentar a manutenção desse modelo como o de melhor desempenho.

Adicionalmente, comparando os dois resultados do modelo baseado em árvore de decisão, verifica-se um resultado próximo em termos das métricas de avaliação, mesmo tendo diferentes tamanhos de amostra e impactando no tempo de execução do modelo.

Por isso, é importante entender se existe diferença estatística significativa entre essas médias para auxiliar na decisão sobre a escolha do tamanho da amostra. Assim, foram realizados testes Kolmogorov-Smirnov para teste de normalidade nas bases e constatou-se que seguem distribuição normal e depois realizou-se um teste T de comparação de médias e verificou-se que não existem evidências para considerá-las diferentes (não rejeita H0). Logo, a estratégia de utilizar uma base menor retorna o mesmo patamar de acurácia com um esforço computacional significativamente menor (aprox. 10 vezes).

Considerando o resultado dos experimentos para o campo de “Marca”, tem-se os resultados apresentados na Tabela 4. Nesse caso, diferentemente da subcategoria, as classificações por similaridade retornaram os melhores resultados nos indicadores de avaliação, porém com um esforço computacional muito alto e também com baixa capacidade de retorno dentro do intervalo estabelecido para o parâmetro de similaridade.

Quadro 2 - Resultados dos experimentos para subcategoria

n	Modelo	% classif.	Acurácia	Precisão	Recall	F-means	t total exe. (ms)	Observações
3.500	TD-IDF/ Distância de Manhatan	3%	N/A	N/A	N/A	N/A	927.240	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TD-IDF/ Distância Euclidiana	17%	N/A	N/A	N/A	N/A	927.092	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TD-IDF/ Similaridade de Cosseno	34%	N/A	N/A	N/A	N/A	921.454	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	One hot encoding/ Cosseno	44%	N/A	N/A	N/A	N/A	245.535	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TF-IDF/ Naive-Bayes	**100%**	0,309	**0,703**	0,068	0,360	105.360	Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 90
	TF-IDF/ Multilayer perceptron	N/A	N/A	N/A	N/A	N/A	N/A	Não foi capaz de executar
	TF-IDF/ Decision Tree	**100%**	**0,729**	0,581	**0,487**	**0,633**	**62.471**	Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda
8.863	TF-IDF/ Naive-Bayes	100%	0,602	**0,678**	0,335	0,562	**757.644**	Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 90
	TF-IDF/ Multilayer perceptron	N/A	N/A	N/A	N/A	N/A	N/A	Não foi capaz de executar
	TF-IDF/ Decision Tree	**100%**	**0,769**	0,530	**0,542**	**0,649**	820.506	**Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda**

Fonte: os autores.

Em seguida, os modelos baseados em árvore de decisão e MLP também obtiveram resultados aceitáveis, mas com o MLP tendo um esforço computacional maior do que a árvore de decisão para o caso de n=3.500. Quando se observa os resultados com o uso do *dataset* completo, verifica-se um desempenho superior, mas também com custo computacional maior.

No entanto, o modelo com base em MLP apresentou um esforço computacional impraticável dentro das possibilidades de aplicação prática. Por isso, ponderando os resultados das métricas de avaliação com a quantidade de observações classificadas e esforço computacional necessário, mais uma vez o modelo baseado em árvore de decisão pareceu fornecer o melhor resultado geral.

Quadro 3 - Resultados dos experimentos para marca

n	Modelo	% classif.	Acurácia	Precisão	Recall	F-means	t total exe. (ms)	Observações
3.500	TD-IDF/ Distância de Manhatan	3%	N/A	N/A	N/A	N/A	997.993	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TD-IDF/ Distância Euclidiana	18%	N/A	N/A	N/A	N/A	998.061	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TD-IDF/ Similaridade de Cosseno	35%	N/A	N/A	N/A	N/A	998.530	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	One hot encoding/ Cosseno	38%	N/A	N/A	N/A	N/A	768.471	Intervalo de similaridade: 0,8 - 1/ Vizinhos: 1
	TF-IDF/ Naive-Bayes	**100%**	0,558	**0,933**	0,188	0,569	80.945	Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 90
	TF-IDF/ Multilayer perceptron	**100%**	0,860	0,732	0,672	0,802	364.715	Sensibilidade neurônios (máx=100) - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 100
	TF-IDF/ Decision Tree	**100%**	0,844	0,698	0,701	0,850	**58.447**	Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda
10.327	TF-IDF/ Naive-Bayes	100%	0,544	0,931	0,170	0,470	**1.077.221**	Probabilidade default: 0,5/ Número máximo de valores único nominais por atributo: 90
	TF-IDF/ Multilayer perceptron	100%	0,895	0,815	0,691	0,850	1.230.810	Sensibilidade neurônios (máx=100) - Número máximo de iterações: 100/ Número de camadas ocultas: 1/ Número de neurônios ocultos por camada: 100
	TF-IDF/ Decision Tree	**100%**	**0,922**	**0,932**	**0,794**	**0,905**	1.216.612	**Sensibilidade Número de nós - Medida de qualidade: Ganho/ Número mínimo de obs por nó: 1/ Método de poda: Sem poda**

Fonte: os autores.

Também, foi possível observar e constatar estatisticamente que existe correlação negativa entre o número de classes de um campo e o desempenho geral dos modelos de aprendizado ao final das rodadas experimentais. Assim, o pior desempenho obtido no campo da subcategoria frente aos demais é parcialmente explicado pelo número maior de classificações que esse campo possui. Dessa maneira, uma boa prática seria tentar reduzir a quantidade de categorias para melhorar os resultados de predição dos modelos.

Por fim, a Tabela 5 resume os resultados do modelo. Considerando as compensações existentes, verifica-se que as estratégias de similaridade tiveram resultados insatisfatórios mesmo quando comparado entre as métricas pertinentes, de porcentagem retornada (73% inferior aos algoritmos de aprendizado de máquina) e tempo de execução (em média 25% superior aos algoritmos de aprendizado de máquina).

Por sua vez, dentre as técnicas de aprendizado de máquina, *Naïve Bayes* teve o pior desempenho geral, provavelmente pelo fato da quantidade elevada de classificações existentes nos campos avaliados, prejudicando as medidas de qualidade de acurácia, recall e F-means. No entanto, cabe destacar que esse algoritmo foi o que obteve o melhor tempo de processando e que, mesmo sendo descartado para esse caso, pode ser investigado em outras aplicações cuja estrutura de categorização dos produtos não seja muito diversificada. Ao final, restaram dois modelos com melhor desempenho médio: as técnicas de árvore de decisão e de MLP (Tabela 5).

Quadro 4 - Resultados globais e médios dos experimentos

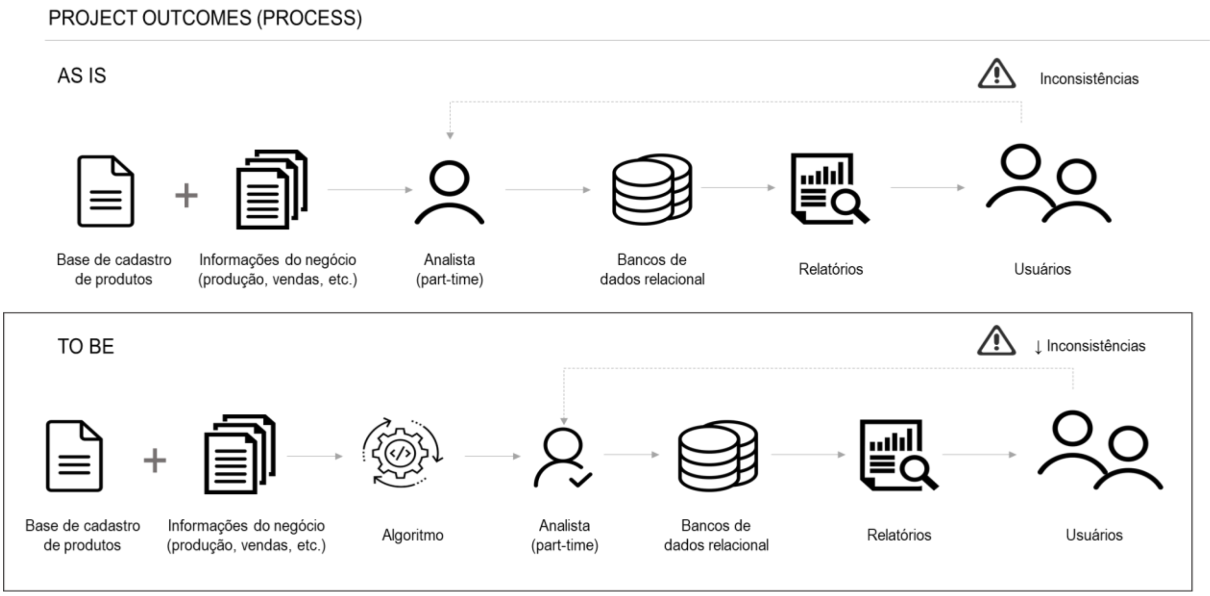
Modelo	Média de % classif.	Média de Acurácia	Média de Precisão	Média de Recall	Média de F-means	Média de t total exe. (ms)
Árvore de decisão	1,000	0,832	0,747	0,666	0,772	519.396
MLP	1,000	0,863	0,739	0,684	0,800	619.915
Naive Bayes	1,000	0,507	0,807	0,251	0,479	389.343
Similaridade	0,267	N/A	N/A	N/A	N/A	612.642

Fonte: os autores.

Apesar de serem satisfatórias por possuírem uma diferença média entre os parâmetros de qualidade de aproximadamente 2,2%, o tempo médio de execução do MLP é aproximadamente 20% mais elevado do que o do algoritmo de árvore de decisão, fazendo que a diferença nas outras métricas seja compensada pela capacidade de implementação deste. Por isso, para esse problema e contexto, o modelo de aprendizado de máquina baseado em árvore de decisão foi eleito como recomendação para o novo processo de classificação de produtos na empresa estudada.

Assim, considerando processo original, a implantação dessa inteligência modifica o processo de negócio de cadastramento e classificação do produto, fazendo com que o papel do analista seja migrado para um avaliador dos resultados das predições obtidas mais do que o executor manual da atividade, e uma análise dos novos códigos cadastrados revelou uma redução na quantidade de erros por cadastro incorreto ou inconsistente de uma média de 20% para 6% (Figura 9).

Figura 9 - *As Is* e *To-Be* do processo de cadastro de produtos



Fonte: os autores.

5 CONCLUSÕES

O presente estudo buscou avaliar diferentes técnicas de aprendizado de mpaquina na capacidade de classificar corretamente produtos de acordo com o texto de suas descrições em um ambiente de elevado nível de inovação. Para isso, foi desenvolvido modelos de aprendizado de máquina no KNIME, que utilizaram basicamente estratégia de bag-of-words, TF-IDF e o uso de algoritmos de aprendizado de máquina e de similaridade para trazer soluções para o problema proposto e essas soluções foram avaliadas em termos de desempenho em diversas métricas tradicionais da área, como precisão, recall, F-means, etc.

Da análise dos resultados, chegou-se à conclusão de que, para o perfil de dados utilizado, os algoritmos de árvore de decisão e MLP foram os que obtiveram os melhores resultados em geral. No entanto, pelo esforço computacional demandado, a escolheu-se o modelo desenvolvido que contempla o uso da árvore de decisão, que revelou potencial de reduzir inconsistências no cadastro da ordem de 20% para 6%, o que é uma melhoria significativa dado o contexto analisado em que se tem um alto volume e alta frequência de entrada de novos produtos para classificação.

Por fim, esse estudo possui limitações que podem ser exploradas em etapas de melhoria do processo. Dentre elas, considerando a relação entre tamanho de amostra, tempo de execução e desempenho das métricas de classificação. Para trabalhos futuros, pode ser investigado se há um número n de amostra que otimiza esses três parâmetros de acordo com o campo predito.

Além disso, como mencionado anteriormente, poder-se-ia revisar a estrutura das classificações de cada campo para reduzir os casos com poucas observações – sem deixá-los demasiadamente genéricos – também seria outra possibilidade para melhorar o desempenho preditivo do modelo baseado em árvore de decisão.

Outro ponto é que, apesar do uso do TF-IDF se mostrar satisfatório para o resultado dessa tarefa, outras possibilidades poderiam ser exploradas, por exemplo, n-grams ou formas combinadas, bem como outros tipos de técnicas de aprendizado de máquina, inclusive os não supervisionados. Além disso, o uso de um software comercial como KNIME também possui seus pontos de vantagens e desvantagens.

Vantagens, pelo fato de ser uma plataforma acessível, *open source* e de relativa simplicidade de uso, e existe a facilidade de reprodução desse estudo e estratégia de modelos para outras empresas, contribuindo para a prática. Como desvantagens, pela sua natureza modularizada, possui limitações sobre explorações mais profundas das técnicas, sendo eventualmente útil para uma etapa de prototipagem, com posterior desenvolvimento de códigos mais específicos para cada contexto, seja em linguagem Phyton ou alguma outra linguagem de programação.

Em suma tais considerações devem proporcionar melhor transparência e compreensão dos dados, e em tempo hábil. O que deve refletir no processo de decisão para categorização de produtos em atividade de cadastro apresentando ganhos como no atendimento de demandas operacionais e financeiras.

REFERÊNCIAS

BADRIYAH, T.; WIJAYANTO, E. T.; SYARIF, I.; KRISTALINA, P. A hybrid recommendation system for E-commerce based on product description and user profile. *In:* SEVENTH INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING TECHNOLOGY (INTECH). IEEE, 2017. https://doi.org/10.1109/INTECH.2017.8102435

BRANDAO, M. S.; GODINHO FILHO, M.; DA SILVA, A. L. Luxury supply chain management: a framework proposal based on a systematic literature review. **International Journal of Physical Distribution & Logistics Management**, 2021. https://doi.org/10.1108/IJPDLM-04-2020-0110

CALVO-VALVERDE, L. A.; MENA-ARIAS, J. A. Evaluation of different text representation techniques and distance metrics using KNN for documents classification. **Tecnología en marcha**, v. 33, n. 1, p. 64-79, 2020.

CHANDRA, B.; MAZUMDAR, S.; ARENA, V. C.; PARIMI, N. Elegant Decision Tree Algorithm for Classification in Data Mining. *In*: WISE WORKSHOPS, 2002.

CHERIYAN, S.; IBRAHIM, S.; MOHANAN, S.; TREESA, S. Intelligent Sales Prediction Using Machine Learning Techniques. *In*: INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRONICS & COMMUNICATIONS ENGINEERING (ICCECE). IEEE, 2018. https://doi.org/10.1109/iCCECOME.2018.8659115

LIMA, F. R. P. SILVA, A. L.; GODINHO FILHO, M.; DIAS, E. M. Systematic review: resilience enablers to combat counterfeit medicines. **Supply Chain Management**: An International Journal, 2018. https://doi.org/10.1108/SCM-04-2017-0155

FARIA, N. C. **Cadastro de Materiais -** Um Tesouro Ignorado pelas Empresas. Disponível em: https://www.guialog.com.br/Y542.htm. Acesso em: 04 abr. 2004.

GOMAA, W.; FAHMY, A. A. A survey of text similarity approaches. **International journal of Computer Applications**, v. 68, n. 13, p. 13-18, 2013. https://doi.org/10.5120/11638-7118

HARRAG, F.; EL-QAWASMEH, E.; PICHAPPAN, P. Improving Arabic text categorization using decision trees. *In*: INTERNATIONAL CONFERENCE ON NETWORKED DIGITAL TECHNOLOGIES, 1., IEEE, 2009.

HARRIS, J. G.; DAVENPORT, T. H. **Competing on analytics: The new science of winning**. Harvard Business Review, 2017.

HASAN, A.; MOIN, S.; KARIM, A.; SHAMSHIRBAND, S. Machine learning-based sentiment analysis for twitter accounts. **Mathematical and Computational Applications**, v. 23, n. 1, p. 11, 2018. https://doi.org/10.3390/mca23010011

HSSINA, B.; MERBOUHA, A.; EZZIKOURI, H.; ERRITALI, M. A comparative study of decision tree ID3 and C4. 5. **International Journal of Advanced Computer Science and Applications**, v. 4, n. 2, p. 13-19, 2014. https://doi.org/10.14569/SpecialIssue.2014.040203

HUANG, A. **Similarity measures for text document clustering.** *In*: PROCEEDINGS OF THE SIXTH NEW ZEALAND COMPUTER SCIENCE RESEARCH STUDENT CONFERENCE (NZCSRSC2008), Christchurch, New Zealand: [s.e], 2008.

IMAM. **Padrão Descritivo de Materiais – PDM.** Disponível em: https://www.imam.com.br/consultoria/artigo/pdf/padrao-descrtivo-de-materiais-pdm.pdf. Acesso em: 4 abr. 2021.

JIN, C.; DE-LIN, L.; FEN-XIANG, M. An improved ID3 decision tree algorithm. *In*: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE & EDUCATION, 4., IEEE, 2009.

JOHNSON, D. E.; OLES, F. J.; ZHANG, T.; GOETZ, T. A decision-tree-based symbolic rule induction system for text categorization. **IBM Systems Journal**, v. 41, n. 3, p. 428-437, 2002. https://doi.org/10.1147/sj.413.0428

JORDAN, M. I.; MITCHELL, T. M. **Machine learning:** Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. https://doi.org/10.1126/science.aaa8415

KIBRIYA, A. M.; FRANK, E.; PFAHRINGER, B.; HOLMES, G. Multinomial naive bayes for text categorization revisited. *In*: AUSTRALASIAN JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. Springer, Berlin, Heidelberg, 2004. https://doi.org/10.1007/978-3-540-30549-1_43

KIM, S. B.; HAN, K. S.; RIM, H. C.; MYAENG, S. H. Some effective techniques for *Naïve Bayes* text classification. **IEEE transactions on knowledge and data engineering**, v. 18, n. 11, p. 1457-1466, 2006. https://doi.org/10.1109/TKDE.2006.180

LAKSHMI, T. M.; MARTIN, A.; BEGUM, R. M.; VENKATESAN, V. P. An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative

Data. **International Journal of Modern Education & Computer Science**, v. 5, n. 5, 2013. https://doi.org/10.5815/ijmecs.2013.05.03

LEAL, R. S. **Métricas Comuns em Machine Learning:** como analisar a qualidade de chat bots inteligentes — métricas (3 de 4). Disponível em: https://medium.com/as-m%C3%A1quinas-que-pensam.

MARTINEZ-MARTIN, N. What are important ethical implications of using facial recognition technology in health care?. **AMA journal of ethics**, v. 21, n. 2, p. E180, 2019. https://doi.org/10.1001/amajethics.2019.180

MIAO, F.; ZHANG, P.; JIN, L.; WU, H. Chinese news text classification based on machine learning algorithm. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT HUMAN-MACHINE SYSTEMS AND CYBERNETICS (IHMSC), 10., 2018. https://doi.org/10.1109/IHMSC.2018.10117

MOORE, M. M.; SLONIMSKY, E.; LONG, A.D.; SZE, R. W.; IYER, R. S. Machine learning concepts, concerns and opportunities for a pediatric radiologist. **Pediatric radiology**, v. 49, n. 4, p. 509-516, 2019. https://doi.org/10.1007/s00247-018-4277-7

NASSIF, A. B.; SHAHIN, I.; ATTILLI, I.; AZZEH, M.; SHAALAN, K. Speech recognition using deep neural networks: A systematic review. **IEEE**, v. 7, p. 19143-19165, 2019. https://doi.org/10.1109/ACCESS.2019.2896880

PALMA NETO, L. G.; NICOLETTI, M. C. Introdução às redes neurais construtivas. São Carlos, SP: Editora da Universidade Federal de São Carlos, 2005.

PAVLYSHENKO, B. M. Machine-learning models for sales time series forecasting. **Data**, v. 4, n. 1, p. 15, 2019. https://doi.org/10.3390/data4010015

PENG, W.; CHEN, J.; ZHOU, H. An implementation of ID3-decision tree learning algorithm, v. 13, 2009. https://doi.org/10.1109/ICCSE.2009.5228509

QIANG, G. An effective algorithm for improving the performance of Naïve Bayes for text classification. *In*: SECOND INTERNATIONAL CONFERENCE ON COMPUTER RESEARCH AND DEVELOPMENT. IEEE, 2010. https://doi.org/10.1109/ICCRD.2010.160

RAD, S. E.; BEHJAT, A. R. Document Classification base on Ensemble Classifiers Support Vector Machine Multi-layer Perceptron and k-Nearest Neighbors. J. Biochem. **Tech**, v. 2, p. 174-182, 2019.

RUECKEL, V.; KOCH, A.; FELDMANN, K.; MEERKAMM, H. Process data management in the whole product creation process. *In*: PROCEEDINGS OF THE NINTH INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED

COOPERATIVE WORK IN DESIGN. IEEE, 2005. https://doi.org/10.1109/CSCWD.2005.194329

SABUNA, P. M.; SETYOHADI, D. B. Summarizing Indonesian text automatically by using sentence scoring and decision tree. *In*: INTERNATIONAL CONFERENCES ON INFORMATION TECHNOLOGY, INFORMATION SYSTEMS AND ELECTRICAL ENGINEERING (ICITISEE), 2., 2017. https://doi.org/10.1109/ICITISEE.2017.8285473

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, v. 34, n. 1, p. 1-47, 2002. https://doi.org/10.1145/505282.505283

SEBASTIANI, F. Text categorization. *In*: ENCYCLOPEDIA of Database Technologies and Applications. IGI Global, p. 683-687, 2005. https://doi.org/10.4018/978-1-59140-560-3.ch112

SHI, L.; WENG, M.; MA, X.; XI, L. Rough set based decision tree ensemble algorithm for text classification. **Journal of Computational Information Systems**, v. 6, n. 1, p. 89-95, 2010.

SINGH, S.; GUPTA, P. Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. **International Journal of Advanced Information Science and Technology (IJAIST)**, v. 27, n. 27, p. 97-103, 2014.

VARIAN, H. Artificial intelligence, economics, and industrial organization. *In*: NATIONAL Bureau of Economic Research, 2018. https://doi.org/10.3386/w24839

VIJAYMEENA, M. K.; KAVITHA, K. A survey on similarity measures in text mining. **Machine Learning and Applications**: an International Journal, v. 3, n. 2, p. 19-28, 2016. https://doi.org/10.5121/mlaij.2016.3103

WANG, Z.; DI, H.; SHAFIQ, M. A.; ALAUDAH, Y.; ALREGIB, G. Successful leveraging of image processing and machine learning in seismic structural interpretation: A review. **The Leading Edge**, v. 37, n. 6, p. 451-461, 2018. https://doi.org/10.1190/tle37060451.1

ZHANG, H.; LI, D. **Naïve Bayes text classifier.** *In*: INTERNATIONAL CONFERENCE ON GRANULAR COMPUTING (GRC 2007), 2007. https://doi.org/10.1109/GrC.2007.40

Artigo recebido em: 24/10/2021 e aceito para publicação em: 21/02/2022
DOI: http://doi.org/10.14488/1676-1901.v21i4.4483