

CLASSIFICAÇÃO AUTOMATIZADA DE PRODUTOS DA NOTA FISCAL ELETRÔNICA DE COMPRAS PÚBLICAS

Bruno Mattos Souza de Souza Melo
Mestre em Engenharia de Sistemas pela Universidade de São Paulo - USP
Tribunal de Contas do Estado do Rio de Janeiro - TCE-RJ

Wellington de Souza Amaral
Mestre em Ciência da Computação pelo Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET-RJ
Tribunal de Contas do Estado do Rio de Janeiro - TCE-RJ

Leonardo Silva de Lima
Doutor em Engenharia de Produção pela Universidade Federal do Rio de Janeiro - COPPE/UFRJ
Universidade Federal do Paraná

Eduardo Bezerra da Silva
Doutor em Engenharia de Sistemas de Computação pela Universidade Federal do Rio de Janeiro - COPPE/UFRJ
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET-RJ

Resumo: O problema de agrupamento e classificação de produtos quanto à sua natureza e similaridade, a partir das informações constantes das notas fiscais eletrônicas (NFE), é comum a todos os órgãos de controle do Brasil. Com o uso da base de dados das NFEs de produtos adquiridos pelo Estado do Rio de Janeiro, da base de dados de medicamentos autorizados pela Agência Nacional de Vigilância Sanitária e de técnicas de mineração de dados, desenvolveu-se uma metodologia para agrupar e classificar os bens e produtos farmacêuticos adquiridos por órgãos da administração pública fluminense. A contribuição imediata da metodologia desenvolvida é o aumento da capacidade analítica dos órgãos de controle de todo o Brasil na fiscalização das despesas relacionadas aos medicamentos adquiridos pela rede pública de saúde.

Palavras-chave: procedimentos analíticos aplicados ao controle externo; mineração de dados; agrupamento e classificação de despesas públicas; análise de grafos.

***Abstract**: The problem of grouping and classifying products as to their nature and similarity, based on the information contained in their electronic invoices (NFE), is common to all public control bodies in Brazil. Using the NFE database of products purchased by the State of Rio de Janeiro, the database of medicines authorized by ANVISA (the Brazilian counterpart of USA's FDA – Food and Drug Administration agency) and data mining techniques, a methodology was developed to group and classify products acquired by the public administration of the State Rio de Janeiro. The immediate contribution of the developed methodology is to increase the analytical capacity of the control bodies all over the country in the inspection of expenses related to drugs purchased by the public administration.*

***Keywords**: analytical procedures applied to external control; data mining; grouping and classification of public expenditures; graph analysis.*

1. INTRODUÇÃO

Nos dias atuais, a maioria das transações comerciais no Brasil é estabelecida por meio da Nota Fiscal Eletrônica (NFE). Para transações comerciais destinadas a órgãos públicos, a emissão da NFE tornou-se obrigatória a partir de 2010 [Brasil, 2020]. Tais dados podem ser alvo para diferentes tipos de análise ou até mesmo ser empregados como base para geração de modelos preditivos. Poder-se-ia obter o preço médio de venda de um determinado produto (ou serviço), informação esta que pode ser utilizada para identificar discrepâncias nas aquisições. Um obstáculo para essa tarefa reside na característica observada nas descrições de itens de NFE. Não é raro encontrar descrições distintas em NFE distintas que referenciam um mesmo produto comercial. Associar tais descrições ao produto a que se referem não é uma tarefa trivial. A NFE é o documento digital, emitido e armazenado eletronicamente, com o intuito de documentar, para fins fiscais, uma operação de circulação de mercadorias ou uma prestação de serviços, ocorrida entre as partes. A NFE contém informações suficientes para identificar o produto ou serviço comercializado. Para esse fim, destacam-se os atributos "Nomenclatura Comum do Mercosul / Sistema Harmonizado (NCM/SH)", a "descrição", "CLEAN" e "CLEANTributavel".
Embora haja, no próprio sistema de nota fiscal eletrônica vigente, a previsão de classificar e individualizar os bens e produtos por meio da NCM/SH, essa classificação ainda não é suficiente para individualizar os produtos a fim de permitir, por exemplo, pesquisa de preços de maneira precisa. Grande parte dos atributos necessários para tal individualização está contida no campo "descrição" da nota fiscal eletrônica no qual as empresas emissoras informam, em linguagem natural, detalhes dos itens, bens e produtos comercializados.
Considere o exemplo da Tabela 1 a seguir, confeccionada a partir da base de dados de NFEs disponíveis ao TCE-RJ. A descrição de alguns produtos de idêntico código de NCM/SH ("3004.90.45"), cuja referência é "Paracetamol; bromoprida", evidencia o uso de linguagem natural na qual se encontram dispersos vários atributos relacionados à composição, embalagem, forma de apresentação e dosagem dos medicamentos que deverão ser utilizados na individualização e agrupamento dos produtos.

Tabela 1: Excerto dos produtos cujo NCM/SH é igual a "3009045" na base de dados da NFe.

Descrição do Produto	NCM/SH
**PARACETAMOL + CODEINA 500MG + 30MG - GEOLAB - Lote: 1710024 / 31/10/2019	30049045
CODEINA+PARACETAMOL 30MG/500MG	30049045
FOSFATO DE CODEINA+PARACETAMOL 30MG	30049045
PARACET + CODEINA 500/30 MG C/96 (A2) PARACETAMOL + FOSFATO DE CODEINA (A2)	30049045
PARACETAMOL + CODEINA (FOSFATO) 500MG + 30MG	30049045
PARACETAMOL + F. CODEINA 30 MG	30049045
AMINOFILINA 24MG/ML AMPOLA 10ML TEUTO	30049045
BROMOPRIDA 10MG	30049045
BROMOPRIDA 10MG (ITEM GENERICO) L: 17D45M Q: 1.600,0000 V: 30/04/19	30049045

Fonte: elaboração dos autores a partir da base de dados de nota fiscal eletrônica.

Dois campos da NFE são de particular relevância para o trabalho aqui proposto. O campo CLEAN corresponde ao número Global Trade Item Number (GTIN) contido na embalagem com código de barras. O campo CLEANTRIBUTAVEL corresponde ao número de GTIN da menor unidade comercializada no varejo identificável por código GTIN. Ele deve ser preenchido com GTIN contido na embalagem com código de barras. No caso de produtos comercializados por lote, nesse campo, será preenchido com o GTIN da menor unidade comercializada no varejo. Apesar de se poder classificar de forma clara e inequívoca um produto pelo GTIN, nem todos os produtos possuem GTIN e o emissor da NFE não é obrigado a preencher os campos CLEAN e CLEANTRIBUTAVEL nas NFE. Verificou-se que o índice de preenchimento desses campos nas NFE é baixo. Devido ao potencial para classificar os produtos e a razão de sua existência, denominaremos os campos "NCM-SH", "CLEAN" e "CLEANTRIBUTAVEL" como "campos classificadores".
Nesse contexto, o objetivo desta pesquisa é investigar formas de associar as descrições de itens de NFE aos seus referidos produtos comerciais. Tal problema foi definido como uma tarefa de classificação em aprendizado de máquina, em que o produto comercial representa a classe à qual uma descrição de item pode pertencer. O conjunto de dados utilizado como base neste projeto é oriundo de registros de aquisições por órgãos públicos brasileiros. O resultado imediato deste processo de agrupamento e classificação é o aumento da capacidade analítica dos órgãos de controle. O problema de agrupamento e classificação de produtos quanto à sua natureza e similaridade, a partir das informações constantes das notas fiscais eletrônicas, é comum a todos os órgãos de controle do Brasil.
Já existem pesquisas anteriores com o mesmo propósito, como é o caso do trabalho de Carvalho et al. (2014), que utiliza técnicas de frequências de palavras na descrição de produtos, Marzagão (2015), que utiliza a técnica de máquinas de vetor de suporte (do inglês, support vector machine) e Gandini (2019), que utiliza processamento de linguagem natural, redução de dimensionalidade e agrupamento. Vale mencionar que todos esses trabalhos mencionados anteriormente não atingiram um ponto em que se pode identificar um produto satisfatoriamente de maneira que se permita montar um banco de preços.

2. DESENVOLVIMENTO

Neste trabalho foi utilizado o processo CRISP-DM para a mineração dos dados. CRISP-DM é a abreviação de CRoss Industry Standard Process for Data Mining (Shaerer, 2000). É o modelo reconhecido como o padrão mundial da indústria para os processos de mineração de dados. Ele descreve as abordagens comumente utilizadas por especialistas em mineração de dados para a solução de problemas, independentemente da área de negócio e das tecnologias aplicadas. Nas subseções que se seguem serão apresentadas as etapas de "análise exploratória dos dados", "preparação dos dados" e "modelagem" do processo CRISP-DM.

2.1. ETAPA DE ANÁLISE EXPLORATÓRIA DOS DADOS

O conjunto de dados analisado corresponde a registros de produtos de NFE emitidas em que incidem o Imposto sobre Circulação de Mercadorias e Serviços (ICMS) para órgãos da administração pública sob a jurisdição do Tribunal de Contas do Estado do Rio de Janeiro, o que inclui o estado do Rio de Janeiro e todos os municípios do Rio de Janeiro, com exceção da capital. Foram analisadas todas as NFEs com data de emissão no ano de 2018. Foi realizada uma análise exploratória sobre esse conjunto de dados com o objetivo de explorar e verificar sua qualidade. O conjunto de dados analisados possui ao todo 4.861.392 registros referentes a produtos. Cada registro guarda diversas informações a respeito do produto comprado, da empresa emissora da nota fiscal, do comprador, do meio de transporte utilizado para a entrega do produto e do local de emissão da nota fiscal.
No período analisado, foram emitidas NFEs para 1.338 órgãos públicos. Foram emitidas 1.076.719 NFEs distintas para esses órgãos. Foram adquiridos 4.745.029 produtos. O valor total de produtos adquiridos nessas NFEs é de R$8.346.025.418,80. Apesar do grande número de NFEs emitidas, a quantidade de empresas fornecedoras é de 6.100 empresas.
Realizou-se uma análise mais focada nos campos "NCM-SH", "CLEAN" e "CLEANTRIBUTAVEL". Esses campos apresentam um maior potencial para classificar e agrupar os produtos. Verificou-se que o percentual de preenchimento dos campos classificadores em cada NFE é bem diferente. Enquanto o "NCM-SH" é preenchido em 100% das NFEs, os campos "CLEAN" e "CLEANTRIBUTAVEL" só são preenchidos, respectivamente, em 24% e 23% das NFEs.
Para cada registro de produto de nota fiscal há uma classificação do NCM-SH correspondente. Nesta seção serão identificadas as classificações em maior número de registros e valores. A Tabela 2 a seguir apresenta um consolidado da quantidade de produtos e os valores por capítulo de NCM-SH. São apresentados os cinco capítulos com maior valor em sua ordem decrescente.

Tabela 2: Consolidação das dez classificações da Nomenclatura Comum do Mercosul (NCM) com maior valor associado.

Descrição Capítulo	Qtde. (%)	Qtde. ordem	Qtde. valor	Valor ordem
produtos farmacêuticos	**10,50%**	**3**	**21,70%**	**1**
preparações alimentícias diversas	3,90%	7	12,10%	2
veículos: automóveis, tratores, ciclos e outros veículos terrestres, suas partes e acessórios	2,20%	12	11,10%	3
instrumentos e aparelhos de óptica, de fotografia, de cinematografia, de medida, de controle ou de precisão; instrumentos e aparelhos médico-cirúrgicos; suas partes e acessórios	10,80%	2	8,10%1	4
combustíveis minerais, óleos minerais e produtos da sua destilação; matérias betuminosas; ceras minerais	13,30%	1	5,30%	5

Fonte: elaboração dos autores a partir da análise exploratória da base de NFE.

O gráfico de dispersão a seguir apresenta no eixo das abscissas o percentual de registros de produtos por classificação NCM-SH e no eixo das ordenadas o percentual do somatório dos produtos para capítulo correspondente. As classificações em maior quantidade foram indicadas com as respectivas descrições.

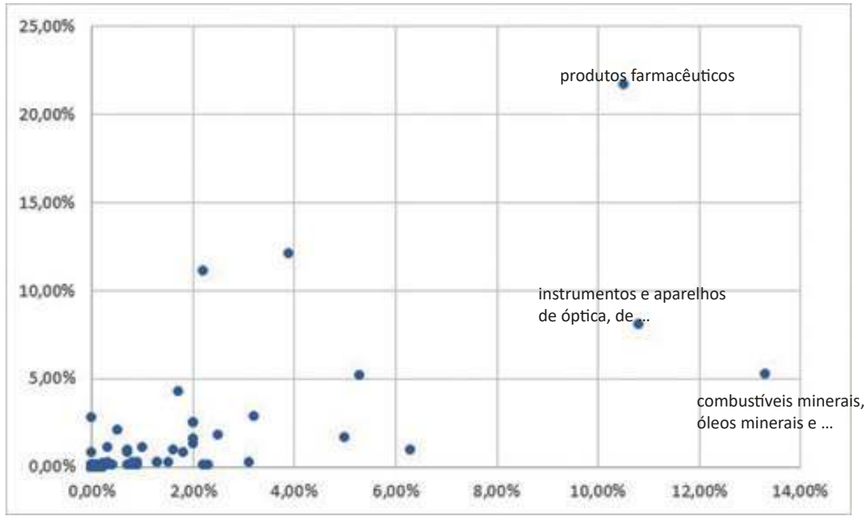


Figura 1: Gráfico de dispersão Relevância Monetária vs. Quantidade Percentual de Registros.

Pela análise da Tabela 2 e do gráfico da Figura 1, o capítulo "produtos farmacêuticos" da classificação NCM-SH corresponde a 22% em valores monetários e a 10% em quantidade de registros da base de dados. Por essa análise, a amostra deve recair sobre o capítulo "produtos farmacêuticos".
Verificou-se que a taxa de preenchimento dos campos "CLEAN" e "CLEANTRIBUTAVEL" varia em função da classificação NCM-SH. A Tabela 3 apresenta as cinco classificações com maior taxa de preenchimento do campo "CLEAN" e "CLEANTRIBUTAVEL"

Tabela 3: As cinco classificações com maior taxa de preenchimento do campo "CLEAN" e "CLEANTRIBUTAVEL".

Descrição capítulo NCM-SH	Percentual de preenchimento do campo CLEAN	Percentual de preenchimento do campo CLEANTRIBUTAVEL
tecidos especiais; tecidos tufados; rendas; tapeçarias; passamanarias; bordados	65,15%	65,07%
produtos farmacêuticos	54,11%	50,01%
armas e munições; suas partes e acessórios	53,36%	52,95%
fibras sintéticas ou artificiais, descontínuas	52,88%	50,05%
obras de couro; artigos de correeiro ou de seleiro; artigos de viagem, bolsas e artigos semelhantes; obras de tripa	51,30%	50,30%

Fonte: elaboração dos autores a partir da análise exploratória da base de NFE.

A taxa de preenchimento do CLEAN e CLEANTRIBUTAVEL de produtos de algumas classificações supera os 50% enquanto a média não é superior a 25%. Portanto, por essa análise, verifica-se que a taxa de preenchimento dos campos "CLEAN" e "CLEANTRIBUTAVEL" varia bastante em função da classificação do produto pela NCM SH.
Dessa maneira, seria possível utilizar os registros de produtos preenchidos com campo CLEAN e CLEANTRIBUTAVEL para validar os resultados do modelo classificador.

2.2. ETAPA DE PRÉ-PROCESSAMENTO

Em função da análise exploratória de dados descrita na etapa anterior, estabeleceu-se, com base em critérios objetivos (quantidade de registros e relevância monetária), o domínio dos "Produtos Farmacêuticos" como sendo o escopo das etapas subsequentes do modelo.
Para o balizamento da tarefa de classificação foi utilizada a base de dados de medicamentos da Agência Nacional de Vigilância Sanitária (ANVISA). A Figura 2 exibe cinco registros da base ANVISA com seus respectivos campos. Esta base reúne todos os medicamentos comercializados no Brasil e contém os seguintes campos de interesse: o campo "princípio ativo" é o princípio ativo do medicamento; o campo "apresentação" é a descrição textual de como o medicamento é apresentado, quanto à sua forma farmacêutica, dosagem e quantidade; o campo "produto" é o nome comercial dado ao medicamento, em determinada apresentação, de um referido laboratório; e o campo "código EAN" é também conhecido como Código de Barras ou GTIN.

	Principio Ativo	EAN	Produto	Apresentacao
0	CEFALOTINA SÓDICA	7896361881450	CEFALOTINA SÓDICA	1G PÓ P/ SOL INJ CT 50 FA VD INC (EMB HOSP)
1	CEFAZOLINA SÓDICA	7896361881405	CEFAZOLINA SÓDICA	1 G PÓ P/ SOL INJ CT 50 FA VD INC (EMB HOSP)
2	CEFOTAXIMA SÓDICA	7896361881412	CEFOTAXIMA SÓDICA	1 G PÓ P/ SOL INJ CT 50 FA VD INC (EMB HOSP)
3	CLORIDRATO DE CIPROFLOXACINO MONOIDRATADO	7896361881313	CLORIDRATO DE CIPROFLOXACINO	500 MG COM REV CT 2 BL AL PLAS INC X 07
4	CEFALEXINA	7896361880019	CEFALEXINA	500 MG COM REV CT BL AL PVC/PVDC INC X 8

Figura 2: Excerto da tabela de medicamentos da ANVISA.

A tarefa de rotulação dos dados da base de NFE consiste essencialmente no cruzamento de seus registros com a base de medicamentos da ANVISA.
A Figura 3 apresenta três registros da base de NFE. Os campos de especial interesse para a presente tarefa são: o campo "Descrição", que é um campo textual que, embora não seja opcional, é de livre preenchimento. Neste sentido, a descrição costuma reunir desde uma simples versão abreviada do "princípio ativo" ou do "produto" até uma composição completa desses campos acrescidos de sua forma de "apresentação"; e o campo "EAN", que corresponde ao "Código EAN" da base ANVISA. No entanto, seu preenchimento é opcional e em grande parte dos registros ele não é informado.

	Numero NF	Numero Item	Quantidade	Valor Unitario	Descricao	EAN
83785	705	7	4000	0.1500	PREDNISONA 5MG	N/I
168075	41560	1	10	0.5600	PROPILRACIL-100MG CX 30 COMP	N/I
301563	568368	11	6	4.2625	TRAMADOL G. NEO 50 MG 10 CA (A2)	7896714217321

Figura 3: Excerto da base de notas fiscais eletrônicas.

Conforme se observa na Figura 3, o campo "Descrição" pode apresentar muito ou pouco detalhe do produto, enquanto o campo EAN nem sempre está disponível por não ser de preenchimento obrigatório.
Para o subconjunto de produtos da base NFE que estão associados a um EAN, é possível identificar se dois itens de nota fiscal correspondem ao mesmo produto. Para isso, esses itens devem apresentar o mesmo valor para o campo EAN. Entretanto, nem todos os itens da base NFE estão associados a um EAN. Para esses produtos, o problema de identificar se dois itens quaisquer de notas fiscais diferentes correspondem ao mesmo produto não é trivial. Nesse caso, a informação que deve ser considerada para esta identificação é o campo "Descrição" do produto comprado constante na base NFE.
Verificou-se que dos 390.341 itens de nota fiscal da base NFE, 211.493 apresentam EAN, dentre os quais somente 12.029 EANs são distintos. Uma análise preliminar na base de NFE identificou que existem casos em que um único EAN está associado com diferentes preenchimentos do campo "Descrição". Com o intuito de verificar qual a frequência dessas ocorrências, produziu-se um histograma, representado pela Figura 4, como um sumário das quantidades de descrições encontradas para um mesmo produto no recorte da base de dados NFE relativo a medicamentos.

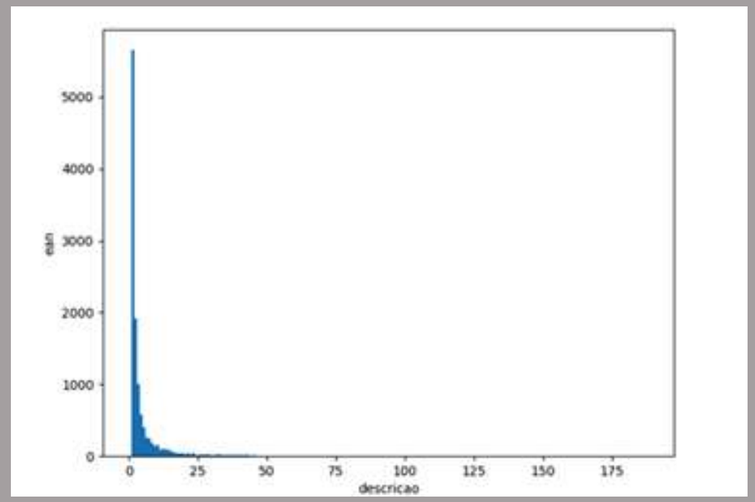


Figura 4: Base NFE: distribuição de EANs com diferentes strings no campo "Descrição".

A partir da Figura 4, pode-se observar que muitos EANs têm apenas uma descrição, o que implica que há uma pequena quantidade de EANs com várias descrições diferentes. Este fato caracteriza o desbalanceamento da distribuição.

2.3. ETAPA DE MODELAGEM

Nesta etapa, foram realizadas as tarefas de classificação e agrupamento pela implementação das rotinas computacionais em linguagem de programação Python. Os experimentos relacionados ao treinamento dos modelos de aprendizado de máquina foram realizados em um computador com a configuração Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz, 32GiB DIMM DDR4 2400MHz, NVIDIA GeForce GTX 1080. Foram utilizadas as bibliotecas do pacote python-Levenshtein, que possui funções para cálculo da Distância Levenshtein (Haapala, 2020); o pacote Imbalanced-learn (Lemaître et al., 2017), que oferece funções para a aplicação de diversas técnicas de balanceamento em conjuntos de dados; e o pacote fastText (Joulin et al., 2016), que é uma biblioteca que permite o treinamento de modelos de classificação em conjuntos de dados de texto.

2.3.1. Classificação por similaridade

Esta abordagem de classificação empregou a Distância Levenshtein para calcular o valor de similaridade entre uma descrição de item de NFE e uma descrição de medicamento da ANVISA. A ideia é que um valor de similaridade alto entre essas descrições é um indicativo de que correspondem ao mesmo produto. As etapas de preparação e construção do algoritmo para realizar essa tarefa são detalhadas a seguir.

Como primeiro passo, foi executado o pré-processamento do conjunto de dados NFE. Nesta etapa, expressões regulares foram utilizadas para remover informações irrelevantes, tais como lote e validade. Além disso, também foram aplicadas transformações dos caracteres para maiúsculo e remoção de duplicatas. Itens sem o atributo EAN registrado também foram removidos, uma vez que esse dado é essencial para a avaliação do classificador.

Foram empregadas expressões regulares para extrair das descrições de itens de NFE o que denotamos por termos principais. Consideramos termos principais aqueles definidos na descrição estruturada pela ANVISA: o princípio ativo ou o nome comercial do medicamento, a concentração, a farmacêutica e a quantidade. Os trechos de texto da descrição não extraídos nesta etapa foram descartados. Durante a etapa de extração dos termos principais, foram realizadas, paralelamente, a transformação, a ordenação e a concatenação destes. Esse processamento teve como objetivo derivar uma nova descrição aos moldes do normatizado pela ANVISA. Por exemplo, um trecho da descrição de item de NFE contendo o texto "C/30 CPR" foi detectado, por meio das expressões regulares, como as informações sobre forma farmacêutica e quantidade. No algoritmo implementado neste projeto, o trecho foi então transformado para os termos "COM" e "30". Alguns exemplos de resultado desta etapa são apresentados na Tabela 4.

Tabela 4: Extração dos termos principais

Descrição	Princípio ativo	Concentração	Forma	Quantidade
ciprofloxacino 200mg cloridrato s.fecha	ciprofloxacino	200mg		
carbolitium cr 450mg c/30 cpr	carbolitium cr	450mg	comprimido	30
mupirocina 20mg/g 15g generico prati, donaduzzi	mupirocina	20mg/g		15g

Fonte: elaboração dos autores a partir da análise exploratória da base de NFE.

Uma vez obtidas as descrições derivadas do conjunto de dados de NFE, o próximo passo foi calcular o valor de similaridade entre estas descrições e cada uma das descrições ANVISA. O produto associado à descrição ANVISA com maior valor de similaridade é então considerado como o produto correto. Para avaliar a taxa de acerto desta abordagem, bastou comparar os EANs e calcular o percentual de acerto.

2.3.2. Classificação por aprendizado de máquina

Nesta abordagem os conjuntos de dados de NFE e Anvisa foram combinados pela união de suas descrições e, ao conjunto combinado, aplicadas técnicas de aumento e balanceamento de dados. O dado produto oriundo do conjunto ANVISA não foi utilizado e as descrições de medicamentos derivadas deste conjunto foram formadas pela concatenação dos campos principio_ativo e apresentacao. Uma consequência desta decisão foi a necessidade de agrupar descrições iguais com EANs distintos (devido à remoção do campo produto). Cada agrupamento foi mapeado para uma chave única a qual substituiu os EANs do conjunto de dados resultante. A biblioteca fastText foi empregada para treinar um classificador utilizando as descrições de medicamentos como exemplos de entrada e as chaves únicas como classe. Estes passos serão mais bem detalhados a seguir.

Iniciando o pré-processamento dos dados, foram aplicadas técnicas de aumento de dados sobre os conjuntos NFE e Anvisa. A necessidade dessa etapa justifica-se pelo histograma apresentado na Figura 4. No histograma, observa-se que muitos EANs possuem apenas uma descrição distinta associada, o que seria ineficiente para o aprendizado de um classificador baseado em aprendizado de máquina.

A primeira técnica empregada foi a coleta de descrições na web. Um algoritmo de coleta foi implementado para obter novas descrições retornadas pelo buscador Google. Em resumo, o algoritmo realiza uma busca no Google dando como entrada uma descrição de medicamento. Os dez títulos de resultado retornados são armazenados em uma lista e são denotados por candidatos. A lista de candidatos é pré-processada, removendo informações irrelevantes (e.g. preço ou nome de fornecedor). Por meio das regras de expressão regular empregadas na abordagem anterior, os termos principais são extraídos tanto da descrição de entrada quanto do candidato. O candidato que possui interseção entre o seu conjunto de termos extraídos e o conjunto de termos extraídos da descrição de entrada é eleito para compor o conjunto de dados aumentado.

A segunda técnica empregada foi a derivação de novas descrições por meio de transformações. Expressões regulares foram empregadas para substituição de termos, inclusão ou remoção de espaços, permutação de palavras, dentre outros. A Tabela 5 mostra três exemplos de transformações aplicadas no termo principal concentração. A coluna da esquerda apresenta uma descrição original de medicamento e a coluna da direita as derivações resultantes. Na primeira derivação foi removido o caractere "/". Na segunda, o espaço entre o dígito e a unidade de medida foi removido. Na terceira derivação, as medidas foram convertidas para unidades equivalentes.

Tabela 5: Derivação por transformação.

Descrição original	Descrição derivada
Seloken 1 mg / ml sol inj x 5 ml	Seloken 1 mg ml sol inj x 5 ml
Seloken 1 mg / ml sol inj x 5 ml	Seloken 1mg ml sol inj x 5 ml
Seloken 1 mg / ml sol inj x 5 ml	Seloken 1G / 1000ml sol inj x 5 ml

Fonte: elaboração dos autores a partir da derivação de novas descrições por meio de transformações de descrições de itens da NFE



Como última etapa de pré-processamento, foi realizado o balanceamento do conjunto de dados. A biblioteca Imbalanced-learn foi empregada nesta tarefa. O conjunto de dados aumentado foi dado como entrada para a biblioteca, que processou o oversampling dos dados. O conjunto de dados resultante alcançou pouco mais de 51 milhões de registros. A biblioteca fastText foi empregada para treinar um classificador, utilizando-se das descrições de medicamentos como exemplos de entrada e as chaves únicas como classes. Devido ao grande número de exemplos não oriundos do conjunto de dados NFE, houve a preocupação de separar um conjunto de teste contendo apenas exemplos deste último. Todas as descrições constantes do conjunto de teste foram removidas do conjunto de treino. O conjunto de teste resultante recebeu 6.922 registros com 2.909 classes distintas.

3. APRESENTAÇÃO DOS RESULTADOS

O modelo de classificação obtido por meio da abordagem de classificação por similaridade obteve uma acurácia de apenas 16% de acertos na base de dados de teste. Este desempenho é muito ruim, razão pela qual este método foi descartado sem que avaliações de desempenho mais robustas ou técnicas de otimização de parâmetros (“ajustes finos”) fossem a ele aplicados. Os motivos causadores de sua precariedade serão abordados na seção “Considerações Finais”.

O modelo obtido utilizando-se a abordagem de classificação por métodos de aprendizado de máquina, treinado pela biblioteca fastText, alcançou uma acurácia superior, de tal modo justificou-se a utilização de uma métrica de avaliação de desempenho mais robusta: o F1 Score. Nesta métrica, que combina os conceitos de precisão[1] e revocação[2], o modelo obteve uma pontuação média, considerando todas as classes de medicamentos, de 0,936, em uma escala compreendida entre 0 e 1, na qual o valor máximo indicaria um modelo perfeito.

A Figura 5 apresenta uma matriz de confusão em forma de mapa de calor para as 20 classes com maior número de descrições de medicamentos distintas associadas. Realizando a análise por classes, somente 573 classes (de um total de 2.909) obtiveram F1 Score menor que 1.

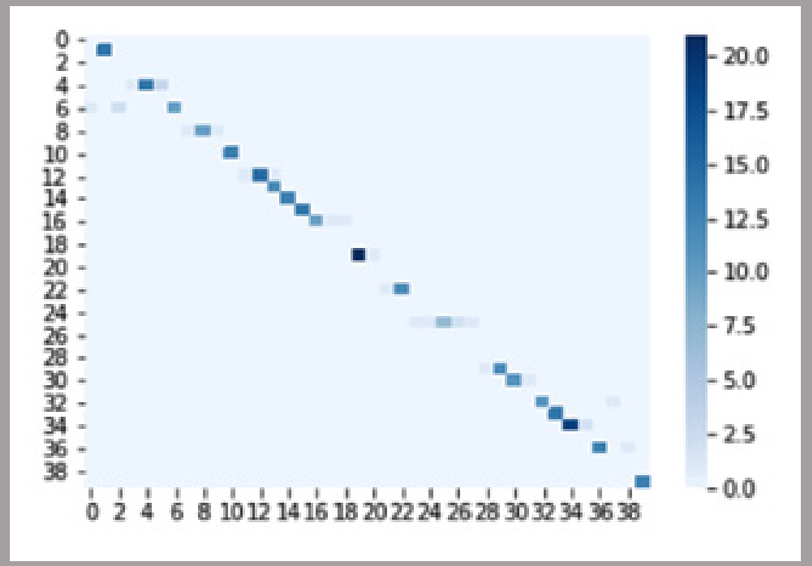


Figura 5: Matriz de Confusão (em modo de mapa de calor) das 20 maiores classes em termos de quantidade de descrições de medicamentos distintas associadas.

A Figura 6 apresenta um histograma para essas últimas no qual o eixo horizontal representa o valor F1 Score e o eixo vertical representa a quantidade de classes.

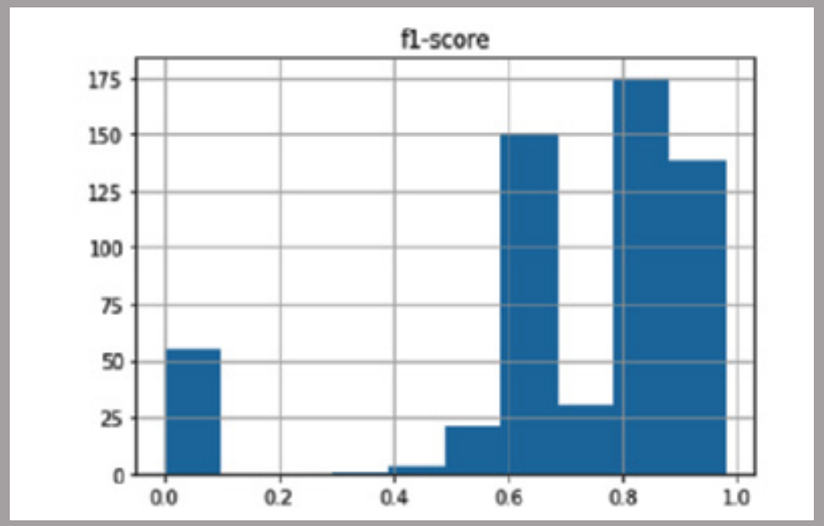


Figura 6: Histograma das 573 classes com F1 Score abaixo de 1, onde o eixo x representa o valor do F1 Score e o eixo y representa a quantidade de classes.

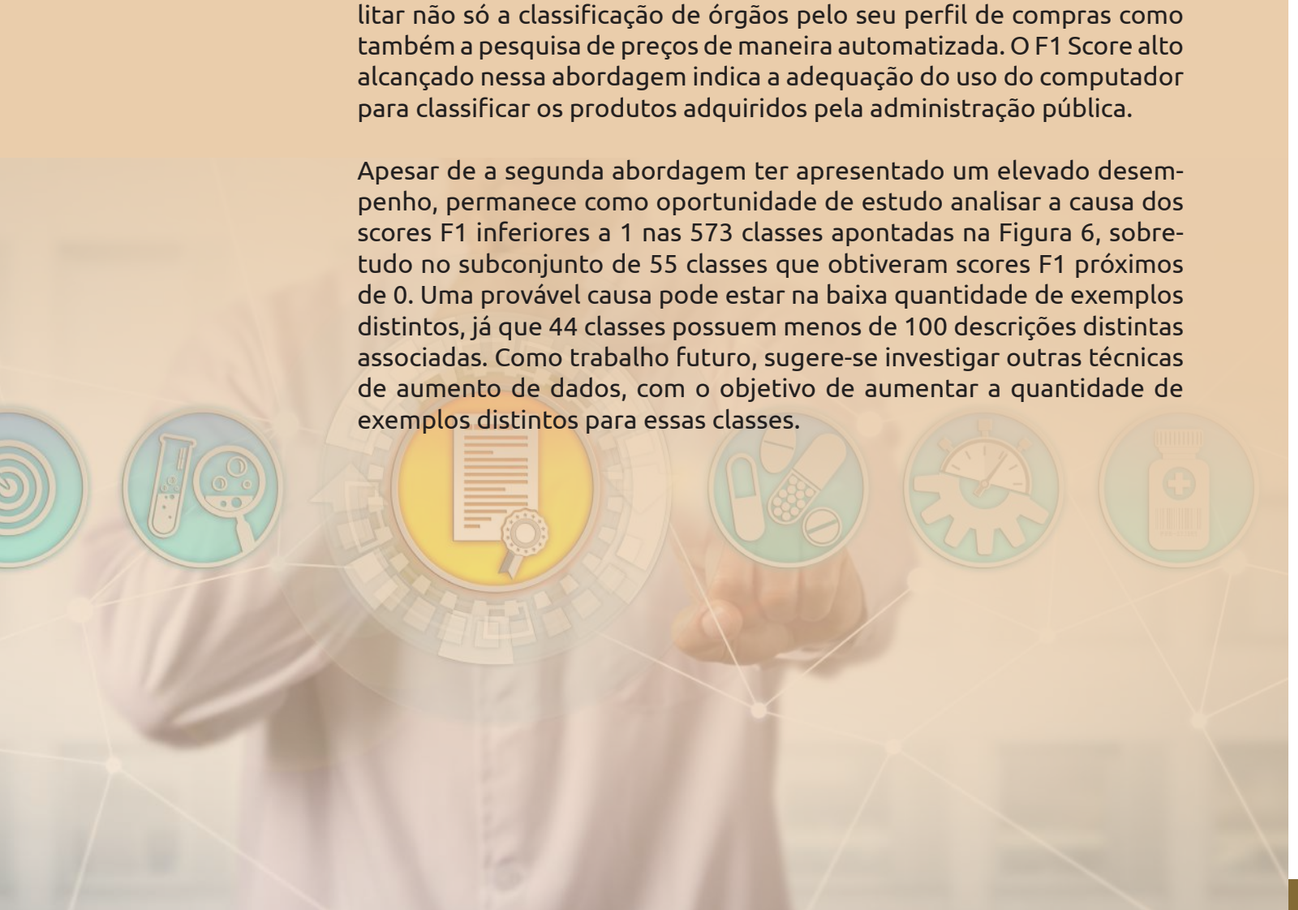
1 Precisão: é o percentual elementos, dentre aquele rotulados para uma classe, que realmente pertencem a ela.
2 Revocação: é o percentual de elementos pertencentes uma classe que foram recuperados ou corretamente identificados pelo modelo.

4. CONSIDERAÇÕES FINAIS

A baixa acurácia da primeira abordagem (classificação por similaridade) tem causa provável na ausência de informações mais detalhadas nas descrições de itens de NFE. Observando os resultados dos experimentos, foi possível identificar vários casos nos quais todos os termos principais extraídos do item de NFE estavam corretos, se comparados aos da descrição predita pelo algoritmo, mas que foram atribuídos à classe incorreta. No conjunto de dados ANVISA é comum existirem vários produtos com o mesmo princípio ativo, concentração e quantidade, mas que possuem identificadores EAN distintos devido ao nível de detalhamento das apresentações dos produtos. Uma vez que o algoritmo de Levenshtein (1966) aponta como descrição mais similar aquela que necessita de menor quantidade de edições, então, no caso anteriormente descrito, produtos com descrições de apresentação menores acabam sendo apontados como o produto mais similar, o que não necessariamente é verdade.

A segunda abordagem (classificação por aprendizado de máquina) alcança o objetivo principal da presente pesquisa: desenvolver um modelo classificador de produtos das compras públicas de maneira a possibilitar não só a classificação de órgãos pelo seu perfil de compras como também a pesquisa de preços de maneira automatizada. O F1 Score alto alcançado nessa abordagem indica a adequação do uso do computador para classificar os produtos adquiridos pela administração pública.

Apesar de a segunda abordagem ter apresentado um elevado desempenho, permanece como oportunidade de estudo analisar a causa dos scores F1 inferiores a 1 nas 573 classes apontadas na Figura 6, sobretudo no subconjunto de 55 classes que obtiveram scores F1 próximos de 0. Uma provável causa pode estar na baixa quantidade de exemplos distintos, já que 44 classes possuem menos de 100 descrições distintas associadas. Como trabalho futuro, sugere-se investigar outras técnicas de aumento de dados, com o objetivo de aumentar a quantidade de exemplos distintos para essas classes.



REFERÊNCIAS

BRASIL. Controladora Geral da União. **Manual da Lei de Acesso à Informação para Estados e Municípios**. 1. ed. Brasília: CGU, Secretaria de Prevenção da Corrupção e Informações Estratégicas, abr. 2013. Disponível em: https://www.gov.br/cgu/pt-br/centrais-de-conteudo/publicacoes/transparencia-publica/brasil-transparente/arquivos/manual_lai_estadosmunicipios.pdf. Acesso em: 13 nov. 2021.

BRASIL. (2020). Ministério da Fazenda. Conselho Nacional de Política Fazendária. **Protocolo ICMS 42, de 3 de julho de 2009**. Estabelece a obrigatoriedade da utilização da Nota Fiscal Eletrônica (NF-e) em substituição à Nota Fiscal, modelo 1 ou 1-A, pelo critério de CNAE e operações com os destinatários que especifica. Disponível em: https://www.confaz.fazenda.gov.br/legislacao/protocolos/2009/pt042_09. Acesso em: 12 nov. 2021.

CARVALHO, R.; PAIVA, E. D.; ROCHA, H. A. da; MENDES, G. L. Using Clustering and Text Mining to Create a Reference Price Database. **Learning and NonLinear Models**, v. 12, n. 1, p. 38–52, jan. 2014.

GANDINI, A. Identificando sobrepreço em compras públicas. In: SEMINÁRIO INTERNACIONAL SOBRE ANÁLISE DE DADOS NA ADMINISTRAÇÃO PÚBLICA, 5., 2019. **Palestra [...]**. Brasília: Tribunal de Contas da União, Instituto Serzedello Corrêa, 2019. Brasília. Disponível em: https://youtu.be/ISqva3yZrgw?t=5273. Acesso em: 4 mar. 2020.

HAAPALA, A. Python scripts to compute the Levenshtein distance. Available at: https://github.com/ztane/python-Levenshtein Accessed in: April, 2021.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; MIKOLOV, T. Bag of Tricks for Efficient Text Classification. **arXiv preprint arXiv:1607.01759**, 9 Aug. 2016. Disponível em: https://arxiv.org/pdf/1607.01759.pdf. Acesso em: 13 nov. 2021.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research 18 p.1-5, 2017.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics--Doklady**, v. 10, n. 8, p. 707-710, Feb. 1966. Disponível em: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf. Acesso em: 16 nov. 2021.

MARZAGÃO, T. Using SVM to pre-classify government purchases. **arXiv preprint arXiv:1601.02680**, 7 Dec. 2015. Disponível em: https://arxiv.org/pdf/1601.02680.pdf. Acesso em: 16 nov. 2021.

SHAERER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000.