UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO
MESTRADO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

João Batista Gonçalves de Brito

SELEÇÃO DE VARIÁVEIS BASEADA NA INTEGRAÇÃO DE RANKING DE IMPORTÂNCIA SVD COM MÉTODOS DE APRENDIZAGEM DE MÁQUINA

Porto Alegre

2020

João Batista Gonçalves de Brito

Seleção de variáveis baseada na integração de ranking de importância SVD com métodos de aprendizagem de máquina

Dissertação submetida ao Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Profissional, na área de concentração em Sistemas de Produção.

Orientador: Prof. Michel José Anzanello, *Ph D*.

Porto Alegre

2020

João Batista Gonçalves de Brito

Seleção de variáveis baseada na integração de ranking de importância SVD com métodos de aprendizagem de máquina

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Profissional e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel José Anzanello, *Ph.D.*

Orientador PMPEP/UFRGS

_____

Profa. Christine Tessele Nodari, *Dra.*

Coordenador PMPEP/UFRGS

Banca Examinadora:

Prof. Alessandro Kahmann, *Dr.* (IMEF/FURG)

Prof. João Luiz Becker, *Ph.D.* (EAESP/FGV)

Prof. Paul Gerhard Kinas, *Ph.D.* (IMEF/FURG)

" Eu fiz aqui apenas um
buquê de flores selecionadas, e nada
trago que seja meu, exceto o laço que as reúne."
- Michel de Montaigne, Essais (1580).

AGRADECIMENTOS

À minha esposa, Thábia, pelo amor, companheirismo, encorajamento e dedicação.

Aos meus pais Luiz (*in memorian*) e Leni pelo amor incondicional e, sobretudo, pelo investimento na minha educação, a maior e mais valiosa herança que poderiam deixar.

À minha irmã Josiane pelo carinho e incentivo.

Ao colega Guilherme pela assistência no desenvolvimento da proposta e à colega Juliana, pelo apoio na implementação deste projeto.

Ao meu orientador Professor Michel José Anzanello, *Ph.D.*, pela oportunidade, orientação e confiança.

Aos professores, técnicos administrativos e demais envolvidos que compõem a estrutura do Programa de Pós-Graduação em Engenharia de Produção na UFRGS pelo ensino de referência e organização.

RESUMO

Métodos para seleção de variáveis são importantes para tornar modelos de aprendizagem de máquina parcimoniosos e mais acurados, eliminando variáveis não-relevantes, ruidosas e altamente correlacionadas. Ademais, esses métodos podem contribuir com redução de custo e aumento da eficácia em atividades que incluem aferições de qualidade em processos industriais e comprovação da autenticidade de amostras de produtos. O presente trabalho propõe duas novas abordagens de seleção de variáveis, sendo cada uma disposta em um artigo. Em relação ao método, um novo ranking de importância de variáveis, baseado na decomposição de valores singulares, é proposto e utilizado para orientar um processo iterativo que compõe subconjuntos e os submente à uma técnica de aprendizagem de máquina. Na sequência, a acurácia do modelo é avaliada; o processo retém as variáveis que promovem ganho de acurácia e descarta as demais. Em termos dos artigos que compõem essa dissertação, no primeiro é aplicado o método de aprendizagem de máquina *k-Nearest Neighbor,* e os experimentos são direcionados à análise forense de identificação de medicamentos falsos. O segundo artigo utiliza o método de aprendizagem de máquina Ensemble Logistic GMDH-NN e executa experimentos sobre dados de processos industriais e propriedades físico-químicas de Biodiesel e Diesel brasileiro. As duas abordagens propostas demonstram desempenho superior em termos de aumento de acurácia e redução do subconjunto de variáveis quando comparadas a métodos reportados pela literatura.

Palavras-chave: Seleção de Variáveis. Ranking de Importância das Variáveis. Decomposição de Valores Singulares. Group Method Data Handling.

ABSTRACT

Methods for feature selection are important to make machine learning models parsimonious and accurate, eliminating non-relevant, noisy and highly correlated features. Moreover, these methods can contribute to cost reduction and increased efficiency in activities that include quality assessments in industrial processes and proving the authenticity of product samples. This paper proposes two new approaches to feature selection, each of which is arranged in an article. Regarding the method, a new ranking of the importance of variables, based on the singular value decomposition, is proposed and used to guide an iterative process that composes subsets and underlies them to a machine learning technique. In the sequence, the accuracy of the model is evaluated; the process retains the variables that promote accuracy gain and discards the others. In terms of the articles that compose this dissertation, in the first one the k-Nearest Neighbor machine learning method is applied, and the experiments are directed to the forensic analysis of falsified drug identification. The second article uses the Ensemble Logistic GMDH-NN machine learning method and performs experiments on industrial process data and physical-chemical properties of Brazilian Biodiesel and Diesel. The two proposed approaches demonstrate superior performance in terms of to improve accuracy and reduction of the subset of variables when compared to methods reported in the literature.

Keywords: Feature Selection. Ranking of importance of features. Singular Value Decomposition. Group Method Data Handling.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações Iniciais

O processo de seleção de variáveis, também conhecido como *feature selection*, é uma importante etapa de pré-processamento para criação de modelos preditivos, pois buscar as melhores variáveis verificando todos os possíveis subconjuntos pode ser uma tarefa inviável, principalmente em bases de dados com alta dimensionalidade (HUANG *et al.*, 2011). Nesse contexto, estudos desenvolvidos nas mais diversas áreas demonstram que um método de seleção de variáveis robusto contribui diretamente no desempenho dos modelos de aprendizagem (LIU; YU, 2005; DASH; LIU, 1997; ASYALI *et al.*, 2006). Para Kumar e Minz (2014), duas abordagens são relatadas no processo de seleção de variáveis.

A primeira é definida como *feature ranking* (FR) ou *individual evaluation* e trata da atribuição de pesos individuais conforme o grau de relevância da variável. A segunda, denominada como *subset evaluation* (SE), refere-se ao processo de avaliar o desempenho dos subconjuntos de variáveis. Diversos estudos propõem o uso da FR como mecanismo para orientar a SE, como descrevem Guyon e Elisseeff (2003). No trabalho de Anzanello *et al.* (2013), o ranking contribui para direcionar o descarte de variáveis (da menos importante para a mais importante), apoiando-se em mecanismos de SE. A criação de um ranking para quantificação da relevância de variáveis pode ser supervisionada ou não supervisionada.

Rankings supervisionados atribuem escores às variáveis considerando a relevância para predição de uma variável dependente (WANG *et al.*, 2017). Em contrapartida, rankings não supervisionados buscam variáveis com alto potencial de discriminação das observações, independentemente da existência de uma variável dependente. Nesse último caso, sua aplicação é tipicamente indicada para definição de clusters de observações em análises exploratórias (HE *et al.*, 2005; BRERETON, 2009; ZHAO; LIU, 2011; BOCKLITZ, 2019).

A presente de dissertação implementa um novo método de FR utilizando como base a decomposição valores singulares (SVD), e o integra com duas

diferentes propostas de SE. Na primeira, apresentada no artigo I (capítulo 2), foi utilizado o método de aprendizagem de máquina *k-Nearest Neighbor* (KNN) (FIX; HODGES, 1951), o qual foi aplicado sobre amostras de dados espectrais de medicamentos. Na segunda proposta, descrita pelo artigo II (capítulo 3), foi utilizado o algoritmo Ensemble Logistic GMDH-NN (EL-GMDH-NN) com dados provenientes de processos industriais e propriedades físico-químicas de Biodiesel e Diesel.

Ensemble Logistic GMDH-NN utiliza a estratégia *stacking ensemble* para combinar diversos algoritmos de aprendizagem de máquina, descritos nessa ótica como *learners*, em um algoritmo de aprendizagem de máquina, definido como *meta learning* (ALFARO *et al.*, 2019). O *meta learner* utilizado foi a rede neural Logistic GMDH-NN - otimizada pela heurística *Group Method of Data* Handling (GMDH) (KONDO et al., 1999). O GMDH promove a auto-organização da arquitetura da rede neural, definindo automaticamente a quantidade de neurônios, camadas e variáveis de entrada (IVAKHNENKO; G., 1968; KONDO, 1998, ARDAKANI; KORDNAEIJ, 2019).

A figura 1.1 apresenta a estrutura geral utilizada nos dois artigos - capítulo 2 e 3, respectivamente - no qual três etapas fundamentais são relatadas: (I) divisão da base de dados em porções de treino e teste; (II) cálculo do ranking de importância das variáveis via SVD, que representa uma inovação em termos de FR; e (III) seleção de variáveis, na qual o índice de importância proposto é utilizado para orientar um processo iterativo de composição dos subconjuntos de variáveis, utilizando inclusão *forward.* No que segue, cada subconjunto de variáveis compõe um modelo de aprendizagem de máquina. Quando a acurácia aumenta, a última variável inserida é mantida no subconjunto de variáveis eleitas. Caso contrário, ela é descartada. O processo segue até que a última variável do ranking de importância tenha sido verificada ou a acurácia tenha atingido 1. Os modelos gerados têm sua acurácia avaliada nos dados inseridos na porção de teste. A proposta de cada artigo diferencia-se, essencialmente, pelo método de aprendizagem de máquina (destacado em amarelo na figura 1.1) e pelos diferentes tipos de bases de dados.

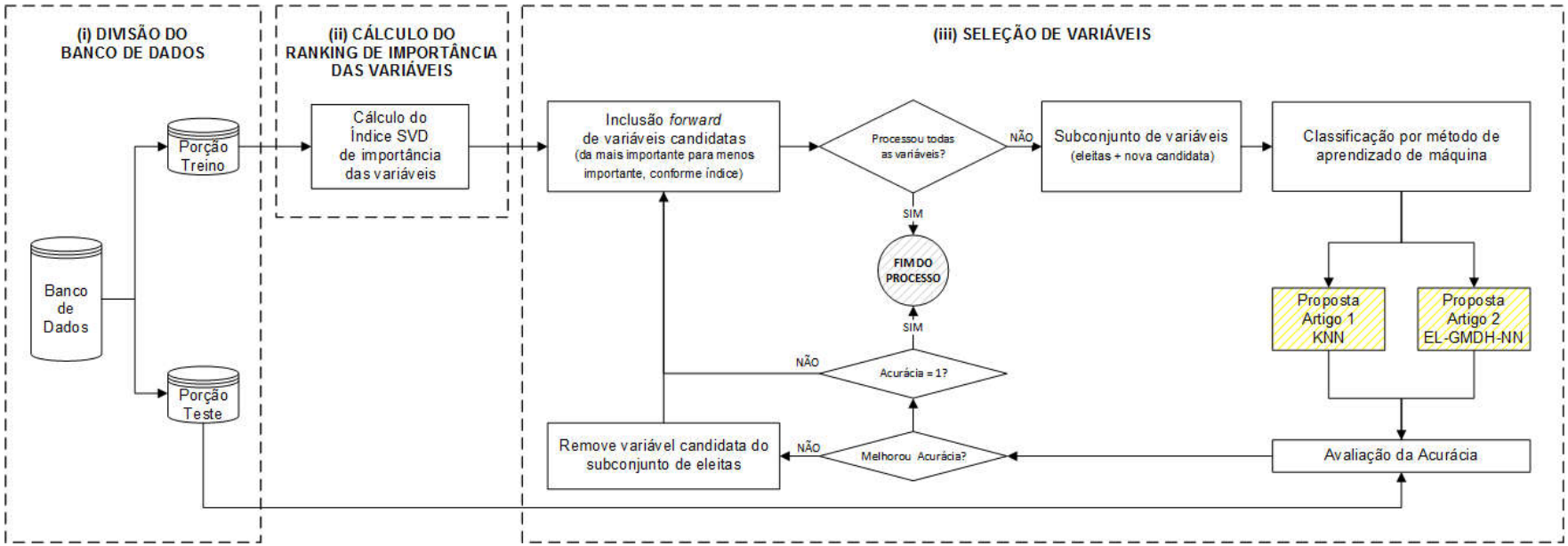

Figura 1.1 - Estrutura geral utilizada nos dois artigos

1.2 Objetivos

O objetivo geral da dissertação é propor uma nova abordagem para seleção de variáveis mais relevantes com vistas à classificação de amostras de medicamentos e produtos industriais.

Como objetivos específicos, listam-se:

- Propor um novo ranking de importância de variáveis baseado na decomposição matricial de valores singulares;
- Integrar o ranking proposto com métodos de aprendizagem de máquina para selecionar o subconjunto ótimo ou subótimo de variáveis com maior poder de discriminação de uma classe binária;
- Avaliar o desempenho da técnica *stacking ensemble* integrando múltiplos métodos de aprendizagem de máquina;
- Avaliar o desempenho de algoritmos preditivos otimizados pela heurística *Group Method of Data Handling* no processo de seleção do subconjunto de variáveis;
- Aplicar a proposta sobre dados espectrais de medicamentos para identificar se uma amostra é autêntica ou falsificada;
- Aplicar a proposta sobre dados de produtos industriais com vistas a enquadrá-los em classes de qualidade.

1.3 Justificativa do Tema

À medida que a tecnologia dos computadores e bancos de dados avança rapidamente, a humanidade confia cada vez mais na computação para acumular, processar e fazer uso dos dados. De tal forma, ferramentas de inteligência como aprendizagem de máquina, descoberta de conhecimento e mineração de dados têm assumido elevada relevância nos dias atuais (KANTARDZIC, 2020). Pesquisadores e profissionais percebem que, para usar essas ferramentas efetivamente, uma das etapas mais importante é o pré-processamento, no qual os dados são processados antes de serem

apresentados a qualquer algoritmo de aprendizagem, descoberta ou visualização em aplicações do mundo real (JONES, 2020).

Métodos para reconhecimento de padrões precisam lidar com amostras que podem conter um grande número de variáveis candidatas à predição (VENKATESH; ANURADHA, 2019). Por isso, a redução da dimensionalidade se torna crucial para viabilizar a criação de modelos preditivos (RAO *et al.*, 2019). Nesse contexto, o processo de seleção de variáveis remove as variáveis irrelevantes e ruidosas para definir o subconjunto ótimo ou subótimo com potencial de melhorar o desempenho de classificação (GUYON; ELISSEEFF, 2003). Para Chandrashekar e Sahin (2014), o processo de seleção de variáveis também contribui para entender os dados, uma vez que, ao identificar as variáveis mais relevantes, é possível extrair interpretações sobre o contexto da informação. Como resultado, há uma grande variedade de áreas no qual o processo é aplicado.

Amostras de dados resultantes do espectro gerado por Reflexão Total Atenuada no Infravermelho com Transformada de Fourier (ATR-FTIR) são frequentemente aplicadas na identificação de falsificações de produtos dos mais diversos setores, como alimentícia, farmacêutica e química, dentre outros (CUSTERS *et al.*, 2015). No processo, centenas ou milhares de comprimentos de onda são identificados e mensurados, produzindo, assim, um banco de dados com alta dimensionalidade (YANG *et al.*, 2019). Grande parte dos comprimentos de onda não contribui como preditor para classificação entre autêntico ou falsificado, sendo um desafio identificar o melhor subconjunto de preditores (ANZANELLO *et al.*, 2015). Ademais, na ambiência industrial, a identificação de aferições que produzem variáveis irrelevantes tem potencial de reduzir custos financeiros. Nessa situação, por exemplo, a seleção de variáveis pode fornecer subsídios para diagnóstico de um processo que faz avaliação de qualidade ou averiguação de autenticidade de amostras e produtos. Desta forma, percebe-se que o tema abordado nesta dissertação encontra justificativas em âmbito prático.

1.4 Procedimentos Metodológicos

O método de pesquisa utilizado pode ser caracterizado como de natureza aplicada, tendo em vista que o conteúdo teórico é explorado na solução de problemas genéricos (CRESWELL, 2010). A dissertação apresenta abordagem quantitativa, pois utiliza métodos de decomposição matricial e aprendizagem de máquina para solução dos problemas apresentados. O conteúdo é desdobrado em dois artigos.

O primeiro artigo propõe um ranking de importância baseado na decomposição matricial *Decomposição de Valores Singulares*. O ranking é integrado ao método de aprendizagem de máquina *k-Nearest Neighbor* para identificar o subconjunto de comprimentos de onda mais relevantes na classificação entre autêntico e falsificado, em amostras espectrais de Cialis® e Viagra®. A proposta toma como base o estudo de Anzanello *et al.* (2013), no qual foram utilizadas amostras de Cialis® e Viagra® autênticas..

No segundo artigo foram utilizados dados amostrais de processos industriais, conforme especificações de Gauchi e Chagnon (2001), e propriedades bioquímicas de Biodiesel e Diesel brasileiro fornecidas por Ferrão *et al.* (2011). Nos dois conjuntos de dados há uma classe que identifica se a amostra se caracteriza como conforme ou não conforme. O processo inicia computando o índice de importância das variáveis, da mesma forma como foi aplicado no primeiro artigo. Em seguida, o índice é integrado com o método *Ensemble Logistic Group Method of Data Handling type Neural Networks*. Essa técnica utiliza a estratégia *stacking ensemble* para combinar múltiplos algoritmos de aprendizagem, definidos como *learners* (*Suport Vector Machine*, *Random Forest*, *Naïve Bayes*, *Regularization Paths for Generalized Linear Models Coordinate Descent* e *Single-hidden-layer neural network*) no meta learning *Logistic Group Method of Data Handling type Neural Networks*. Entende-se que a combinação dos classificadores pode conduzir a resultados mais precisos, dado o caráter compensatório das vantagens e desvantagens de cada método.

1.5 Estrutura da Dissertação

A dissertação está organizada em quatro capítulos. O primeiro trata sobre as considerações iniciais, os objetivos gerais e específicos, a justificativa do tema, os procedimentos metodológicos, a estrutura da dissertação e as delimitações do estudo.

O segundo capítulo traz o primeiro artigo da dissertação. O capítulo é dividido em um resumo, introdução, método proposto, experimentos, resultados, discussões e conclusão.

O terceiro capítulo abarca o segundo artigo da dissertação. A segmentação do capítulo é descrita pelas seções de introdução, métodos, resultados e discussões e conclusão.

O capítulo final é composto pelas considerações finais e possíveis desdobramentos do trabalho.

1.6 Delimitações da Pesquisa

O presente estudo possui as seguintes delimitações:

- Não serão apresentadas novas ferramentas de aprendizagem de máquina ou decomposição matricial, restringindo-se a combinar tais ferramentas de forma a gerar novas abordagens para seleção de variáveis;
- Somente métodos supervisionados de aprendizagem de máquina são utilizados nos processos de seleção de variáveis;
- O novo ranking de importância de variáveis não é comparado diretamente com outros rankings; e
- Os bancos de dados utilizados restringem-se a comprimentos de onda de medicamentos e de dados de processos industriais e propriedades químicas de amostras de biodiesel e diesel brasileiro.

1.7 Referências

ALFARO, E.; GÁMEZ, M.; GARCÍA, N. *Ensemble classification methods with applications in R*. [S.l.]: John Wiley & Sons, 2019. ISBN 9781119421573.

ANZANELLO, M. *et al.* Multicriteria wavenumber selection in cocaine

classification. *Journal of Pharmaceutical and Biomedical Analysis*, v. 115, p. 562 – 569, 2015. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708515301060>.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. *Chemometrics and Intelligent Laboratory Systems*, v. 97, n. 2, p. 111–117, 2009. ISSN 0169-7439.

ANZANELLO, M. J. *et al.* A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. *Journal of Pharmaceutical and Biomedical Analysis*, v. 83, n. Supplement C, p. 209–214, 2013. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708513001970>.

ARDAKANI, A.; KORDNAEIJ, A. Soil compaction parameters prediction using gmdh-type neural network and genetic algorithm. *European Journal of Environmental and Civil Engineering*, Taylor & Francis, v. 23, n. 4, p. 449–462, 2019. Disponível em: <https://doi.org/10.1080/19648189.2017.1304269>.

ASYALI, M. H. *et al.* Gene expression profile classification: a review. *Current Bioinformatics*, Bentham Science Publishers, v. 1, n. 1, p. 55–73, 2006.

BOCKLITZ, T. Chemometrics: data driven extraction for science. *Analytical and Bioanalytical Chemistry*, v. 411, n. 14, 2019. ISSN 1618-2650. Disponível em: <https://doi.org/10.1007/s00216-019-01786-2>.

BRERETON, R. G. *Chemometrics for Pattern Recognition*. 1. ed. [S.l.]: Wiley, 2009. ISBN 0470987251,9780470987254.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. Computers & Electrical Engineering, Elsevier, v. 40, n. 1, p. 16–28, 2014. Disponível em: <https://doi.org/10.1016/j.compeleceng.2013.11.024>.

CRESWELL, J. W. Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: *Projeto de pesquisa métodos qualitativo, quantitativo e misto*. [s.l.] Penso Editora, 2010.

CUSTERS, D. *et al.* ATR-FTIR spectroscopy and chemometrics: An interesting tool to discriminate and characterize counterfeit medicines. *Journal of Pharmaceutical and Biomedical Analysis*, v. 112, p. 181–189, 2015. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708514005342>.

DASH, M.; LIU, H. Feature selection for classification. *Intelligent Data Analysis*, v. 1, n. 1, p. 131–156, 1997. ISSN 1088-467X.

FAN, J.; LI, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006. Disponível em: <https://arxiv.org/abs/math/0602133>.

FERRÃO, M. F. *et al.* Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. *Fuel*, v. 90, n. 2, p. 701–706, 2011. ISSN 00162361.

FIX, E.; HODGES, J. L. Discriminatory analysis, nonparametric discrimination. *Technical Report 4*, USAF School of Aviation Medicine, 1951. Disponível em: <https://apps.dtic.mil/docs/citations/ADA800276>.

GAUCHI, J.-P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, v. 58, n. 2, p. 171–193, oct 2001. ISSN 01697439. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743901001587>.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. The *Journal of Machine Learning Research*, v. 3, p. 1157–1182, 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944968>.

HE, X.; CAI, D.; NIYOGI, P. Laplacian score for feature selection. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2005. (NIPS'05), p. 507–514. Disponível em: <http://dl.acm.org/citation.cfm?id=2976248.2976312>.

HUANG, Q. *et al.* Exploiting local coherent patterns for unsupervised feature ranking. *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), v. 41, n. 6, p. 1471–1482, 2011. ISSN 1083-4419. <https://ieeexplore.ieee.org/document/5887432>.

IVAKHNENKO; G., A. The Group Method of Data of Handling; A rival of the method of stochastic approximation. *Soviet Automatic Control*, v. 13, p. 43–55, 1968. Disponível em: <https://ci.nii.ac.jp/naid/10004319713/>.

JONES, B. *Avoiding Data Pitfalls*: How to steer clear of common blunders when working with data and presenting analysis and visualizations. [S.l.]: Wiley, 2020. ISBN 9781119278160.

KANTARDZIC, M. *Data Mining: Concepts, models, methods, and algorithms*. 3rd. ed. [S.l.]: Wiley-IEEE Press, 2020. ISBN 1119516048,9781119516040.

KONDO, T. GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem. In: *Proceedings of the 37th SICE Annual Conference. International Session Papers*. Soc. Instrum. Control Eng, 1998. p. 1143–1148. Disponível em: <http://ieeexplore.ieee.org/document/742993/>.

KONDO, T.; PANDYA, A.; ZURADA, J. Logistic GMDH-type neural networks and their application to the identification of the X-ray film characteristic curve. In: *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*. IEEE, 1999. v. 1, p. 437–

442. ISBN 0-7803-5731-0. Disponível em: <http://ieeexplore.ieee.org/document/814131/>.

KUMAR, V.; MINZ, S. Feature selection: A literature review. Smart CR, v. 4, p. 211–229, 2014. <https://www.cc.gatech.edu/~hic/CS7616/Papers/Kumar-Minz-2014.pdf>.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 4, p. 491–502, 2005. ISSN 1558-2191. <https://ieeexplore.ieee.org/document/1401889>.

RAO, H. *et al.* Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, v. 74, p. 634–642, 2019. ISSN 1568-4946. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1568494618305933>.

VENKATESH, B.; ANURADHA, J. A review of feature selection and its methods. *Cybernetics and Information Technologies*, Sciendo, Berlin, v. 19, n. 1, p. 3–26, 2019. ISSN 1314-4081. Disponível em: <https://content.sciendo.com/view/journals/cait/19/1/article-p3.xml>.

YANG, G. *et al.* Review of anti-counterfeiting of prints based on infrared spectroscopy. In: ZHAO, P. *et al.* (Ed.). *Advances in Graphic Communication, Printing and Packaging*. Singapore: Springer Singapore, 2019. p. 150–156. ISBN 978-981-13-3663-8. Disponível em: <ttps://link.springer.com/chapter/10.1007/978-981-13-3663-8_22>.

ZHAO, Z.; LIU, H. *Spectral Feature Selection for Data Mining*. [S.l.]: CRC Press, 2011. ISBN 9781000023077

2 PRIMEIRO ARTIGO

SELEÇÃO DE COMPRIMENTOS DE ONDA BASEADA EM DECOMPOSIÇÃO DE VALORES SINGULARES PARA CLASSIFICAÇÃO DE AMOSTRAS

Artigo publicado no periódico Forensic Science International – Qualis A2. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073820300530>

Resumo: A disseminação de medicamentos falsificados é um risco à saúde pública. Técnicas como Reflexão Total Atenuada no Infravermelho com Transformada de Fourier (ATR- FTIR) são comumente adotadas para detecção de falsificações. Contudo, o espectro gerado pelo ATR-FTIR tipicamente resulta em centenas de comprimentos de onda (COs), reduzindo o desempenho dos métodos de classificação na discriminação entre medicamentos autênticos e falsificados. O presente trabalho propõe um novo método para selecionar um subconjunto reduzido composto por COs que aprimoram o desempenho do classificador. A decomposição matricial de valores singulares (SVD) é utilizada para gerar um índice de importância dos COs, o qual é integrado à técnica de classificação k-Nearest Neighbor (KNN). Na sequência, um processo iterativo cria modelos KNN adicionando os COs em ordem decrescente do índice de importância. Os COs que gerarem incremento de acurácia são selecionados. Quando aplicada a dados ATR-FTIR de Cialis®, a abordagem proposta reteve, em média, 0,51% dos COs com 100% acurácia na classificação; nas classificações de Viagra® o método produziu classificações perfeitas com, em média, 0,17% dos COs originais.

Palavras-chave: Seleção de comprimentos de onda. Medicamentos falsificados. SVD. KNN. ATR-FTIR.

2.1 Introdução

A produção e comercialização de medicamentos falsificados representam um problema crescente à saúde pública (BEEN *et al.*, 2011) em decorrência da falta de garantias sobre os controles das dosagens farmacêuticas e condições de produção (ANZANELLO *et al.*, 2014). A disseminação das falsificações é impulsionada pelo fácil acesso aos recursos necessários para sua fabricação e por seu amplo suprimento à disposição pela internet (SACRÉ *et al.*, 2011). Por isso, agências de segurança pública buscam constantemente métodos forenses que aumentem a precisão e reduzam custos na identificação das falsificações.

Autoridades de controle utilizam diferentes técnicas analíticas para avaliar a pureza dos medicamentos, as quais incluem processamento de imagens do medicamento (JUNG *et al.*, 2012), dados de controle físico dos medicamentos (ORTIZ *et al.*, 2012b), perfil inorgânico por espectroscopia de fluorescência de raios-X (ORTIZ *et al.*, 2012a), perfil orgânico por espectroscopia de massa por ionização com elétron e espectroscopia por infravermelho (ORTIZ *et al.*, 2013). Os dados extraídos por essas técnicas são então acoplados a métodos multivariados, adaptados à avaliação de amostras (SALARI; YOUNG, 1998). Essas técnicas geralmente resultam em um grande número de COs. No entanto, análises forenses geralmente dão origem a um número limitado de amostras, sendo comum a existência de mais COs que observações. À luz disso, métodos estatísticos multivariados, como *Análise de Componentes Principais* (PCA), índice de similaridade e análise de agrupamento hierárquico são aplicados para identificar os COs com maior poder discriminatório, possibilitando diversas naturezas de análise (FAN; LI, 2006; JOHNSTONE; TITTERINGTON, 2009; ANZANELLO *et al.*, 2013).

O estudo de Yu *et al.* (2020) segmentou o processo de seleção de COs em três etapas. Primeiramente, o autor aplicou Intervalo Parcial Mínimos Quadrados (iPLS) para selecionar intervalos de COs; em seguida, refinou a seleção apurando a importância de cada um dos COs, utilizando Parcial Mínimos Quadrados (PLS); por fim, selecionou os COs aplicando *Variable combination population analysis* (VCPA). Por sua vez, Anzanello *et al.* (2013) apresentam um método para seleção de COs e identificação de medicamentos autênticos no qual é proposto um índice baseado em PCA para classificar os COs de acordo com o maior poder de discriminação. O processo de seleção dos COs apoiou-se em *backward elimination*, começando com todos os comprimentos disponíveis e removendo-os, um a um, do índice mais baixo para o mais alto. O critério de decisão de retenção do comprimento de onda foi a acurácia resultante da aplicação da técnica de classificação KNN (FIX; HODGES, 1951).

Alternativamente ao PCA, o SVD (GOLUB; KAHAN, 1965) pode ser utilizado para compor o índice de importância dos COs, apresentando diversas vantagens sobre o PCA. Elliott *et al.* (1999) propuseram o uso de SVD para a análise em espectrometria de ressonância magnética nuclear alegando que,

diferentemente do PCA, o SVD permite análises em conjuntos de dados espectrais com qualquer quantidade de variação na fase espectral. Barkhuijsen *et al.* (1985) usaram SVD como uma alternativa à transformada de Fourier na análise de sinais da espectroscopia de Ressonância Magnética Nuclear (NMR). Entre suas vantagens, quando comparado ao PCA, o SVD permite trabalhar com matrizes retangulares, enquanto o PCA requer o uso de matrizes quadradas (ABDI; WILLIAMS, 2010). Por isso, normalmente o conjunto de dados é transformado em uma matriz de covariância ou correlação antes de prosseguir para o PCA, tornando, assim, o resultado dependente do quão bem os dados são representados na matriz quadrada gerada pelo pré-processamento (DHAR; SHIMAMURA, 2015). Além disso, Skillicorn (2007) ressalta que o cálculo da matriz de correlação representa um custo computacional adicional, que inclusive pode ser alto dependendo do banco de dados em análise. O autor também relata como vantagem sobre o PCA o fato de que o SVD analisa os atributos de forma conjunta, enquanto o PCA conduz tal análise de maneira independente. Por fim, Barkhuijsen *et al.* (1985) relataram maior estabilidade numérica na distinção entre sinal e ruído quando o SVD é aplicado.

A seleção de variáveis é um grande aliado na identificação de adulterações e falsificações, tornando-o um excelente método para a área das ciências forenses. Pereira *et al*. (2016) utilizam com sucesso o método *Ordered Predictor Selection* (OPS) integrado com análise discriminante de mínimos quadrados parciais (PLS-DA) para a seleção das variáveis de espectrometria de massa em spray de papel (PS-MS) e na identificação de cervejas falsificadas. Kahmann *et al*. (2018) propõem um método de seleção de COs para identificação e quantificação da cocaína e seus adulterantes, no qual demonstram que um menor número de COs contribui para uma análise mais rápida e precisa. Além de reduzir o tempo de análise da amostra, a identificação específica de COs pode levar ao uso de equipamentos portáteis, facilitando o trabalho *in situ*. Estudos utilizando espectrômetros portáteis, calibrados por uma seleção de variáveis, têm mostrado excelentes resultados para estimar a maturação das uvas (GIOVENZANA *et. al.*, 2014), teor de iodo em óleo comestível (YAN *et. al.*, 2018) e qualidade do biodiesel (MÁQUINA *et. al*, 2017).

O presente artigo propõe um novo método para a seleção dos COs mais relevantes para classificar amostras de medicamentos em autênticas ou falsificadas. A estrutura tem como proposta encontrar a maior acurácia de classificação com o número mínimo de COs retidos. Para tanto, a matriz de dados original é decomposta usando SVD e um índice de importância para cada comprimento de onda é calculado. O índice de importância orienta a composição de subconjuntos de COs, que são submetidos ao KNN para classificação. O modelo de KNN responsável pela maior acurácia revela o subconjunto de COs recomendado para classificação.

2.2 Método proposto e experimentos

2.2.1 Decomposição de Valores Singulares (SVD) e k-Nearest Neighbor (KNN)

SVD de uma matriz A é a fatorização de A no produto de três matrizes: U, Δ e V^T. As matrizes U e V^T contêm os vetores singulares esquerdos e direitos, respectivamente. A matriz Δ é diagonal, com valores reais não negativos, de acordo com equação 2.1 (YANAI *et al.*, 2011). De tal forma que, A é a matriz de dados com valores reais positivos, U é a matriz com os vetores singulares ortogonais esquerdos, Δ é a matriz de valores singulares, e V^T é a matriz com os vetores singulares ortogonais direitos.

$$A = U\Delta V^T \tag{2.1}$$

A quantidade de variância explicada, γ, requer o cálculo dos autovalores λ, de acordo com a equação 2.2 (YANAI *et al.*, 2011), onde j é o índice de valores singulares e autovalores correspondentes, r é o número total de valores singulares, λ_j é o *j*-ésimo autovalor, e μ_j é o *j*-ésimo valor singular.

$$\lambda_j = \mu_j^2 (j = 1, \dots, r) \tag{2.2}$$

O *j*-ésimo valor singular, μ_j, é determinado de forma que a variância entre os valores singulares seja maximizada, enquanto a quantidade de variância

explicada por cada valor singular é representada pelo coeficiente γ_j, de acordo com equação 2.3.

$$\gamma_j = \frac{\lambda_j}{\sum_{j=1}^{r} \lambda_j} \tag{2.3}$$

Por sua vez, a técnica de classificação KNN é um método supervisionado de reconhecimento de padrões (SAMMUT; WEBB, 2011) que, ao usar uma partição do banco de dados como treinamento, reconhece as relações entre os COs e classes (autêntico ou falsificado). O KNN classifica cada elemento da partição restante de acordo com a maioria dos k vizinhos mais próximos. O número de vizinhos, k, é definido de forma a maximizar uma medida de desempenho de classificação.

2.2.2 Método proposto

O método proposto é dividido em três estágios, como descreve a figura 2.1: (i) divisão do bando de dados em porções de treino e teste; (ii) criação do índice de importância dos COs; e (iii) processo iterativo para seleção dos COs.

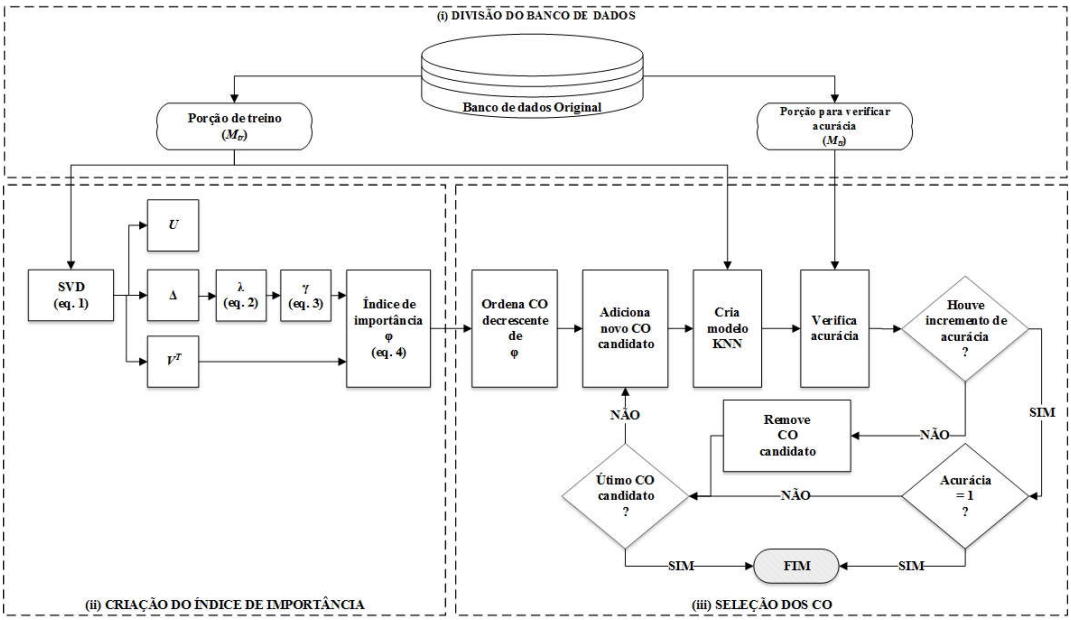

Figura 2.1 - Fluxograma do método proposto

Uma matriz com M observações descritas por N COs é aleatoriamente dividida em duas partições: M_{tr} para treino e M_{ts} para verificação da acurácia, com $M_{tr} + M_{ts} = M$. A amostra de treino é utilizada para computar o SVD. A decomposição SVD (equação 2.1) gera a matriz com os vetores singulares direitos V^T, contendo cada vetor singular v_j, e a matriz com valores singulares Δ, com cada valor singular μ_j na diagonal principal. Então, calculam-se os autovalores de acordo com equação 2.2 para encontrar o coeficiente de variância explicada para cada dimensão latente SVD, γ_j (equação 2.3). Finalmente, calcula-se o índice de importância de cada comprimento de onda, φ_a, conforme equação 2.4. Os COs com maior φ são considerados mais relevantes para discriminação.

$$\varphi_a = \sum_{j=1}^{r} \gamma_j \; v_{aj}^2, a = 1, \dots, n \tag{2.4}$$

Um processo iterativo é então iniciado. O KNN é executado utilizando como preditor o comprimento de onda com maior e como variável de resposta a classe que identifica autenticidade (autêntico ou falsificado). A cada ciclo de iteração um novo comprimento de onda é adicionado na ordem decrescente de φ_a, compondo assim o subconjunto de preditores para execução do KNN. Calcula-se a acurácia com a partição de verificação; se houver incremento da acurácia o comprimento de onda adicionado será mantido no subconjunto de preditores; em caso contrário, será removido. As iterações seguem até que a acurácia seja de 100% ou tenha testado o último comprimento de onda (com o menor φ).

2.2.3 Plano amostral para coleta das amostras

As amostras autênticas de Cialis® TAD 20mg foram compostas por 28 comprimidos, 8 dos quais foram fornecidos por Eli Lilly do Brasil Ltda Laboratórios (São Paulo, Brasil); as 20 amostras restantes pertencentes a oito lotes diferentes foram compradas na Dimed S/A Distribuidora de Medicamentos (Porto Alegre, Brasil). As amostras autênticas de Viagra® SLD 50mg foram compostas por 25 comprimidos, dos quais 19 eram de seis lotes diferentes

comprados na Dimed S/A Distribuidora de Medicamentos (Porto Alegre, Brasil) e seis foram fornecidos pela Pfizer Ltda Laboratórios (São Paulo, Brasil). Para as amostras falsificadas, 104 comprimidos foram fornecidos pela Polícia Federal do Brasil (Porto Alegre, Brasil). Todas as amostras foram analisadas por ATR-FTIR.

As amostras de medicamentos falsificados e autênticos foram maceradas em argamassa de porcelana antes da análise. Algumas amostras utilizaram um revestimento de filme que foi removido após o esmagamento. Para as amostras que não possuíam esse filme, seu revestimento foi homogeneizado durante o processo de maceração. As amostras foram analisadas usando um espectrômetro infravermelho por transformada de Fourier Nicolet 380 (Nicolet Instrument Co., Madison, Wisconsin, EUA) equipado com um detector DTGS (sulfato de triglicina deuterado) acoplado a um amostrador ATR de diamante Smart Orbit.

Os espectros foram obtidos no modo de transmitância e posteriormente convertidos em absorvância. As amostras foram lidas em triplicado, na faixa de 4000-525cm^{-1}, com 16 varreduras e resolução espectral de 4cm^{-1}. Os dados espectrais foram obtidos usando o software EZ OMNIC, versão 7.2a (Nicolet Instrument Co.). O cristal de ATR foi limpo com acetona entre cada análise e o espectro de fundo foi obtido a cada hora. Nenhum pré-tratamento foi utilizado nos espectros.

Os dados espectroscópicos foram pré-processados para corrigir e eliminar ruídos de aferição. O filtro de suavização Savitzky-Golay (polinômio de 2ª ordem e largura da janela 13) foi usado para reduzir o ruído (KRAKOWSKA *et. al.*, 2016). A variação normal padrão (SVN) também foi aplicada a fim de minimizar os efeitos multiplicativos da dispersão e do tamanho das partículas das amostras (PARHIZKAR *et. al.,* 2017).

2.2.4 Experimentos

O conjunto de amostras foi dividido aleatoriamente em duas partições: treino, M_{tr}, e verificação da acurácia, M_{ts}. Três proporções diferentes para treinamento e verificação da acurácia foram usadas neste procedimento de divisão: 60-40%, 75-25% e 90-10%. Dez mil simulações foram realizadas para

cada proporção, separando aleatoriamente as observações entre os conjuntos de treinamento e verificação da acurácia. Em cada simulação, três métricas de desempenho sobre a classificação foram calculadas: sensibilidade, que é a proporção corretamente classificada como autênticas; especificidade, que é a proporção corretamente classificada como falsificadas; e acurácia, que é a proporção de amostras classificadas corretamente. O número de k vizinhos foi definido como três, via validação cruzada. Todos os experimentos computacionais usaram a linguagem de programação R, versão 3.6.2 (R CORE TEAM, 2019), com os pacotes: caret (KUHN, 2020), magrittr (BACHE e WICKHAM, 2014), tidyverse (WICKHAM, 2019), openxlsx (SCHAUBERGER, 2019) e progress (CSÁRDI, 2019).

2.3 Resultados e discussões

A tabela 2.1 mostra a acurácia, sensibilidade, especificidade e porcentagem médias de COs retidos para as três partições do conjunto de dados Cialis®; o desvio padrão é apresentado entre parênteses.

Tabela 2.1 - Médias de desempenho das diferentes partições de treino-teste para Cialis®

Métricas de desempenho	Partição do banco de dados % treino–% teste para Cialis®					
	60-40%		75-25%		90-10%	
	Anzanello *et al.* (2013)	Método proposto	Anzanello *et al.* (2013)	Método proposto	Anzanello *et al.* (2013)	Método proposto
Acurácia	0.9806 (.0140)	1 (0)	0.9861 (.0128)	1 (0)	0.9897 (.0199)	1 (0)
Sensibilidade	0.9485 (.0485)	1 (0)	0.9610 (.0414)	1 (0)	0.9708 (.0586)	1 (0)
Especificidade	0.9943 (.0090)	1 (0)	0.9968 (.0080)	1 (0)	0.9988 (.0071)	1 (0)
CO Retidos (%)	1.84	0.92	1.82	0.73	1.41	0.51

O método proposto utilizou 0,51% dos 661 COs originais para produzir classificações perfeitas (isto é, acurácia, sensibilidade e especificidade médias iguais a 1). Esse resultado é superior à proposta de Anzanello *et al.* (2013), a qual exigia pelo menos 1,84% dos COs para atingir uma precisão média de 0,9806, sensibilidade de 0,9485 e especificidade de 0,9943, e respectivos desvios padrão de 0.0199, 0.0586 e 0,0071. Na partição de 60-40%, usando 0,92% dos COs, o método proposto alcançou classificações perfeitas. Isso sugere que maiores proporções de treinamento fornecem ao método

informações mais relevantes para estimar o índice de importância dos COs, contribuindo para identificação do subconjunto de variáveis. Também se observa maior estabilidade dos resultados entre as simulações da proposta, tendo em vista que com classificações perfeitas a variabilidade foi zero.

A tabela 3.2 mostra os COs selecionados com mais frequência. Nos experimentos, os COs mais retidos são do intervalo 999cm^{-1} até 1059cm^{-1}, para todas as porções treino-teste. O número médio de COs retidos é menor que os encontrados por Anzanello *et al.* (2013), sugerindo que o método proposto com base no SVD pode produzir subconjuntos de COs menores, porém informativos.

Tabela 2.2 - Percentual frequência dos COs retidos nos subconjuntos de Cialis®

Comprimentos de onda (cm^{-1})	Partições do banco de dados		
	Treino 60%, Teste 40%	Treino 75%, Teste 25%	Treino 90%, Teste 10%
1022	98.54	99.87	99.94
1024	81.79	83.63	86.21
1020	82.46	68.64	40.10
1032	65.88	53.12	26.41
1028	58.62	39.90	14.16
1026	48.98	37.12	18.26
1030	44.85	28.53	9.65
1018	31.26	21.12	8.49
1034	22.13	11.86	3.83
1016	12.91	5.60	3.14
1014	9.04	2.18	0.69
1012	5.10	1.06	0.04
1059	3.05	1.01	0.03
999	2.01	0.69	0.04
1001	1.64	0.37	0.01
1009	1.06	0.39	0.06
1005	1.11	0.25	0.01
1055	0.40	0.13	0.01
1038	0.17	0.04	0.01

A figura 2.1 mostra os COs com a maior frequência de retenção para a partição de 90-10%. Além do número reduzido de COs retidos, também é notável a pequena variação nos COs retidos. Isso sugere robustez e consistência do índice de importância proposto. As regiões espectrais retidas estão relacionadas

a deformações de =C-H fora do plano e trechos axiais dos limites de C-O e C-N encontrados nas estruturas de taladafil (COLTHUP *et al.*, 1990).

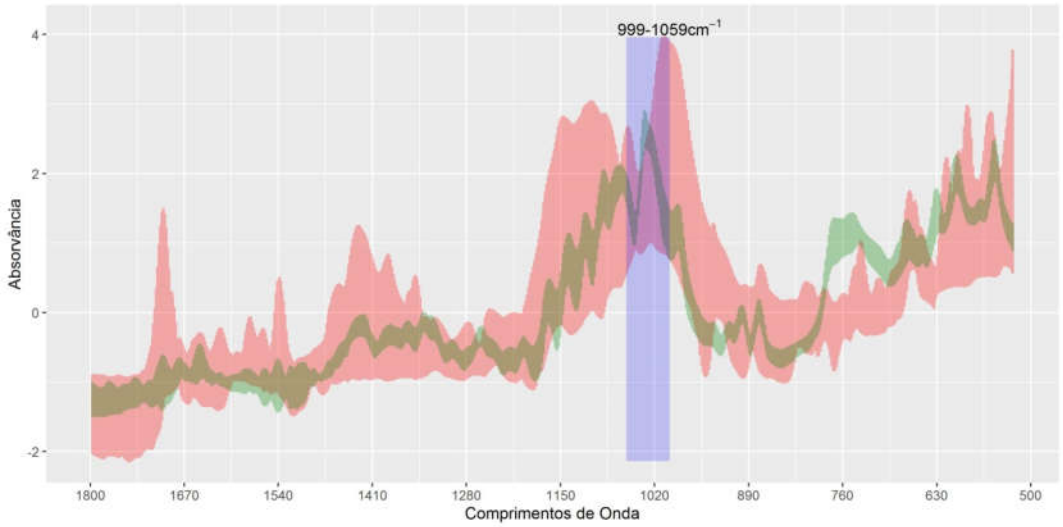

Figura 2.2 - Espectros de Cialis® representativos dos subconjuntos autênticos (em verde), falsificados (em vermelho) e dos COs retidos (tom azul).

O método também demonstra desempenho superior nas amostras de Viagra® em comparação aos resultados de Anzanello *et al.* (2013), como pode ser visto na tabela 3. A porcentagem média de COs na partição de 60-40% foi de 0,23%, com acurácia, sensibilidade e especificidade iguais a 1. Na proporção de 90-10%, o método obteve os melhores resultados, com percentual médio de COs de 0,17% e classificações perfeitas. Para a proporção de 75 a 25%, a porcentagem média de COs obtida foi de 0,21%, com acurácia, sensibilidade e especificidade iguais a 1. Em todas as aferições a variabilidade das simulações foi menor nos resultados do método proposto. Isso demonstra que a abordagem sugerida gerou subconjuntos de variáveis que, além de resultarem em precisão superior na classificação, conduziram a resultados mais estáveis.

Tabela 2.3 - Médias de desempenho das diferentes porções de treino-teste para Viagra®

Métricas de desempenho	Partição do banco de dados % treino–% teste para Viagra®					
	60-40%		75-25%		90-10%	
	Anzanello *et al.* (2013)	Método proposto	Anzanello *et al.* (2013)	Método proposto	Anzanello *et al.* (2013)	Método proposto
Acurácia	0.9100 (.0378)	1 (0)	0.9267 (.0366)	1 (0)	0.9349 (.0477)	1 (0)
Sensibilidade	0.9256 (.0585)	1 (0)	0.9349 (.0624)	1 (0)	0.9068 (.1087)	1 (0)
Especificidade	0.9013 (.0590)	1 (0)	0.9223 (.0561)	1 (0)	0.9503 (.0618)	1 (0)
CO Retidos (%)	7.72	0.23	6.81	0.21	6.05	0.17

À medida que o tamanho da partição de treino aumenta também é verificado um aumento no desempenho, indicando que o SVD usado para gerar o índice de importância do COs se beneficia de mais informações. Além disso, a variabilidade nas métricas de desempenho (acurácia, sensibilidade e especificidade) são menores que os resultados de Anzanello *et al.* (2013), indicando que o método proposto é robusto às variações nas amostras de treinamento.

A tabela 3.4 mostra os COs com alta frequência de inclusão estão no intervalo de 553cm^{-1} à 557cm^{-1}. Estes comprimentos de onda estão relacionados a deformação angular do tipo tesoura do SO2 e/ou estiramento do C-H aromático, ambos presentes no ingrediente sildenafila (COLTHUP *et al.*, 1990).

Tabela 2.4 - Percentual frequência dos COs retidos nos subconjuntos de Viagra®

CO (cm^{-1})	Partições do banco de dados		
	Treino 60%, Teste 40%	Treino 75%, Teste 25%	Treino 90%, Teste 10%
557	100	100	100
553	17.61	14.65	8.59
555	21.79	10.45	1.51

Na figura 2.3 a banda do espectro de COs com maior frequência de seleção é destacada em azul, na faixa de COs 553-557cm^{-1}.

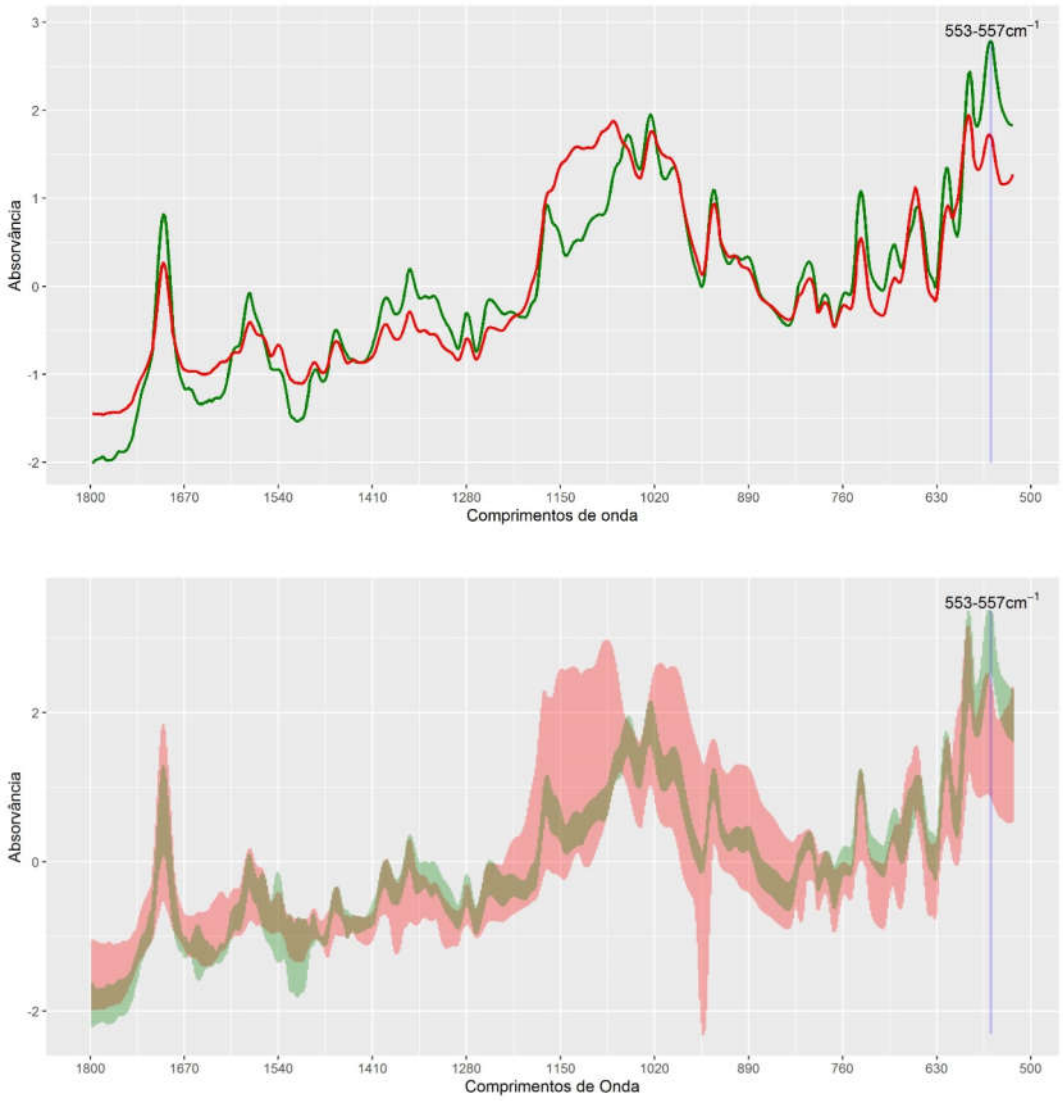

Figura 2.3 - Espectros de Viagra® representativos dos subconjuntos autênticos (em verde), falsificados (em vermelho) e dos COs retidos (tom azul).

Pode-se afirmar que a transformação da matriz realizada diretamente nos dados contribuiu para o aumento do desempenho. Outra vantagem dessa estrutura em comparação com Anzanello *et al.* (2013) está no fato de que os COs foram selecionados com base em um processo de inclusão *forward*, e os COs que não contribuíssem para a acurácia da classificação eram removidos do subconjunto selecionado.

2.4 Conclusões

Técnicas analíticas tradicionais, como a espectroscopia (ATR-FTIR), resultam em um grande número de COs que tendem a reduzir o desempenho dos modelos de classificação. Por isso, a seleção do COs se torna altamente relevante. A estrutura proposta inova ao utilizar a decomposição SVD para calcular a importância de cada COs. Em vez de aplicar o PCA como em Anzanello *et al.* (2013) para estimar esse índice de importância, o método proposto se baseia em SVD.

Diferentemente do PCA – que depende de quão bem os dados são representados na matriz quadrada gerada no pré-processamento - o SVD trabalha com matrizes retangulares que permitem o uso de dados sem pré-processamento.

O uso de métodos de seleção de variáveis nos dados do ATR-FTIR é de grande importância, pois torna a análise mais eficiente. A seleção correta dos COs permite uma classificação mais precisa das amostras, característica necessária no campo da ciência forense. Além disso, a redução dos COs necessários para a classificação das amostras permite a utilização de dispositivos portáteis mais baratos, por tratarem de intervalos curtos e específicos de COs, o que ajudaria o trabalho de campo das autoridades policiais em fiscalizações.

Amostras autênticas e falsificadas de Cialis® e Viagra® foram utilizadas para validar a proposta. Os resultados mostram que um subconjunto bastante reduzido de COs alcançou uma alta acurácia na classificação das amostras. Além disso, demonstrou-se que não apenas os principais ingredientes dos medicamentos são elementos importantes, mas também os excipientes mostram-se relevantes para a correta classificação dos medicamentos nas classes autêntica ou falsificada.

Sugere-se, entre possíveis estudos futuros, comparar o desempenho do índice de importância das variáveis com outros índices da literatura, como Laplacian Score (HE et al., 2005), Information Gain, Mutual Information, Gini Index (AGGARWAL, 2014). Também, propõem-se experimentos com outros

métodos de aprendizagem de máquina, considerando, inclusive, bases de dados multiclasse.

2.5 Referências

ABDI, H.; WILLIAMS, L. J. Análise de Componentes Principais. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 2, n. 4, p. 433–459, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101>.

ANZANELLO, M. J. *et al.* PLS-DA wavenumber selection for the categorization of medicine samples based on multiple criteria. *Forensic Science International*, v. 242, n. Supplement C, p. 111–116, 2014. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073814002709>.

ANZANELLO, M. J. *et al.* A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. *Journal of Pharmaceutical and Biomedical Analysis*, v. 83, n. Supplement C, p. 209–214, 2013. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708513001970>.

BACHE, S. M.; WICKHAM, H. magrittr: A Forward-Pipe Operator for R. 2014. R package version 1.5. Disponível em: <https://CRAN.R-project.org/package=magrittr>.

BARKHUIJSEN, H. *et al.* Application of linear prediction and Decomposição de Valores Singulares (LPSVD) to determine NMR frequencies and intensities from the FID. *Magnetic Resonance in Medicine*, John Wiley and Sons, v. 2, n. 1, p. 86–89, 1985. ISSN 0740-3194,1522-2594. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910020111>

BEEN, F. *et al.* Profiling of counterfeit medicines by vibrational spectroscopy. *Forensic Science International*, v. 211, n. 1, p. 83–100, 2011. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073811002106>.

COLTHUP, N. B.; DALY, L. H.; WIBERLEY, S. E. *Introduction to Infrared and Raman Spectroscopy*. Third edit. [S.l.]: Academic Press, 1990. ISBN 978-0-12-182554-6.

CSÁRDI, G.; FITZJOHN, R. progress: Terminal Progress Bars. 2019. Disponível em: <https://cran.r-project.org/package=progress>.

ELLIOTT, M. A. *et al.* Spectral quantitation by Análise de Componentes Principais using complex Decomposição de Valores Singulares. *Magnetic Resonance in Medicine*, John Wiley and Sons, v. 41, n. 3, p. 450–455, 1999. ISSN 0740-3194,1522-2594.

FIX, E.; HODGES, J. L. Discriminatory Analysis, Nonparametric Discrimination. Technical Report 4, *USAF School of Aviation Medicine*, 1951. Disponível em: <https://www.jstor.org/stable/1403797?seq=1>.

GIOVENZANA, V.; BEGHI, R.; MALEGORI, C.; CIVELLI, R.;GUIDETTI, R. Wavelength selection with a view to a simplified handheld optical system to estimate grape ripeness, *American Journal of Enology and Viticulture*. 65 (2014) 117–123. doi:10.5344/ajev.2013.13024. Disponível em: <https://www.ajevonline.org/content/65/1/117>.

GOLUB, G.; KAHAN, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B*, Numerical Analysis, Society for Industrial and Applied Mathematics, v. 2, n. 2, p. 205–224, 1965. ISSN 0887459X. Disponível em: <http://www.jstor.org/stable/2949777>.

HE, X.; CAI, D.; NIYOGI, P. Laplacian score for feature selection. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge*, MA, USA: MIT Press, 2005. (NIPS'05), p. 507–514. <https://dl.acm.org/doi/10.5555/2976248.2976312>.

JUNG, C. R. *et al.* A new methodology for detection of counterfeit Viagra and Cialis tablets by image processing and statistical analysis. *Forensic Science International*, v. 216, n. 1, p. 92–96, 2012. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037907381100449X>.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; MARCELO, M. C. A.; FERRÃO, M. F.; ORTIZ, R. S.; MARIOTTI, K. C. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. *Journal of Pharmaceutical and Biomedical Analysis*. 152 (2018) 120–127. doi: 10.1016/j.jpba.2018.01.050. Disponível em: <https://doi.org/10.1016/j.jpba.2018.01.050>

KRAKOWSKA, B. CUSTERS, D. DECONINCK, E. DASZYKOWSKI, M. Chemometrics and the identification of counterfeit medicines—A review, Journal of *Pharmaceutical and Biomedical Analysis*. Anal. 127 (2016) 112–122. doi:10.1016/j.jpba.2016.04.016. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0731708516302035?via%3Dihub>.

KUHN, Max. caret: Classification and Regression Training. 2020. R package version 6.0-85. Disponível em: <https://CRAN.R-project.org/package=caret>.

LIANG, L.; Liang, L.; Wei, L.; Fang, G.; Xu, Feng; Deng, Yongjun; Shen, K.; Tian, Q.; Wu, T.; Zhu, B. Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v. 225, p. 117515, 2020. ISSN 1386-1425. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1386142519309059>.

MÁQUINA, A. D. V.; SOUZA, L. M.; BUIATTE, J. E.; SANTOS, D. Q.; NETO, W.B. Fast Quantitative and Qualitative Monitoring of Mafurra Biodiesel Content Using Fourier Transform Mid-Infrared Spectroscopy, Chemometric Tools, and Variable Selection, *Energy & Fuels*. 31 (2017) 571–577. doi:10.1021/acs.energyfuels.6b02079. Disponível em: <https://pubs.acs.org/doi/full/10.1021/acs.energyfuels.6b02079>.

ORTIZ, R. S. *et al.* Counterfeit Cialis® and Viagra® fingerprinting by ATR-FTIR spectroscopy with chemometry. *Forensic Science International*, v. 226, n. 1, p. 282–289, 2013. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073813000649>.

ORTIZ, R. S. *et al.* Fingerprinting of sildenafil citrate and tadalafil tablets in pharmaceutical formulations via X-ray fluorescence (XRF) spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, v. 58, n. Supplement C, p. 7–11, 2012. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708511005231>.

ORTIZ, R. S. *et al.* Physical profile of counterfeit tablets Viagra® and Cialis®. *Brazilian Journal of Pharmaceutical Sciences*, scielo, v. 48, n. 3, p. 487–495, 2012. ISSN 2175-9790. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-82502012000300016>.

PARHIZKAR, E.; GHAZALI, M.; AHMADI, F.; SAKHTEMAN, A. PLS-LS-SVM based modeling of ATR-IR as a robust method in detection and qualification of alprazolam, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 173 (2017) 87–92. doi:10.1016/j.saa.2016.08.055. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S138614251630508X>.

R CORE TEAM. R: *A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <https://www.r-project.org/>.

SACRÉ, P.-Y. *et al.* Impurity fingerprints for the identification of counterfeit medicines—A feasibility study. *Analytica Chimica Acta*, v. 701, n. 2, p. 224–231, 2011. ISSN 0003-2670. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0003267011007598>.

SALARI, A.; YOUNG, R. E. Application of attenuated total reflectance FTIR spectroscopy to the analysis of mixtures of pharmaceutical polymorphs. *International Journal of Pharmaceutics*, v. 163, n. 1, p. 157–166, 1998. ISSN 0378-5173. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378517397003785>.

SAMMUT, C.; WEBB, G. I. *Encyclopedia of Machine Learning*. 1st editio. ed. [S.l.]: Springer, 2011. ISBN 9780387307688.

SCHAUBERGER, P.; WALKER, A. openxlsx: Read, Write and Edit xlsx Files. 2019. Disponível em: <https://cran.r-project.org/package=openxlsx>.

WICKHAM, H. tidyverse: Easily Install and Load the 'Tidyverse'. 2019. Disponível em: <https://cran.r-project.org/package=tidyverse>.

YAN, H.; ZHANG, J.; GAO, J.; HUANG, Y.; XIONG, Y.; MIN, S. Towards improvement in prediction of iodine value in edible oil system based on chemometric analysis of portable vibrational spectroscopic data, *Scientific reports*. vol. 8, n. 14729. 2018. doi:10.1038/s41598-018-33022-9. Disponível em: <https://www.nature.com/articles/s41598-018-33022-9>.

YANAI, H.; TAKEUCHI, K.; TAKANE, Y. *Projection matrices, generalized inverse matrices, and Decomposição de Valores Singulares*. 1. ed. [S.l.]: Springer-Verlag

New York, 2011. (Statistics for Social and Behavioral Sciences). ISBN 9781441998866.

YU, H. Y.; Yun, Y.; Zhang W.; Chen H.; Liu, D.; Qiuping, Z.; Chen, W.; Chen, W. Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near-infrared spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v. 224, p. 117376, 2020. ISSN 1386-1425. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1386142519307668>.

ZHANG, F.; REN, H. W.; LI, J. P. Study of the Structural Properties of Microcrystalline Cellulose (MCC) Particles from Distillers Grains (DG) by XRD, FTIR and SEM. In: *Progress in Environmental Protection and Processing of Resource*. [S.l.]: Trans Tech Publications Ltd, 2013. (Applied Mechanics and Materials, v. 295), p. 339–344.

3 SEGUNDO ARTIGO

MÉTODO HÍBRIDO DE CLASSIFICAÇÃO BINÁRIA BASEADO EM ENSEMBLE LOGISTIC GMDH-TYPE NEURAL NETWORKS

Resumo: Processos industriais e químicos geram grandes volumes de variáveis, as quais podem apresentar níveis de ruído e correlações que dificultam o monitoramento e o controle de tais processos. Nesse contexto, a seleção do subconjunto de variáveis com maior poder preditivo representa um desafio. O presente estudo propõe um novo método para identificação do subconjunto de variáveis com maior poder de classificação de amostras industriais. Para tanto, a decomposição de valores singulares (SVD) é integrada com a técnica de classificação Ensemble Logistic GMDH-type Neural Networks (EL-GMDH-NN). O SVD é utilizado para calcular o índice de importância de cada variável. Um processo iterativo cria modelos adicionando variáveis na ordem decrescente do índice de importância. As variáveis que aumentam a acurácia de classificação são selecionadas. Quando aplicada a bancos de dados industriais, a abordagem proposta conduziu, nos melhores casos, a classificações 100% acuradas com somente 2.72% das 95 variáveis iniciais no banco de dados OXY. Na base de dados referente ao ponto de fulgor do biodiesel e diesel, 98.8% de acurácia com somente 0.33% das 1.738 variáveis. Todos os resultados se mostraram superiores quando comparados com Regressão Logística Stepwise, Logistic GMDH e a proposta de Anzanello et al. (2009).

Palavras-chave: Ranking não supervisionado. Stacking Ensemble. Group Method Data Handling.

3.1 Introdução

O processo de seleção de variáveis não é trivial, pois raramente as variáveis são totalmente independentes, sendo comum possuírem redundância e baixo nível de informação (HUNTER, 2000). De acordo com Anzanello *et al*. (2009), processos industriais e químicos geram um elevado número de variáveis, muitas delas correlacionadas. Bancos de dados podem ter variáveis irrelevantes com potencial para interferir nas variáveis reais, introduzindo heterogeneidade nos dados e gerando dependência (KHAIRE; DHANALAKSHMI, 2019). No entanto, a verificação de todos os subconjuntos possíveis de varáveis pode ser extremamente onerosa (precisamente $2^n - 1$, onde n é o número de variáveis) e rapidamente convergir para um problema NP-difícil, em razão do grande

número de subconjuntos possíveis (JENSEN; SHEN, 2008). Assim, métodos de estatística multivariada, decomposição matricial e técnicas de aprendizagem de máquina são comumente aplicados para encontrar os subconjuntos de dados que reduzem o ruído, valores incompletos, redundantes ou ausentes dos bancos de dados em análise (MATHEUS *et al*., 1993; DEOGUN *et al*., 1997; DEOGUN *et al*., 1998; ERIKSSON; WOLD, 2010). Como resultado, a identificação das variáveis mais relevantes contribui para a eficiência dos modelos preditivos, melhorando a precisão, reduzindo custo computacional e, especialmente em processos industriais, tem potencial de identificar variáveis que geram custo de aferição mas não oferecem retorno prático (LIU; MOTODA, 1998). Por esses motivos, a seleção de variáveis tem sido aplicada em diversas áreas.

Kira e Rendell (1992) utilizam a seleção de variáveis para eliminar testes de diagnóstico médico redundantes, reduzindo assim o custo e o tempo do processo. Huang *et al*. (2011) propõem uma nova abordagem de seleção de variáveis para identificar genes com potencial biomarcador multicâncer. Ming e Zhao (2018) descrevem a seleção de variáveis como chave para detecção e diagnóstico de falhas em processos industriais químicos. Em análises de Biodiesel e Diesel, Ferrão et al. (2011) propõem um modelo que seleciona as bandas de espectro mais adequadas para análises da qualidade utilizando propriedades físico-químicas. Para os autores, identificar as melhores bandas contribui para aprimorar o monitoramento da qualidade, o que se reflete na queima adequada do combustível nos motores, evitando depreciação acelerada e gerando economia de combustível.

Modelos para seleção de variáveis podem implementar estratégias que mesclam diferentes técnicas. Li *et al*. (2014) combinam estatística com aprendizagem de máquina para definir as variáveis mais importantes em processos financeiros do gerenciamento de risco corporativo. Anzanello *et al*. (2013) propõem um método de seleção de variáveis com uma abordagem heurística integrada, no qual aplica-se inicialmente a análise de componentes principais nos dados originais para criar um índice de importância das variáveis. Em seguida, modelos *k*-nearest neighbor são criados com a backward elimination das variáveis. Por fim, é selecionado o subconjunto de variáveis que compõem o modelo mais acurado. Comparado ao PCA, o SVD tem a vantagem

de usar matrizes retangulares sem a necessidade de pré-processamento para transformação de dados em matrizes de covariância ou matrizes de correlação (GOLUB; KAHAN, 1965). Ademais, Skillicorn (2007) destaca que o SVD analisa os atributos de forma conjunta, o que resulta em mais sensibilidade na identificação do ruído sobre todo o conjunto de dados. Da mesma forma, no âmbito da aprendizagem de máquina, métodos como Redes Neurais Artificiais (ANN) são uma alternativa para o KNN, em geral apresentando desempenho superior (BROMLEY; SACKINGER, 1991; MOOSAVIAN *et al*., 2013).

O presente artigo implementa um novo método, denominado SVD-GMDH, que seleciona variáveis para classificação binária de amostras oriundas de processos industriais. A proposta utiliza o SVD para calcular o índice de importância de cada variável e o integra ao método *Ensemble Logistic GMDH-type Neural Networks* (EL-GMDH-NN). O EL-GMDH-NN tem como principal vantagem agregar o desempenho de múltiplos algoritmos de aprendizagem, combinando-os em um único algoritmo de aprendizagem, denominado *meta learning*, como a rede neural logística Logistic GMDH-NN (KONDO *et al.*, 1999; ROKACH, 2010). A Logistic GMDH-NN otimiza o número de neurônios, camadas e variáveis de entrada mais relevantes (KONDO, 1998). Essa característica confere desempenho superior frente aos métodos tradicionais de redes neurais (ONWUBOLU, 2009; BRAGA *et al*., 2012; AHMDI *et al*, 2019). Ademais, quando o GMDH-NN é utilizado como *meta learning*, seu desempenho tende a ser ainda mais destacável (XIAO; HE, 2008; XIAO *et al.*, 2014).

Para analisar o desempenho da proposta SVD-GMDH foram aplicados experimentos sobre diversos conjuntos de dados caracterizados por uma classe binária de qualidade: conforme ou não conforme. Parte dos bancos se referem a processos de produção industrial, incluindo produção de nylon, emulsão na indústria de papel, polimerização em um processo de látex, produção de óxido de titânio e produção de antibióticos. Nesse coletivo de dados, a identificação do subconjunto de variáveis que mais contribui para correta classificação, entre um produto conforme ou não conforme, também fornece informações que possibilitam a revisão do processo de aferição. Essa análise pode levar a redução de custos no processo de apuração da qualidade, caso variáveis possam ser descartadas e não aferidas.

O segundo conjunto de dados analisado está relacionado a propriedades físico-químicas do Biodiesel e Diesel brasileiro. Nessa conjuntura, um algoritmo que classifica com eficiência amostras entre conformes ou não conformes oferece confiabilidade ao processo de classificação. Biodiesel e Diesel de má qualidade refletem diretamente em desperdício de combustível, desgaste acelerado do motor e custos com manutenção do veículo (Ferrão *et al.*, 2011). Além disso, tanto para o conjunto do Biodiesel e Diesel quanto para os processos industriais, a classificação incorreta de conformidade tende a gerar leso material e financeiro ao cliente (BRASIL, 1990) e pode inclusive impactar na imagem de uma marca (PASQUALINI, 2019).

3.2 Método proposto – SVD-GMDH

O desenvolvimento da proposta é organizado em três fases: (i) divisão do banco de dados, (ii) criação do índice de importância, e (iii) seleção das variáveis, conforme figura 3.1.

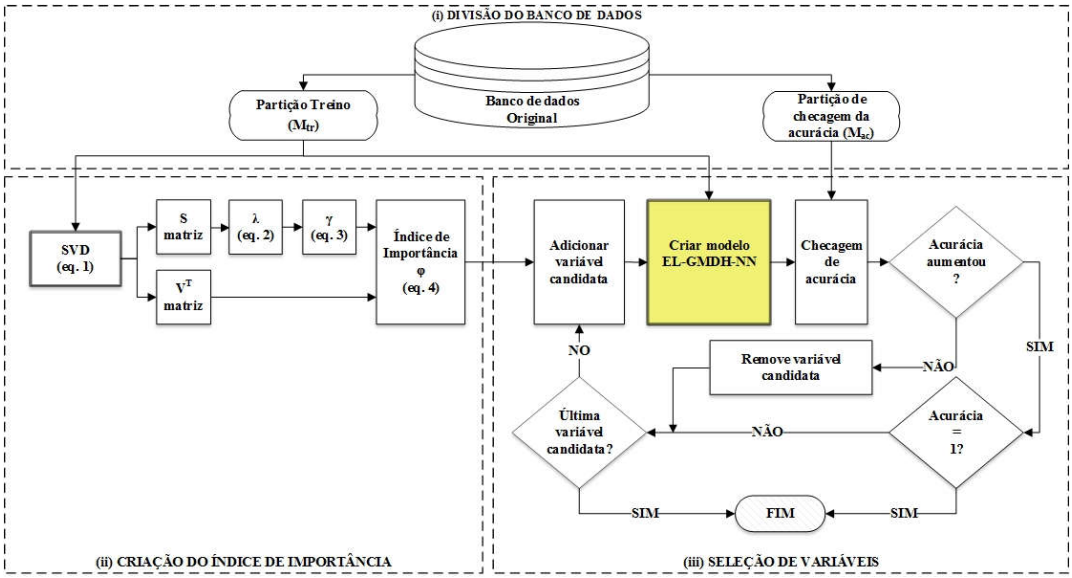

Figura 3.1 - Fluxograma do modelo SVD-GMDH

3.2.1 Divisão do banco de dados

Considera-se uma matriz com M observações, descritas por uma classe binária “0”, que indica observação com qualidade não conforme e “1”, que representa observação com qualidade conforme, contendo N variáveis com valores reais. Tal matriz é randomicamente dividida em duas partições: M_{tr} para treino e M_{ac} de teste, para teste, com $M_{\mathrm{tr}} + M_{\mathrm{ac}} = M$.

3.2.2 Criação do índice de importância

Skillicorn (2007) define o SVD como um método de decomposição matricial que transforma uma matriz $\mathrm{A}_{|\mathrm{mn}|} \in \mathbb{R}_{|\mathrm{mn}|}$, com M linhas e N colunas, e rank r, de forma a expor a quantidade de variância de um conjunto de dados utilizando variáveis latentes. Para o autor, a interpretação mais intuitiva é a geométrica: os dados da matriz $\mathrm{A}_{|\mathrm{mn}|}$ são transformados de tal forma que a variância é maximizada em um eixo; o método então posiciona os dados em um novo eixo que é, em relação ao primeiro, ortogonal e com variância maximizada; o método continua adicionando novos eixos ortogonais com variância maximizada a todos os anteriores, até r da matriz de dados. A fatoração SVD da matriz $\mathrm{A}_{|\mathrm{mn}|}$ é representada pelo produto de três matrizes $\mathrm{U}_{|\mathrm{rr}|}$, $\mathrm{S}_{|\mathrm{rr}|}$ e $\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|}$ (equação 3.1). As matrizes $\mathrm{U}_{|\mathrm{rr}|}$ e $\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|}$ são ortogonais esquerda e direita, respectivamente, e a matriz $\Delta_{|\mathrm{rr}|} \in \mathbb{R}_{+}$ é diagonal.

$$\mathrm{A}_{|\mathrm{mn}|} = \mathrm{U}_{|\mathrm{rr}|}\Delta_{|\mathrm{rr}|}\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|} \tag{3.1}$$

Onde $\mathrm{A}_{|\mathrm{mn}|}$ é a matriz original dos dados, $\mathrm{U}_{|\mathrm{rr}|}$ é a matriz com as coordenadas dos objetos no espaço estendido pelos novos eixos, $\Delta_{|\mathrm{rr}|}$ é a matriz de valores singulares com fator de escala que indica a importância relativa de cada novo eixo e, $\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|}$, é a matriz com os novos eixos. A partir do SVD calcula-se a quantidade de variância explicada, γ. Para calcular λ, é necessário calcular os autovalores λ, de acordo com equação 3.2.

$$\lambda_{\mathrm{j}} = \mu^2_{\mathrm{j}}(\mathrm{j} = 1, \ldots, \mathrm{r}) \tag{3.2}$$

onde, r é o número de valores singulares, j é o índice de valores singulares e autovalores correspondentes, λ_j é o *j*-ésimo autovalor e μ_j é o *j*-ésimo valor singular.

O *j*-ésimo valor singular μ_j é determinado de tal forma que a variância entre os valores singulares é maximizada. A quantidade de variação explicada por cada dimensão singular é representada pelo coeficiente γ_j, como na equação 3.3.

$$\gamma_j = \frac{\lambda_j}{\sum_{j=1}^{r} \lambda_j} \tag{3.3}$$

Então, o índice de importância é calculado, conforme equação 3.4.

$$\varphi_a = \sum_{j=1}^{r} \gamma_j \, v_{aj}^2, a = 1, \ldots, N \tag{3.4}$$

Por fim, o índice de importância é aplicado para ordenar as variáveis de forma decrescente (maiores valores de índice sugerem variáveis mais relevantes para o processo de classificação).

3.2.3 Seleção de variáveis

A fase de seleção de variáveis ocorre em um processo iterativo, no qual cada variável candidata é adicionada ao modelo Ensemble Logistic GMDH-NN de acordo com a abordagem de inclusão *forward*. A acurácia do modelo é verificada com a partição M_{ac}. Se houver aumento na acurácia, a variável candidata é mantida no subconjunto de variáveis selecionadas. Quando todas as variáveis forem testadas ou o subconjunto em análise atingir acurácia de 100%, o processo é encerrado. Como resultado, é obtido o subconjunto de variáveis recomendado para predição. O modelo EL-GMDH-NN utiliza a metodologia ensemble e é descrito, conforme figura 3.2, em cinco passos: (i) Divisão da partição de treino M_{tr} em sub-partições M_{sub-tr} e M_{sub-ts}; (ii) *Learners*: algoritmos de aprendizagem de máquina treinados com o subconjunto de variáveis selecionadas e a nova variável candidata, adicionada conforme ordem do índice de importância; (iii) *Predictions*: executar as predições utilizando

a sub-partição $M_{\mathrm{sub-ts}}$; (iv) Ensemble Logistic GMDH-NN: treinamento do modelo com a predição de cada *learner* e a binária classe da partição $M_{\mathrm{sub-ts}}$.

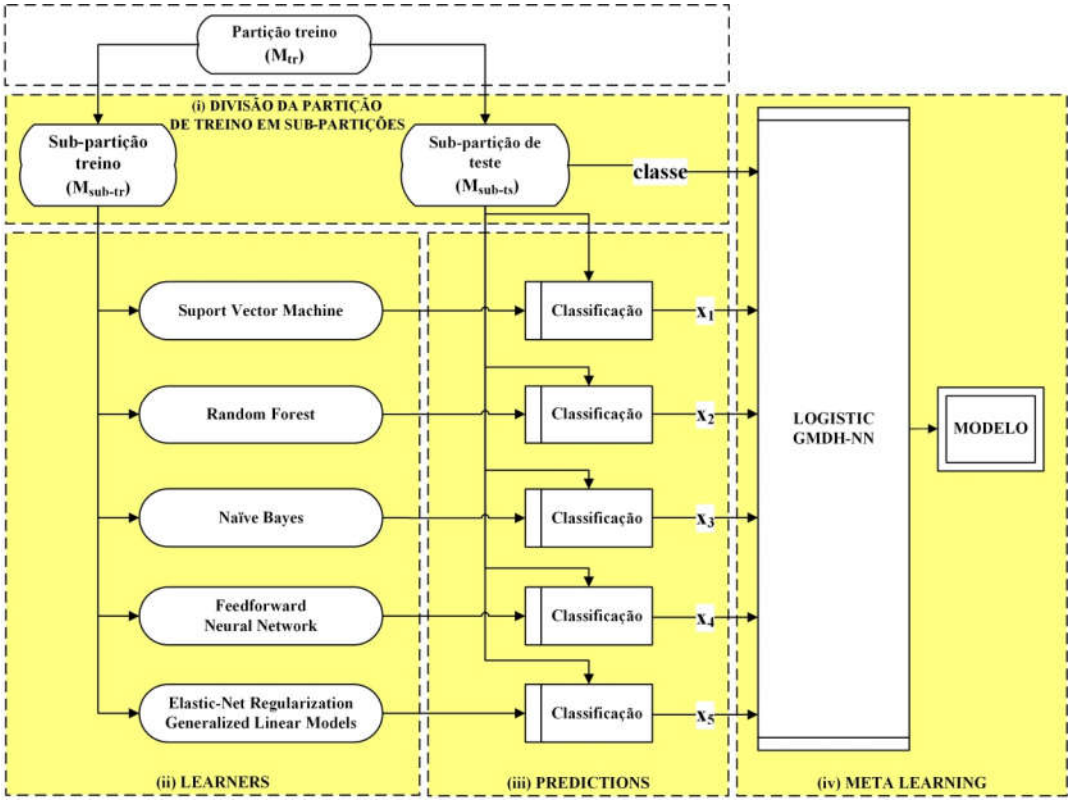

Figura 3.2 - Criação do modelo EL-GMDH-NN

3.2.3.1 *Learners* e *Predictions*

A criação do modelo Ensemble Logistic GMDH-NN inicia com divisão aleatória da partição M_{tr} nas sub-partições $M_{\mathrm{sub-tr}}$ e $M_{\mathrm{sub-ts}}$, na razão de 70-30%. Então a partição $M_{\mathrm{sub-tr}}$, que contém somente as variáveis eleitas e a nova candidata, é utilizada para modelos *learners*: *Suport Vector Machine* (SVM) (WANG, 2005), *Random Forest* (RF) (BREIMAN, 2001), *Naïve Bayes* (NB) (SURHONE *et al.*, 2010), *Regularization Paths for Generalized Linear Models Coordinate Descent* (GLMNET) (FRIEDMAN *et al.*, 2009) e *Single-hidden-layer neural network* (NNET) (RIPLEY, 2007). Os *learners* treinados são utilizados para predição da sub-partição $M_{\mathrm{sub-ts}}$. O *meta learning* Logistic GMDH-NN é criado utilizando como classe de resposta à classe da sub-partição M_{ts} e como

preditores o resultado da predição de cada um dos *learners*. A predição de cada um dos *learners*, SVM, RF, NB, GLMNET e NNET, corresponde a uma variável independente do *meta learning*, representada, respectivamente, por x_1, x_2, x_3, x_4 e x_5.

3.2.3.2 Meta learning - Logistic GMDH-NN

Logistic GMDH-NN é um modelo que integra o método de redes neurais artificiais Feedforward com a heurística de auto-organização GMDH (IVAKHNENKO, 1970). Na primeira camada do modelo de rede neural, cada variável preditiva é representada por um neurônio, então pares de neurônios são conectados via polinômio quadrático Kolmogorov–Gabor, resultando em $n^n - 1$ combinações, sendo n o número de variáveis de entrada (NARIMAN- ZADEH et al., 2002). Em seguida, um critério externo (geralmente erro médio quadrado (MSE)), também conhecido como Criterion for Regularity (CR) é calculado para cada conjunto de testes, de acordo com a equação 3.5 (FERNÁNDEZ; LOZANO, 2010). Conforme os autores, somente os neurônios com o menor valor de CR são mantidos. Em seguida, uma nova camada é criada usando as saídas selecionadas e o processo começa novamente formando pares de entradas. Essas etapas são repetidas para gerar novas camadas até que o critério de erro pare de diminuir. Quando isso ocorre, a saída obtida será a que atingiu o menor valor de critério de erro na camada anterior.

$$CR = \frac{1}{T} \sum_{n=1}^{T} (\widehat{y_n} - y_n)^2 \tag{3.5}$$

onde T é o número de vetores do subconjunto de teste; $\widehat{y_n}$ é o valor predito para a *n*-ésima variável; e y_n é o valor atual da *n*-ésima variável. Os parâmetros da rede são estimados utilizando séries de Volterra (VOLTERRA, 1930; MASOUMNEZHAD *et al.*, 2016). A equação 3.6 demonstra a conexão entre as entradas e saídas do modelo.

$$\hat{y} = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \cdots \tag{3.6}$$

onde $\hat{y}$ é a predição; $\hat{y} \in \{0, 1\}$; m o número de entradas $a, b, c, d, e, \cdots$ são os coeficientes da função; x é o vetor de entrada variáveis e $i, j, k, \cdots$ são os índices das variáveis. A saída $\hat{y}$ é arredondada para zero quando valor for menor que 0,5, caso contrário é assumido o valor 1.

A acurácia do modelo EL-GMDH-NN indica se o processo deve continuar, adicionar uma nova variável ou ser finalizado. Esse passo utiliza a partição M_{ac} para extrair a predição dos *learners* treinados e submeter às predições ao *meta learning* EL-GMDH-NN. O processo é finalizado quando a acurácia atinge o valor 1 ou quando não há mais variáveis candidatas (o que acontecer antes).

3.3 Experimentos, resultados e discussões

3.3.1 Amostras

O método proposto SVD-GMDH foi aplicado sobre nove bancos de dados padronizados. Cinco foram obtidas por Gauchi e Chagnon (2001) e Wold *et al.* (2001) e contêm dados relacionados a aferições de processos industriais, como a produção de nylon (ADPN), emulsão na indústria de papel (GRANU), polimerização em um processo de látex (LATEX), produção de óxido de titânio (OXY) e produção de antibióticos (SPIRA). As variáveis preditoras a serem selecionadas descrevem temperaturas, pressões e concentração de reagentes químicos dos processos analisados. A variável de resposta (classe) indica uma característica do produto, como viscosidade ou pureza, e é definida em dois níveis (conforme ou não conforme), de acordo com especificações de Gauchi e Chagnon (2001) e Wold *et al.* (2001). Os outros quatro bancos analisados foram fornecidos por Ferrão *et al.* (2011) e são relacionados a propriedades físico-químicas do biodiesel e diesel brasileiro, como temperatura de transição do estado líquido para o gasoso (BOILING POINT), velocidade de combustão e compressão para ignição (CETANE NUMBER), temperatura de ignição instantânea em contato com fogo (FLASH POINT) e viscosidade do combustível (VICOSITY). As classes foram definidas com base na mediana da variável resposta. Valores acima da mediana foram considerados conformes e valores abaixo ou iguais à mediana foram considerados como não conformes em relação à qualidade. A tabela 3.1 relaciona o nome de cada banco analisado, natureza

da aplicação, quantidade de variáveis, quantidade de observações, quantidade de observações classificadas como conforme e como não conforme e número de observações em cada uma das partições de treino-checagem de acurácia.

Tabela 3.1 - Bancos de dados de processos industriais e propriedades físico-químicas do Biodiesel e Diesel brasileiro

Banco de dados	Natureza da aplicação	Número total de variáveis	Número total de observações	Conformidade		Número de observações					
						Partição de treino			Partição de teste		
				Sim	Não	60%	75%	90%	40%	25%	10%
ADPN	Produção de nylon	100	71	51	20	43	53	64	28	18	7
GRANU	Emulsão na indústria de papel	78	29	15	14	17	22	26	12	7	3
LATEX	Polimerização em um processo látex	117	262	184	78	157	197	236	105	66	26
OXY	Produção de óxido de titânio	95	25	18	7	15	19	23	10	6	3
SPIRA	Produção de antibióticos	96	145	95	50	87	109	131	58	36	15
BOILING POINT (C°)	Temperatura de transição do estado líquido para o gasoso do Biodiesel e Diesel e brasileiro	401	113	60	53	68	85	102	45	28	11
CETANE NUMBER	Indicador de velocidade de combustão e compressão para ignição do Biodiesel e Diesel e brasileiro	401	113	59	54	68	85	102	45	28	11
FLASH POINT (C°)	Temperatura de ignição instantânea em contato com fogo do Biodiesel e Diesel e brasileiro	1738	85	44	41	51	64	77	34	21	9
VICOSITY	Viscosidade do combustível do Biodiesel e Diesel e brasileiro.	401	116	59	57	70	87	104	46	29	12

3.3.2 Experimentos

Os resultados foram comparados com o estudo de Anzanello *et al.* (2009), o qual implementou um novo índice de importância das variáveis baseado em Mínimos Quadrados Parciais (PLS). No estudo dos autores, a proposta foi comparada com outras abordagens e combinada com técnicas de aprendizagem de máquina, tais como: *k-Nearest Neighbor* (KNN), *Probabilistic Neural Network* (PNN) e *Suport Vector Machine* com regressão PLS. Ademais, para comparar

os resultados do método, foram utilizadas outras duas técnicas: *Logistic Stepwise Regression* (*bidirectional elimination*) e Logistic GMDH-NN (KONDO *et al.*, 1999). Os experimentos foram conduzidos com porções de treino-checagem de acurácia gerados aleatoriamente de forma estratificada e proporcional por classe nos conjuntos 60-40%, 75-25% e 90-10%. No estudo de Anzanello *et al.* (2009) não estão incluídos experimentos com partições 60%-40% e 75-25%, dessa forma há lacunas de comparação para essas partições. Foram executadas de 200 a 10.000 simulações independentes, no qual o critério de parada foi a convergência da média de acurácia das últimas dez iterações. Para implementar a proposta foi utilizada a linguagem de programação R (versão 3.6.2) (R CORE TEAM, 2019), com o pacote GMDH2 (versão 1.4) (DAG *et al.*, 2019).

3.3.3 Resultados e discussões

Foram apurados como resultados o percentual médio de variáveis retidas (% variáveis retidas), percentual médio de acurácia (% acurácia) e percentual médio do coeficiente de variação da acurácia (% CF) em percentual.. A tabela 3.2 demonstra o desempenho médio do SVD-GMDH para cada banco de dados em cada partição de treino-checagem de acurácia.

Tabela 3.2 - Desempenho médio do SVD-GMDH

Banco de dados	Porções do banco de dados % treino–% checagem da acurácia					
	60-40%		75-25%		90-10%	
	% Acurácia (% CF)	% Variáveis retidas	% Acurácia (%CF)	% Variáveis retidas	% Acurácia (%CF)	% Variáveis retidas
ADPN	91,95 (5,06)	6,45	93,25 (4,87)	5,61	96,64 (5,75)	3,81
GRANU	88,3 (5,1)	2,34	90,28 (5,86)	1,99	94,2 (5,94)	1,37
LATEX	83,98 (5,34)	2,08	85,34 (6,28)	1,88	91,97 (6,95)	1,41
OXY	99,63 (0,98)	0,55	99,72 (1,09)	0,43	99,8 (1,47)	0,33
SPIRA	90,42 (4,24)	2,31	93,06 (4,47)	2,05	97,62 (4,79)	1,59
BOILING POINT	89,7 (6,3)	6,46	93,35 (3,82)	6,05	100 (0)	4,1
CETANE NUMBER	86,3 (3,58)	8,39	87,47 (4,48)	7,5	90,93 (5,86)	5,45
FLASH POINT	100 (0)	3,99	100 (0)	3,28	100 (0)	2,72
VICOSITY	83,24 (4,74)	8,54	86,02 (5,27)	7,39	91,98 (8,32)	5,62

O melhor resultado foi obtido na amostra FLASH POINT, para o qual SVD-GMDH gerou 100% de acurácia, com percentual médio de variáveis variando de

3,99%, para porção 60-40%, à 2,72%, na porção 90-10%. Na amostra BOILING POINT, a porção 90-10% também atingiu 100% de acurácia utilizando somente 4,1% das variáveis. Também destaca-se o resultado obtido na amostra OXY, que mesmo possuindo somente 25 observações, alcançou 100% de acurácia na porção 90-10%. Em geral, a acurácia foi superior a 90% em 63% dos experimentos, sendo que o menor resultado foi de 83,24% de acurácia para a amostra VISCOSITY (porção 60-40%). Da mesma forma, os resultados mostraram-se positivos em relação à variabilidade da acurácia, com 85% dos apontamentos apresentando valores inferiores a 6%. Na contramão, a variabilidade mais expressiva foi de 8,32% para porção 90-10% da amostra VICOSITY. Com relação ao percentual de variáveis retidas, 74% dos resultados foram formados por subconjuntos com menos de 6% das variáveis. O resultado mais impactante foi com a amostra de OXY, no qual, em média, somente 0,33% das variáveis foram necessárias para formar um subconjunto de preditores que gerasse, em média, 99,8% de acurácia. Outro resultado destacável ocorreu com o banco de dados FLASH POINT, que possui a mais alta dimensionalidade das bases analisadas (1738 variáveis). O método utilizou, em média, somente 3,9%, para porção 60-40%, 3,28%, na porção 75-25% e 2,72% das variáveis, com a porção 90-10%, para compor subconjuntos de preditores que produzissem classificações perfeitas. A tabela 3.3 traz o desempenho do SVD-GMDH comparado a outras técnicas.

Tabela 3.3 - Comparação de desempenho do SVD-GMDH

Banco de dados	% Treino	Anzanello *et al.* (2009)		Regressão Logística S*tepwise*		Logistic GMDH-NN		SVD-GMDH	
		% Acurácia	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas
ADPN	60			69,82 (13,43)	6,76	79,54 (9,29)	8,06	91,95 (5,06)	6,45
	75			73,14 (14,46)	8,11	82,23 (11,14)	7,93	93,25 (4,87)	5,61
	90	87,00	8,00	74,10 (19,50)	8,62	80,82 (17,53)	8,46	**96,64 (5,75)**	**3,81**
GRANU	60			62,12 (7,00)	7,11	73,13 (10,66)	2,31	88,3 (5,1)	2,34
	75			55,55 (11,85)	5,84	74,42 (11,84)	2,23	90,28 (5,86)	1,99
	90	87,00	7,70	66,66 (0)	7,58	71,29 (22,97)	2,15	**94,2 (5,94)**	**1,37**
LATEX	60			83,83 (5,13)	18,62	70,89 (10,03)	2,17	83,98 (5,34)	2,08
	75			85,68 (4,46)	21,80	71,57 (10,93)	2,03	85,34 (6,28)	1,88
	90	83,00	18,5	81,40 (4,59)	20,09	70,89 (17,14)	2	**91,97 (6,95)**	**1,41**
OXY	60			66,66 (24,74)	2,74	98,76 (1,57)	0,63	99,63 (0,98)	0,55

	75			76,00 (7,20)	1,75	97,64 (2,37)	0,78	99,72 (1,09)	0,43
	90	73,00	6,30	83,33 (23,09)	2,11	**100 (0)**	**0,62**	99,8 (1,47)	0,33
SPIRA	60			71,92 (8,50)	16,73	70,09 (7,61)	2,36	90,42 (4,24)	2,31
	75			58,80 (15,96)	17,96	77,93 (8,66)	2,72	93,06 (4,47)	2,05
	90	90,00	4,20	61,60 (21,70)	26,42	82,29 (10,38)	2,03	**97,62 (4,79)**	**1,59**
BOILING POINT	60			70,48 (7,19)	3,45	70,24 (11,12)	10,62	89,7 (6,3)	6,46
	75			82,75 (0)	4,12	68,27 (9,5)	8,38	93,35 (8,82)	6,05
	90			87,74 (4,81)	4,25	83,33 (21,91)	3,85	**100 (0)**	**4,1**
CETANE NUMBER	60			64,49 (4,24)	3,99	78,1 (6,33)	5,98	86,3 (3,58)	8,39
	75			65,51 (0)	3,62	76,68 (0,99)	8,07	87,47 (4,48)	7,5
	90			70,83 (17,85)	5,49	81,7 (4,66)	7,29	**90,93 (5,86)**	**5,45**
FLASH POINT	60			55,88 (31,58)	100	100 (0)	9,61	100 (0)	3,99
	75			86,36 (10,52)	100	85,71 (0)	8,98	100 (0)	3,28
	90			77,77 (14,28)	100	82,46 (20,74)	10,03	**100 (0)**	**2,72**
VISCOSITY	60			80,85 (5,41)	100	74,34 (4,77)	8,88	83,24 (4,74)	8,54
	75			55,17 (6,25)	100	77,3 (5,42)	7,85	86,02 (5,27)	7,39
	90			58,33 (14,29)	100	74,29 (9,22)	7,44	**91,98 (8,32)**	**5,62**

Em todas as comparações o desempenho do SVD-GMDH mostrou-se superior com relação a acurácia, com exceção da base OXY, no qual a porção 90-10% obteve 99,8%, enquanto que no método Logistic GMDH-NN as classificações foram perfeitas. No entanto, também é possível constatar que foram necessárias, aproximadamente, o dobro de variáveis, com 0,62% contra 0,33% do SVD-GMDH. Esse resultado, em especial, indica que o índice de importância de variáveis calculado a partir do SVD desempenha um papel importante na orientação da composição dos subconjuntos. Sobre essa ótica, e ponderando que a diferença de acurácia não é extremamente pequena (0,02%), pode-se considerar que o SVD-GMDH também mostra-se superior nesse resultado.

No conjunto de dados relacionados aos processos industriais (ADPN, GRANU, LATEX, OXY e SPIRA), verificou-se que o segundo método com melhor desempenho foi Anzanello *et al.* (2009). Todavia, enfatiza-se que o SVD-GMDH demonstrou desempenho superior, mesmo ao comparar a partição 60-40% (com acurácia de 90,86%) com 90-10% de Anzanello *et al.* (2009), que registrou 84% de acurácia utilizando uma porção maior da amostra para treinamento. Dentre

todos os métodos, para todas as bases de dados, o pior desempenho foi obtido pelo método Regressão Logística *Stepwise*.

O processo de seleção de variáveis tem como principal objetivo a definição de um subconjunto ótimo ou subótimo de preditores, como relata Guyon e Elisseeff (2003). Porém, um grande benefício que pode ser obtido, principalmente em processos industriais, é a identificação das variáveis que não são relevantes. Essa informação pode nutrir análises para otimizar as aferições de qualidade, aumentando a eficácia do processo e reduzindo custos de produção. Tomando como base essa abordagem, a tabela 3.4 descreve, para as bases de dados relacionadas a processos industriais, o percentual de variáveis que não esteve presente em nenhum subconjunto de variáveis selecionadas nas simulações.

Tabela 3.4 - Variáveis não selecionadas

Banco de dados	% de variáveis não selecionadas		
	Porção 60-40%	Porção 75-25%	Porção 90-10%
ADPN	15,00	19,00	39,00
GRANU	26,92	33,33	41,03
LATEX	8,55	11,11	24,79
OXY	51,58	55,79	60,00
SPIRA	2,08	5,21	10,42

Como destaque, a tabela 3.4 demonstra que, mesmo na porção 60-40%, que fornece a menor fração de dados para treino, no banco de dados OXY 51,58% das variáveis não foram utilizadas para compor subconjuntos de preditores. Quando foi utilizada a porção que oferece mais dados ao treino, 90-10%, a identificação das variáveis mais importantes torna-se mais precisa, o que levou ao apontamento de mais variáveis a serem descartadas como preditores. No resultado menos expressivo, 2,08% das variáveis não são relevantes, um resultado que, apesar de ser mais modesto, caracteriza uma informação importante para otimização do processo de qualidade. Como os preditores são relacionados a uma classe que indica a conformidade de qualidade, é possível utilizar essa informação para investigar se as variáveis tidas como não importantes estão sendo utilizadas em outras etapas do processo produtivo, e, se não estiverem sendo utilizadas, podem ser descartadas do processo de aferição da qualidade, tornando a atividade mais eficaz e reduzindo custos.

3.4 Conclusões

O processo de seleção de variáveis tem papel fundamental para criação de modelos preditivos mais acurados e parcimoniosos, como relatam Guyon e Elisseeff (2003). O modelo proposto, SVD-GMDH, apresenta uma inovação ao implementar um novo ranking de importância das variáveis, baseado em SVD, que elenca as variáveis com base no poder de discriminação das observações. O ranking orienta, de forma iterativa, a inclusão das variáveis em subconjuntos de preditores utilizados no método de classificação Ensamble Logistic GMDH. Esse método utiliza a estratégia *stacking ensemble* para combinar múltiplos algoritmos de aprendizagem de máquina. O SVD-GMDH mostrou-se eficiente para definir o subconjunto ótimo ou subótimo de predição quando comparado a outros modelos para seleção de variáveis, como Anzanello et al. (2009), Regressão Logística *Stepwise* e Logistic GMDH-NN.

Um dos primeiros benefícios que pode ser obtido com a aplicação do SVD-GMDH é a melhora da acurácia na classificação de qualidade das amostras entre conforme e não conforme. No conjunto de dados relacionados ao Biodiesel e Diesel brasileiros, classificações incorretas podem levar ao mercado um produto prejudicial ao consumidor. Como consequência, um produto não conforme classificado de forma errada como conforme pode gerar desde um maior consumo de combustível até depreciação acelerada ou falha no motor. Para compreender o impacto de um produto não conforme no mercado, deve-se considerar que, somente no Brasil, há 2 milhões de caminhões ativos, conforme dados da Revista Exame (2019). Ademais, outro fato importante, mas visto de forma indireta, está na imagem da marca do Biodiesel e Diesel brasileiro, que pode, inclusive, ecoar negativamente na esfera internacional.

Com relação aos processos industriais, além da melhora na precisão da classificação, o SVD-GMDH, ao reduzir os subconjuntos de preditores importantes, também aponta um conjunto maior de variáveis não importantes. Nesse sentido, o método coopera para gerar informações que podem ser utilizadas na análise do processo de aferição da qualidade, tendo como consequências positivas a maior eficácia e redução de custos.

Sobre a estrutura do método, um aspecto de destaque é a utilização do algoritmo indutivo GMDH. A heurística GMDH processa automaticamente inter-relações nos dados, selecionar uma estrutura ideal de modelo ou rede e aumentar a precisão dos algoritmos existentes. O GMDH aplicado em redes neurais consegue definir automaticamente a quantidade de camadas, neurônios e quais são as variáveis de entrada necessárias (excluindo as desnecessárias), o que elimina a necessidade de definir os parâmetros do modelo arbitrariamente.

Cabe ressaltar que, mesmo apresentando os melhores resultados, o método torna-se pouco parcimonioso ao envolver diversos algoritmos. Nesse contexto, caberiam estudos futuros sobre a integração do SVD com o GMDH no sentido de melhorar a heurística do GMDH, reduzindo a complexidade do método.

3.5 Referências

AHMDI, M.H.; Ramezanizadeh, M.; Alhuyi Nazari, M.; Kheradmand, S.; Shamshirband, S. Carbon Dioxide Emission Prediction of Four CIS Countries by Applying a Correlation and GMDH Artificial Neural Network. *Preprints 2019*, 2019. ISSN 2019060227. doi 10.20944/preprints201906.0227.v1. <https://www.preprints.org/manuscript/201906.0227/v1>

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. Chemometrics and Intelligent Laboratory Systems, v. 97, n. 2, p. 111–117, 2009. ISSN 0169-7439.

ANZANELLO, M. J. *et al.* PLS-DA wavenumber selection for the categorization of medicine samples based on multiple criteria. Forensic Science International, v. 242, n. Supplement C, p. 111–116, 2014. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073814002709>.

ANZANELLO, M. J. *et al.* A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. Journal of Pharmaceutical and Biomedical Analysis, v. 83, n. Supplement C, p. 209–214, 2013. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708513001970>.

BATTITI, R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, v. 5, n. 4, p. 537–550, 1994. ISSN 10459227. Disponível em: <http://ieeexplore.ieee.org/document/298224/>.

BRAGA, A. L. S.; LLANOS, C. H.; COELHO, L. dos S. Comparing artificial neural network implementations for prediction and modeling of dynamical systems. In: . [s.n.], 2012. v. 5, p. 602–609. *ABCM Symposium Series in Mechatronics*, Section III – Emerging Technologies and AI Applications. Disponível em:

<http://abcm.org.br/symposium-series/REFERÊNCIAS 39 SSM_Vol5/Section_III_Emerging_Technologies_and_AI_Applications/02451.pdf>.

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Institui o Código Civil. Presidência da República Casa Civil: Subchefia para Assuntos Jurídicos, Brasília, DF, ano 139, n. 8, p. 1-74, 1990. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l8078.html. Acesso em: 14 jan. 2020.

BREIMAN, L. Random Forests. Machine Learning, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125. Disponível em: <http://link.springer.com/10.1023/A: 1010933404324>.

BRILL, F.; BROWN, D.; MARTIN, W. Fast generic selection of features for neural network classifiers. IEEE Transactions on Neural Networks, v. 3, n. 2, p. 324–328, 1992. ISSN 10459227. Disponível em: <http://ieeexplore.ieee.org/document/125874/>.

BROMLEY, J.; SACKINGER, E. Neural-Network and k-Nearest-neighbor Classifiers. AT&T Bell Laboratories, aug 1991. Disponível em: <http://oro.open.ac.uk/35666/>.

CASTELLANO, G.; FANELLI, A. M. Variable selection using neural-network models. Neurocomputing, Elsevier, v. 31, n. 1-4, p. 1–13, mar 2000. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231299001460>.

CIBAS, T. *et al.* Variable selection with optimal cell damage. In: ICANN '94. London: Springer London, 1994. p. 727–730. Disponível em: <http://link.springer.com/10.1007/ 978-1-4471-2097-1{_}>.

DAG, O.; KARABULUT, E.; ALPAR, R. GMDH2: Binary Classification via GMDH-Type Neural Network Algorithms—R Package and Web-Based Tool. [S.l.], 2019. v. 12, n. 2, 649 p. Disponível em: <https://cran.r-project.org/package=GMDH2>.

DEOGUN, J. S. *et al.* Feature selection and effective classifiers. Journal of the American Society for Information Science, v. 49, n. 5, p. 423–434, 1998. ISSN 00028231.

DEOGUN, J. S. *et al.* Data Mining: Research Trends, Challenges, and Applications. IN ROUGHS SETS AND DATA MINING: ANALYSIS OF IMPRECISE DATA, p. 9–45, 1997. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.337>.

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. Journal of Chemometrics, John Wiley and Sons, v. 24, n. 11-12, p. 779–789, 2010. ISSN 0886-9383,1099-128X.

EXAME, O Brasil tem caminhões em excesso – e terá ainda mais. *Revista Exame*. 2019. Disponível em <https://exame.abril.com.br/economia/o-brasil-tem-caminhoes-em-excesso-e-tera-ainda-mais/>.

FERNÁNDEZ, F. H.; LOZANO, F. H. Gmdh algorithm implemented in the intelligent identification of a bioprocess. In: ABCM Symposium Series in Mechatronics.–2010.–4.–P. [S.l.: s.n.], 2010. p. 278–287.

FERRÃO, M. F. *et al.* Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. Fuel, v. 90, n. 2, p. 701–706, feb 2011. ISSN 00162361.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. [S.l.], 2009. Disponível em: <https://web.stanford.edu/{~}hastie/Papers/glmnet.p>.

FUKUNAGA, K. Introduction to statistical pattern recognition. [S.l.]: Academic Press, 1990. 591 p. ISBN 9780080478654.

GAUCHI, J.-P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics and Intelligent Laboratory Systems, v. 58, n. 2, p. 171–193, oct 2001. ISSN 01697439. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743901001587>.

GOLUB, G.; KAHAN, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis, Society for Industrial and Applied Mathematics, v. 2, n. 2, p. 205–224, 1965. ISSN 0887459X. Disponível em: <http://www.jstor.org/stable/2949777>.

HUANG, Q. *et al.* Exploiting local coherent patterns for unsupervised feature ranking. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 41, n. 6, p. 1471–1482, 2011. ISSN 1083-4419.

HUNTER, A. Feature Selection Using Probabilistic Neural Networks. Neural Computing & Applications, Springer-Verlag London Limited, v. 9, n. 2, p. 124–132, jul 2000. ISSN 0941-0643. Disponível em: <http://link.springer.com/10.1007/s005210070023>.

IVAKHNENKO, A. G. Heuristic self-organization in problems of engineering cybernetics. Automatica, Elsevier, v. 6, n. 2, p. 207–219, 1970.

JAIN, A.; ZONGKER, D. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 19, n. 2, p. 153–158, 1997. ISSN 01628828. Disponível em: <http://ieeexplore.ieee.org/document/574797/>.

JENSEN, R.; SHEN, Q. Computational intelligence and feature selection : rough and fuzzy approaches. [S.l.]: IEEE Press, 2008. 339 p. ISBN 9780470377888.

KHAIRE, U. M.; DHANALAKSHMI, R. Stability of feature selection algorithm: A review. [S.l.]: King Saud bin Abdulaziz University, 2019.

KIRA, K.; RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, 1992. (AAAI'92), p. 129–134. ISBN 0-262-51063-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1867135.1867155>.

KONDO, T. GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem. In: Proceedings of the 37th SICE Annual Conference. International Session Papers. Soc. Instrum.

Control Eng, 1998. p. 1143–1148. Disponível em: <http://ieeexplore.ieee.org/document/742993/>.

KONDO, T.; PANDYA, A.; ZURADA, J. Logistic GMDH-type neural networks and their application to the identification of the X-ray film characteristic curve. In: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028). IEEE, 1999. v. 1, p. 437–442. ISBN 0-7803-5731-0. Disponível em: <http://ieeexplore.ieee.org/document/814131/>.

LI, H. *et al.* Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine. Applied Soft Computing Journal, v. 19, p. 57–67, 2014. ISSN 15684946.

LIU, H.; MOTODA, H. Feature extraction, construction and selection: A data mining perspective. [S.l.]: Springer Science & Business Media, 1998. v. 453.

MAO, J.; JAIN, A. K. Artificial neural networks for feature extraction and multivariate data projection. IEEE Transactions on Neural Networks, v. 6, n. 2, p. 296–317, 1995. ISSN 10459227. Disponível em: <http://ieeexplore.ieee.org/document/363467/>.

MASOUMNEZHAD, M.; JAMALI, A.; NARIMAN-ZADEH, N. Robust GMDH-type neural network with unscented Kalman filter for non-linear systems. Transactions of the Institute of Measurement and Control, SAGE Publications Sage UK: London, England, v. 38, n. 8, p. 992–1003, 2016.

MATHEUS, C. J.; Piatetsky Shapiro, G.; CHAN, P. K. Systems for Knowledge Discovery in Databases. IEEE Transactions on Knowledge and Data Engineering, v. 5, n. 6, p. 903–913, 1993. ISSN 10414347.

MING, L.; ZHAO, J. Feature selection for chemical process fault diagnosis by artificial immune systems. Chinese Journal of Chemical Engineering, Chemical Industry Press, v. 26, n. 8, p. 1599–1604, 2018. ISSN 10049541.

MINING, D. Elements of Machine Learning. Morgan Kaufmann, 2009. v. 27. 83–85 p. ISSN 03436993. ISBN 9780387848570. Disponível em: <http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf>.

MOOSAVIAN, A. *et al.* Comparison of Two Classifiers; K-Nearest Neighbor and Artificial Neural Network, for Fault Diagnosis on a Main Engine Journal-Bearing. Shock and Vibration, v. 20, n. 2, p. 263–272, 2013. ISSN 1070-9622. Disponível em: <http://www.hindawi.com/journals/sv/2013/360236/>.

NARIMAN-ZADEH, N. *et al.* Modelling of explosive cutting process of plates using GMDH-type neural network and Decomposição de Valores Singulares. Journal of Materials Processing Technology, 2002. ISSN 09240136.

PASQUALINI, M. *Experience Economy, Brands and Leadership: Reframe Brands, Marketing and Business*. [S.l.]: Amazon Digital Services LLC - Kdp Print Us, 2019. Shared Perspective Series. ISBN 9781094698755.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2019. Disponível em: <https://www.r-project.org/>.

RIPLEY, B. D. Pattern recognition and neural networks. [S.l.]: Cambridge University Press, 2007. 403 p. ISBN 0521717701.

ROKACH, L. Pattern classification using ensemble methods. [S.l.]: World Scientific Pub. Co, 2010. 225 p. ISBN 9789814271073.

SKILLICORN, D. B. Understanding complex datasets : data mining with matrix decompositions. [S.l.]: Chapman & Hall/CRC Press, 2007. 236 p. ISBN 1584888326.

SURHONE, L. M.; TIMPLEDON, M. T.; MARSEKEN, S. F. Naive Bayes Classifier: Classifier (mathematics), Bayes' Theorem, Probability Theory, Bayesian Inference, Bayesian Probability, Empirical Bayes Method, Statistics, Conditional Probability. [S.l.]: Betascript Publishing, 2010. ISBN 9786130334468.

VOLTERRA, V. Theory of Functionals and of Integral and Integro-differential Equations. [S.l.: s.n.], 1930. ISBN 0486442845.

WALL, M. E.; RECHTSTEINER, A.; ROCHA, L. M. Decomposição de Valores Singulares and Análise de Componentes Principais. In: . A Practical Approach to Microarray Data Analysis. Boston, MA: Springer US, 2003. p. 91–109. ISBN 978-0-306-47815-4. Disponível em: <https://doi.org/10.1007/0-306-47815-3_5>.

WANG, L. Support vector machines : theory and applications. [S.l.]: Springer, 2005. 431 p. ISBN 9783540243885.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, Elsevier, v. 58, n. 2, p. 109–130, oct 2001. ISSN 0169-7439. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0169743901001551>.

XIAO, J.; HE, C. Adaptive selection of classifier ensemble based on gmdh. In: 2008 International Seminar on Future Information Technology and Management Engineering. IEEE, 2008. p. 61–64. ISBN 978-0-7695-3480-0. Disponível em: <http://ieeexplore.ieee.org/document/4746442/>.

XIAO, J.; HE, C.; WANG, S. A classifier ensemble model based on gmdh-type neural network for customer targeting. In: . [s.n.], 2014. p. 259–269. Disponível em: <http://link.springer.com/10.1007/978-3-642-40078-0_22>.

4 CONSIDERAÇÕES FINAIS

O processo de definição do subconjunto ótimo ou subótimo de variáveis independentes contribui diretamente para o desempenho dos algoritmos preditivos, da mesma forma que impacta diretamente em ganho computacional. Ademais, a identificação das variáveis descartáveis, como comprimentos de onda (COs) ou aferições do processo de produção, pode orientar o processo de coleta das variáveis, direcionando o aperfeiçoamento do processo ao focar nas variáveis relevantes. A presente dissertação apresentou dois novos métodos para seleção de variáveis que tomam como base o modelo proposto por Anzanello *et al*. (2013), diferenciando-se do mesmo pela aplicação da abordagem *forward* na adição de variáveis candidatas e pelo uso de um novo ranking para quantificação da importância das variáveis baseado em SVD.

Tradicionalmente, métodos de decomposição matricial são aplicados para minimizar a dimensionalidade de um banco de dados, permitindo reduzir grande parte da variabilidade em poucas dimensões, como componentes principais (PCA) ou valores singulares (SVD). Entretanto, para chegar às dimensões, ainda é necessário submeter todas as variáveis originais. Nesse aspecto, é importante observar que não está sendo feita uma seleção de variáveis originais, mas sim uma transformação, que precisará ser reexecutada toda vez que houver uma atualização da base de dados. Contudo, esses métodos capturam a variabilidade dos dados de forma a permitir a apuração de um ranking de importância sobre as variáveis preditivas sem utilizar a classe, ou seja, um ranking não supervisionado. Essa aparente desvantagem traz como benefício a identificação das variáveis com maior poder de discriminação das observações de uma amostra. Em especial, o uso do SVD apresenta duas grandes vantagens frente ao PCA. Primeiramente, por utilizar matrizes retangulares, ele dispensa transformações para gerar uma matriz quadrada, como cálculos de covariância ou correlação. Assim, o SVD não depende de quão bem um dado foi representado pela transformação. A outra vantagem é que o SVD analisa de forma conjunta os atributos, e o PCA, por outro lado, processa a análise de forma independente. Isso confere maior sensibilidade na distinção entre sinal e ruído. Essas vantagens mostraram-se benéficas na identificação dos comprimentos de

onda mais relevantes, como foi observado nos experimentos com amostras de Cialis® e Viagra®, apresentados no artigo 1.

A falsificação de medicamentos é um problema que vai além do prejuízo financeiro para os laboratórios e governo (impostos), pois representa risco à saúde pública. Nesse cenário, não há garantias sobre o controle da qualidade e dosagem dos ingredientes, assim como sobre todo o processo de fabricação. Além do mais, o problema tem sido potencializado pela redução do custo e fácil acesso aos equipamentos e ingredientes de fabricação. Todavia, a identificação de um medicamento falsificado pode ser automatizada com algoritmos de aprendizagem de máquina, como o KNN, mas seu desempenho depende fundamentalmente da qualidade dos preditores. Nesse sentido, a proposta da dissertação, descrita no artigo 1, demonstrou desempenho superior ao método de Anzanello *et* al. (2013), tido como referência. Em amostras de Cialis® e Viagra® foram identificadas bandas específicas, contendo os comprimentos de onda que mais discriminavam as observações. Como benefício, além de melhorar o desempenho do classificador em relação à ganho computacional e aumento da acurácia, a identificação de comprimentos de onda em bandas específicas permite o uso de dispositivos especializados em coletar as respectivas faixas de comprimentos de onda. Essa vantagem contribui para o uso de equipamentos mais acessíveis, em relação ao custo, e portáveis, para operações em campo de órgãos fiscalizadores.

Com base nos resultados obtidos no primeiro artigo, no segundo artigo o método foi atualizado com a abordagem de aprendizagem de máquina Ensemble Logistic GMDH-NN (EL-GMDH-NN), em substituição ao KNN. O novo método, batizado como SVD-GMDH, foi aplicado sobre nove bancos de dados, sendo cinco relacionados a processos industriais e quatro relativos a propriedades físico-químicas do Biodiesel e Diesel brasileiros. Todos os bancos de dados são representados por uma classe binária que indica conformidade ou não conformidade em relação a qualidade. Comparando os resultados com Regressão Logística *Stepwise*, Logistic GMDH-NN e o modelo de Anzanello *et al*. (2009), o método proposto demonstrou desempenho superior, com maior acurácia em menores subconjuntos de variáveis. Contudo, há um aumento de complexidade na proposta por parte do EL-GMDH-NN. Isso ocorre em

decorrência da estratégia *stacking ensemble*, no qual diversas técnicas de aprendizagem de máquina são combinadas e exercem papel de *learners,* e da escolha do *meta learning* Logistic GMDH-NN.

O método de aprendizagem de máquina Logistic GMDH-NN é uma rede neural, direcionada à predição logística, que foi otimizada pela heurística *Group Method of Data Handling* (GMDH). Essa otimização confere à rede auto-organização estrutural ao definir automaticamente a quantidade de neurónios, camadas e variáveis de entrada. Em experimentos para comparação no segundo artigo, o método foi aplicado diretamente sobre os dados, resultando no desempenho mais próximo da proposta SVD-GMDH. Ainda assim, o risco de produzir resultados incorretos na classificação de qualidade (entre conforme ou não conforme) compensa o aumento de complexidade. Amostras de Biodiesel e Diesel classificadas como conformes incorretamente irão refletir nos consumidores como aumento do consumo de combustível, depreciação acelerada do motor, possibilidade de defeitos e, inclusive, desgaste da imagem da marca. Da mesma maneira, classificações incorretas sobre os processos industriais serão refratadas desde o ambiente de produção até o consumidor final.

A evolução da proposta da dissertação parte do artigo 2 e aponta para dois possíveis tópicos a serem abordados em implementações futuras. Em primeira instância está o desafio de reduzir a complexidade. A heurística do GMDH contempla um processo de seleção de variáveis com potencial de ser aperfeiçoado caso seja fundamentado no ranking de importância de variáveis extraído a partir do SVD. Outro aspecto que pode ser abordado futuramente é a apuração da incerteza em relação ao ranking de importância. No lugar de atribuir escores de importância para cada variável, poderiam ser calculadas distribuições de probabilidade. Sobre essa ótica, seria possível identificar empates técnicos, no qual, em decorrência da amostra, não haveria evidências suficientes para apontar variáveis mais ou menos importantes. Utilizando esse diagnóstico seria possível considerar o método utilizado para ranquear a importância das variáveis inadequado para o conjunto de dados ou, até mesmo, concluir que a amostra não apresenta evidência suficientes para que seja atribuído um índice de importância sobre as mesmas.