UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Juliana de Abreu Fontes

ABORDAGENS DE SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO E REGRESSÃO EM DADOS ESPECTRAIS PARA CONTROLE DA QUALIDADE

Porto Alegre
2020

Juliana de Abreu Fontes

Abordagens de seleção de variáveis para classificação e regressão em dados espectrais para controle da qualidade

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

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

Porto Alegre

2020

Juliana de Abreu Fontes

Abordagens de seleção de variáveis para classificação e regressão em dados espectrais para controle da qualidade

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

_____

Prof. Michel José Anzanello, *Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. Alejandro Germán Frank, *Ph.D.*
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Flávio Sanson Fogliatto, *Ph.D*. (PPGEP/UFRGS)

Professor Marcelo Xavier Guterres, Dr. (ITA)

ProfessoraVera Lucia Duarte Ferreira, Dr. (UNIPAMPA)

Fontes, Juliana.***Abordagens de seleção de variáveis para classificação e regressão em dados espectrais para controle da qualidade***, 2020. Dissertação (Mestrado em Engenharia) – Universidade Ferderal do Rio Grande do Sul, Brasil.

RESUMO

Técnicas espectroscópicas têm sido amplamente empregadas na resolução de problemas referentes à verificação de autenticidade e padrões de qualidade de produtos. No entanto, tais técnicas tendem a gerar um elevado número de variáveis (comprimentos de onda – COs) ruidosas e altamente correlacionadas, reforçando a importância do uso de técnicas que permitam remover as variáveis não informativas e garantir a construção de modelos consistentes de classificação e predição, diminuindo tanto o risco de inferências como o custo computacional. Esta dissertação propõe sistemáticas para seleção de COs com vistas à classificação de produtos e predição de propriedades químicas.Os métodos aqui propostos mesclam diferentes técnicas de aprendizado de máquina para definir os subconjuntos de variáveis mais importantes para as predições. Para tanto, inicialmente faz-se uma investigação sobre métodos de seleção de variáveis por meio de uma pesquisa bibliográfica. Em seguida, visando predizer propriedades químicas das amostras de misturas de combustível, faz-se uso de conceitos químicos advindos da Lei de Lambert-Beer para a geração de índices de importância de variáveis; subconjuntos de variáveis são então construídos por meio de uma abordagem direta com redes neurais artificiais (*Artificial Neural Networks*– ANN). Por fim, utiliza-se o método estatístico qui-quadrado (χ^2) combinado com a ferramenta de classificação floresta aleatória (*Random Forest*– RF) para selecionar o subconjunto de COs que resulte na maior acurácia média com vistas à classificação de amostras de alimentos e drogas (lícitas e ilícitas) em autênticas ou não-autênticas, segundo sua identidade e/ou origem.A aplicação dos métodos propostosem bancos reais possibilitou predições mais robustas, bem como redução do número de variáveis retidas nos modelos.

Palavras-chave: Autenticidade de Produtos,Seleção de COs, Classificação, Regressão, Espectroscopia.

Fontes, Juliana.***Feature selection approaches for classification and regression in spectral data for quality control***, 2020. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

ABSTRACT

Spectroscopic techniques have been widely used in solving problems related to authenticity verification and product quality standards. However, the result of these techniques tends to generate a high number of variables (wavelengths) noisy and highly correlated, reinforcing the importance of using techniques that allow removing non-informative variables and ensure the construction of consistent classification and prediction models, reducing both the risk of inferences and computational cost. This dissertation purposes systematics for the selection wavelengths in order to classify products and predict chemical properties. The methods proposed here merge different machine learning techniques to define the subsets of wavelengths most important to predictions. Therefore, an investigation is initially carried out on methods of variable selection through a bibliographic research. Then, in order to predict chemical properties of fuel mixture samples, chemical concepts from the Lambert-Beer law are used for the generation of variable importance indexes; subsets of variables are then constructed through a direct approach with artificial neural networks (ANN). Finally, the chi-square statistical method (χ^2)combined with the random forest classification tool (RF) is used to select the subset of wavelengths that results in greater average accuracy aiming to classify food and drug samples (lawful and illicit), in authentic or not authentic, according to their identity and/ or origin. The application of the methods proposed in real banks allowed the realization of more robust predictions, as well as the reduction of the number of variables retained in the models.

Keywords: Product Authenticity, Wavelength Selection, Classification, Regression, Spectroscopy.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

NIR	*Near-Infrared Spectroscopy*
FTIR	*Fourier-transform infrared*
HATR – FTIR	*Horizontal attenuated total reflectance Fourier transform infrared spectroscopy*
COs	Comprimentos de onda
VR	Variável Resposta
LASSO	*Least Absolute Shrinkage and Selection Operator*
CFS	*Correlation Based Feature Selection*
WBS	*Weighted Bootstrap Sampling*
LS	*Least Squares*
MIC	*Maximal Information Coeficiente*
QP	*Quadratic Programming*
SVM	*Support Vector Machine*
RFE	*Recursive Feature Elimination*
KW	*Kruskal-Wallis*
LDA	*Linear Discriminant Analysis*
χ^2	*Chi-squared*
PLS	*Partial least squares*
SFS	*Sequential Forward Feature Selection*
SBS	*Sequential Backward Feature Selection*
SFFS	*Sequential Forward Floating Selection*
SBFS	*Sequential Backward Floating Selection*
RelieF	*RELevance In Estimating Features*
ID3	*Iterative Dichotomiser 3*
CART	*Classification and Regression Tree*
iPLS	*Interval Partial Least Squares*
biPLS	*Backward Interval Partial Least Squares*
SiPLS	*Synergy Interval Partial Least Squares*
GARGS	*Genetic Algorithm-Based Region Selection*
RPLS	*Regression Partial Least Squares*
UVE	*Uninformative Variable Elimination*
MC	*Monte Carlo*
GLM	*Generalized Linear Models*
BD	*Bhattacharyya Distance*
$\boldsymbol{\chi^2}$	Qui-quadrado
ReF	*RelieF*
LS	*Laplacian Score*
GI	Ganho de Informação
RG	Razão de Ganho
IG	Índice Gini
NL	Não linear

SUMÁRIO

1. Introdução

1.1 Considerações Iniciais

Uma ampla gama de produtos tem sido alvo de falsificações, acarretando diversos problemas ao meio ambiente, à saúde e à segurança pública, bem como em termos financeiros (DORTA et al., 2018; REID; O'DONNELL; DOWNEY, 2006; WHO, 1999). Tais impactos negativos têm motivado a intensificação de pesquisas voltadas à identificação e rastreamento de falsificações (LOPES; WOLFF, 2009). Nessa perspectiva, a aplicação de ferramentas analíticas para controle de qualidade tem se mostrado crucial em diversos processos produtivos, garantindo tanto a segurança ao consumir o produto final quanto a adequação do produto ou processo à legislação vigente (KELLY, 2003).

Com relação à processos químicos, uma das técnicas analíticas instrumentais de maior uso laboratorial, tanto em indústrias quanto no meio acadêmico, é a espectroscopia (VAZ JUNIOR, 2010). Trata-se de uma técnica rápida e versátil de aquisição de espectros que contêm informações químicas de determinada matriz, apresentando aplicação tanto na determinação qualitativa quanto quantitativa de substâncias moleculares de diversas naturezas (OKAMOTO, 2014). Sendo assim, é capaz de identificar um composto ou investigar a composição de uma amostra, a fim de identificar possíveis adulterações (CÂMARA et al., 2017; GAYDOU; KISTER; DUPUY, 2011).

Contudo, as técnicas analíticas de espectroscopia não fornecem diretamente a informação desejada. Normalmente a propriedade de interesse é determinada por meio de um método de referência, sendo necessário a construção de um modelo de calibração multivariada, no qual se correlacionam matematicamente os espectros com as respectivas variáveis de interesse (MORGANO et al., 2007). Em vista disso, técnicas multivariadas e de inteligência artificial têm sido empregadas na análise de espectros. Além disso, técnicas espectroscópicas tendem a gerar elevado número de variáveis ruidosas e altamente correlacionadas, o que ressalta a importância do uso de técnicas que permitam remover as variáveis não informativas e garantir a construção de modelos consistentes de classificação e predição,diminuindo o risco de inferências não confiáveis e reduzindo o custo computacional(ANZANELLO et al., 2013).

Estratégias de seleção de variáveis em espaços de alta dimensão (geralmente contento centenas ou milhares de variáveis, como é o caso dos dados espectroscópicos) têm atraído

considerável atenção na pesquisa de mineração de dados nos últimos anos. Essa é uma das principais estratégias de redução da dimensionalidade (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015). O objetivo principal da seleção de variáveis é identificar o subconjunto de variáveis preditoras que contém as informações mais significativas para a construção dos modelos. A seleção de variáveis é importante para evitar *overfitting* (sobre ajuste) e melhorar o desempenho do modelo, fornecer modelos mais rápidos e com custo computacional menor e proporcionaruma profunda compreensão sobre os processos que deram origem aos dados (COCCHI; BIANCOLILLO; MARINI, 2018; SAEYS; INZA; LARRAÑAGA, 2007).

A presente dissertação é composta por três artigos abordando seleção de variáveis com o propósito de classificar produtos em categorias que dizem respeito à sua qualidade/autenticidade, bem como predizer propriedades químicas utilizando técnicas analíticas. O primeiro artigo apresenta uma revisão da literatura sobre métodos de seleção de variáveis, incluindo suas etapas, principais características e estudos relevantes. No segundo artigo é proposta uma sistemática para seleção de comprimentos de onda (COs) em espectroscopia que integra uma metodologia de ranqueamento baseada em conceitos químicos, visando atender às características espectrais dos dados para construir subconjuntos de variáveis por meio de uma abordagem direta com redes neurais artificiais (*Artificial Neural Networks* – ANN). Para avaliar o desempenho do método, empregam-se três conjuntos de dados com o objetivo de prever três importantes propriedades do diesel: ponto de ebulição, número de cetano e ponto de fulgor. Os resultados foram ainda comparados com outras abordagens da literatura voltadas à seleção de COs. O terceiro artigo propõe a utilização de florestas aleaórias (*Random Forests*– RF) como alternativa à calibração multivariada em seis bancos dedados de espectroscopia de natureza binária e multiclasse. Seis diferentes técnicas de ranqueamento foram testadas para orientar a seleção das variáveis na ferramenta classificadora RF, de acordo com a abordagem *forward inclusion stepwise*. O método visa atender à ciência forense na identificação de amostras adulteradas em produtos alimentícios e em drogas (lícitas e ilícitas).Os modelos construídos pela melhor combinação de técnicas foram comparados com modelosda literatura para as mesmas bases de dados.

1.2 Objetivos

O objetivo principal da dissertação é a proposição de novas abordagens para seleção de comprimentos de onda com vistas à predição de propriedades químicas de produtos e classificação de produtos em categorias associadas à qualidade ou autenticidade.

Os seguintes objetivos específicos são apresentados:

- Avaliar as principais técnicas de análise multivariada de dados, sua organização e contribuição para o processo de seleção de variáveis;
- Criar um índice de importância de variáveis com base na lei de Lambert-Beer com vistas a mensurar a relevância das variáveis para fins de predição de propriedades de amostras;
- Avaliar o desempenho da técnica floresta aleatória na calibração multivariada de dados espectroscópicos;
- Avaliar o desempenho de diferentes metodologias de ranqueamento de importância de variáveis;
- Comparar os resultados dos métodos propostos a outras abordagens voltadas à seleção de variáveis; e
- Avaliar o desempenho das proposições em dados de NIR com diferentes dimensionalidades e tipos de variável de resposta (discreta e contínua).

1.3 Justificativa do Tema e dos Objetivos

A adulteração de produtos, apesar de ilegal, demonstra ser uma atividade altamente lucrativa, devido, principalmente, ao baixo custo de produção. Tipicamente, os produtos introduzidos no mercado apresentam composição similar aos autênticos, porém fora dos padrões de qualidade e segurança necessários fazendo uso de etiquetas, embalagens ou instruções falsificadas (MARUCHECK et al., 2011; WHO, 1999), podendo acarretar diversos problemas relacionados ao meio ambiente (CUNHA et al., 2019; KRISHNA et al., 2019; LI et al., 2005)e à saúde pública (KAHMANN et al., 2018), bem como impactos de cunho financeiro (CÂMARA et al., 2017).

O avanço de tecnologias de análise baseadas em espectroscopia tornou possível a extração de informações precisas referentes à composição de uma amostra de forma não-destrutiva, rápida e precisa. Dessa maneira, é possível realizar a identificação de amostras adulteradas e impedir a circulação de produtos não autênticos no mercado, além de juntar informações que permitam interromper a atividade fraudadora. Entretanto, a grande quantidade de dados gerada através dessas tecnologias traz consigo a necessidade de utilização de técnicas de análise de dados para extrair informação útil (uma vez que, dados por si só não transmitem nenhuma mensagem que possibilite o entendimento sobre determinada situação). Além disso, cenários caracterizados por grandes volumes de dados requerem a utilização de técnicas de redução de dimensionalidade que permitam analisar os dados sem a perda de informações relevantes (GAUCHI; CHAGNON, 2001), diminuindo o risco de inferências não confiáveis assim como o custo computacional.

Desta forma, a sistemática aqui proposta encontra respaldo prático ao proporcionar aos profissionais – que necessitam lidar com bancos com tendência crescente de dimensionalidade – solucionar problemas por meio da redução do espaço de busca de hipóteses, melhorando o desempenho e simplificando os resultados da modelagem de dados espectrais. Por fim, a seleção de variáveis também possibilita diminuição dos custos de experimentos e coletas de dados, reduzindo despesas associadas a procedimentos laboratoriais.

No âmbito acadêmico, observa-se um constante interesse pelo desenvolvimento de abordagens com vistas à seleção de variáveis em diversas áreas de conhecimento. Muitas abordagens contemporâneas de modelagem têm se apoiado em ferramentas robustas de predição/classificação por meio de algoritmos de aprendizado de máquina baseados em inteligência artificial (IA) tais como: Redes Neurais Artificiais (*Artificial Neural Networks* – ANN) e Florestas Aleatórias (*Random Forests* – RF). Abordagens desta naturezapodem se aproximar dos relacionamentos não-lineares e complexos encontrados na natureza(ATIQUZZAMAN; KANDASAMY, 2018). Além disso seu desempenho não é significativamente influenciado por variações nos dados, como por exemplo, *outliers*. Por estas razões, tais abordagens tem se mostrado superiores no processo de modelagem em comparação a outros modelos como por exemplo,a regressão PLS tradicional que, embora seja capaz de lidar com grande número de variáveis correlacionadas e acometidas por ruído (ZIMMER; ANZANELLO, 2014), fornece apenas modelos preditivos lineares, podendo levar

adesempenhos preditivos ruins na presença de uma forte relação não linear (NL) entre variáveis independentes e dependentes (LAVOIE; MUTEKI; GOSSELIN, 2019). Índices de importância também vem sendo bastante empregados como ferramenta de seleção preliminar, em técnicas de aprendizado de máquina em etapa subsequente, de forma a guiar a eliminação das variáveis irrelevantes. Dentre as diversas aplicações sugeridas por abordagens para a seleção de variáveis tem-se: controle de qualidade (ANZANELLO et al., 2015; KAHMANN et al., 2017; SOARES et al., 2017), análise forense(KAHMANN et al., 2018), análise financeira (PAIVA et al., 2019), sistemas de segurança (ROLDÁN et al., 2020) e saúde (SINGH, 2019).

O desenvolvimento deste estudo se justifica em razão doaprimoramento de técnicas existentes através da combinação de novas técnicas e ferramentas que permitam a identificação de variáveis espectroscópicas relevantes para compor modelos robustos com vistas à classificação e predição de propriedades químicas.

1.4 Procedimentos Metodológicos

Em relação aos objetivos, a presente dissertação é classificada como pesquisa exploratória, uma vez que busca construir hipóteses para resolver os problemas a partir da sua análise. Quanto à natureza, é caracterizada como pesquisa aplicada, dado que a fundamentação teórica é explorada com vistas à solução de problemas genéricos (CRESWELL, 2010; GIL, 2002). A dissertação apresenta ainda abordagem quantitativa, em razão do uso de análises estatísticas e modelagem matemática para solução dos problemas apresentados. Além disso, o estudo faz uso de procedimentos de pesquisa bibliográfica.

No primeiro artigo é realizada uma pesquisa bibliográfica sobre métodos de seleção de variáveis, com intuito de apresentar uma visão geral sobre o tema. Além das etapas necessárias ao desenvolvimento destes métodos, são apresentadas as principais características, tipos de abordagens e estudos representativos reportados pelaliteratura.

No segundo artigo,um método proposto para seleção de comprimentos de onda para predição de propriedades químicas de amostras de diesel/biodiesel inicia com a identificação dos comprimentos de onda com maior variabilidade. Para isso, os conjuntos de dados são ordenados em ordem crescente em relação à variável dependente contínua y (absorbância), e são divididos em quatro quartis (Q1, Q2, Q3 e Q4). Em seguida, calcula-se a absorbância média de cada comprimento de onda das amostras inseridas em Q1; o mesmo é feito para as amostras

inseridas em Q4. Um vetor contendo as diferenças das médias entre os quartis extremos é criado. Para a geração do *ranking* basta listar este vetor em ordem decrescente. A etapa final consiste na criação dos subconjuntos. Nesta etapa, as variáveis são iterativamente testadas no modelo ANN, de acordo com o *ranking* gerado anteriormente, seguindo a abordagem de inserção de variáveis, uma por vez, em um conjunto inicialmente unitário (método *forward*). O melhor subconjunto de variáveis é identificado para cada banco de dados baseado na raiz do erro quadrático médio de predição (*Root Mean Square Error* – RMSE). O desempenho do método é ainda comparado a outros métodos de seleção de COs da literatura.

O terceiro artigo tem por objetivo identificaramostras adulteradas em produtos alimentícios e em drogas (lícitas e ilícitas). Para isso, propôs-se a utilização da ferramenta de aprendizado de máquina RF para a calibração multivariada em seis bancos de dados de espectroscopia de natureza binária e multiclasse. Seis diferentes técnicas de ranqueamento de variáveis de acordo com a sua relevância (qui-quadrado, *relief-F, laplacian score*, ganho de informação, razão de ganho e índice gini) foram testadas para orientar a inserção *forward*das variáveis na ferramenta classificadora RF. A partir de um subconjunto inicialmente vazio, as variáveis são inseridas uma a uma, da mais para a menos importante, na ferramenta de classificação. Se houver aumento na acurácia, a variável candidata é mantida no subconjunto de variáveis selecionadas; caso contrário, é eliminada e a próxima variável do ranqueamento é adicionada ao subconjunto. Quando todas as variáveis forem testadas ou o subconjunto em análise atingir acurácia de 100%, o processo é encerrado. O desempenho do modelo construído pela melhor combinação de técnicas foi comparado comoutras abordagens para seleção de COs da literatura.

1.5 Estrutura da Dissertação

A dissertação encontra-se dividida em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos e as justificativas, bem como o método de pesquisa adotado. A estrutura do trabalho e a delimitação do estudo finalizam o capítulo.

No segundo capítulo é apresentado o primeiro artigo, que apresenta uma revisão bibliográfica de métodos de seleção de variáveis, com intuito de apresentar uma visão geral sobre o tema. Além das etapas necessárias ao desenvolvimento destes métodos, são

apresentadas as principais características, tipos de abordagens e alguns estudos representativos das abordagens listadas.

O terceiro capítulo apresenta o segundo artigo, o qual visa predizer propriedades químicas das amostras de misturas de combustível Diesel/Biodisel com vistas ao monitoramento da qualidade do combustível. Para este fim, 3 bases de dados foram utilizadas. É proposto um método para seleção dos comprimentos de onda baseado em conceitos da lei de *Lambert-Beer* para o ordenamento dos comprimentos de onda, os quaissão inseridos na ferramenta preditiva ANN usando uma abordagem *forward*. O método proposto é comparado com métodos tradicionais de seleção de variáveis.

O quarto capítulo apresenta o terceiro artigo, que objetiva classificar amostras de alimentos e drogas (lícitas e ilícitas),de seis bases de dados, quanto à sua autenticidade (seja pela região de origem ou adulteração). Seis bases de dados foram utilizadas. Trata-se de um método que engloba classificação binária e multiclasse. É proposto um método para seleção dos COs utilizando o classificador RF para obter subconjuntos de variáveis através da abordagem de seleçãode inserção ou remoção de variáveis, uma por vez, em um conjunto inicialmente unitário. Seis técnicas de ranqueamento – qui-quadrado, *relief-F*, *laplacian score*, ganho de informação, razão de ganho e índice gini – são utilizadas para ordenar os COs. A combinação de cada técnica de ranqueamento com a ferramenta RF é analisada por meio da acurácia gerada pelo melhor subconjunto de comprimentos de onda resultante para cada método (ranqueamento + RF).

Por fim, o quinto e último capítulo traz a conclusão do trabalho, na qual são avaliados os principais resultados frente aos objetivos almejados e as delimitações citadas. Essa seção traz ainda sugestões para desdobramentos futuros.

1.6 Delimitações do Estudo

Constituem-se em limitaçõesdo presente estudo:

- A aplicação das técnicas restringe-se a dados de natureza espectral;
- Os métodos desenvolvidos utilizam somentetécnicas baseadas em índices de importância aliadas a ferramentas de aprendizado de máquina, não sendo utilizadas outras abordagens de seleção;

- O trabalho não propõe novas ferramentas de classificação ou regressão, restringindo-se a combinar tais ferramentas de forma a gerar novas abordagens para seleção de variáveis; e
- Somente métodos supervisionados são abordados nas sistemáticas de seleção de variáveis propostas, sendo necessária prévia informação sobre a classe de cada amostra (ou variável dependente) a ser modelada.

1.7 Referências

ANZANELLO, M. J.; FU, K.; FOGLIATTO, F. F.; FERRÃO, M. F. HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point. **Chemometrics and Intelligent Laboratory Systems**, 2015.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGERB, R. P.; MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209–214, 2013.

ATIQUZZAMAN, M.; KANDASAMY, J. Robustness of Extreme Learning Machine in the prediction of hydrological flow series. **Computers and Geosciences**, v. 120, p. 105–114, 2018. Elsevier Ltd.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. **Feature Selection for High-Dimensional Data.** Springer ed. 2015.

CÂMARA, A. B. F.; DE CARVALHO, L. S.; DE MORAIS, C. L. M.; et al. MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends. **Fuel**, 2017.

COCCHI, M.; BIANCOLILLO, A.; MARINI, F. Chemometric Methods for Classification and Feature Selection. **Comprehensive Analytical Chemistry**. v. 82, p.265–299, 2018.

CRESWELL, J. W. **Projeto de pesquisa métodos qualitativo, quantitativo e misto**.

2010.

CUNHA, D. A.; NETO, Á. C.; COLNAGO, L. A.; CASTRO, E. V. R.; BARBOSA, L. L. Application of time-domain NMR as a methodology to quantify adulteration of diesel fuel with soybean oil and frying oil. **Fuel**, 2019.

DORTA, D. J.; YONAMINE, M.; COSTA, J. L. DA; MARTINIS, B. S. DE. **Toxicologia forense**. 2018.

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. **Anais...** , 2001.

GAYDOU, V.; KISTER, J.; DUPUY, N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. **Chemometrics and Intelligent Laboratory Systems**, 2011.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4º ed. São Paulo: Atlas S.A., 2002.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018.

KAHMANN, A.; ANZANELLO, M. J.; MARCELO, M. C. A.; POZEBON, D. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 140, p. 348–360, 2017.

KELLY, S. D. **Using stable isotope ratio mass spectrometry (IRMS) in food authentication and traceability**. 2003.

KRISHNA, S. M.; ABDUL SALAM, P.; TONGROON, M.; CHOLLACOOP, N. Performance and emission assessment of optimally blended biodiesel-diesel-ethanol in diesel engine generator. **Applied Thermal Engineering**, v. 155, p. 525–533, 2019.

LAVOIE, F. B.; MUTEKI, K.; GOSSELIN, R. A novel robust NL-PLS regression methodology. **Chemometrics and Intelligent Laboratory Systems**, v. 184, p. 71–81, 2019.

LI, D. G.; ZHEN, H.; XINGCAI, L.; WU-GAO, Z.; JIAN-GUANG, Y. Physico-chemical properties of ethanol-diesel blend fuel and its effect on performance and emissions of diesel engines. **Renewable Energy**, v. 30, n. 6, p. 967–976, 2005.

LOPES, M. B.; WOLFF, J. C. Investigation into classification/sourcing of suspect counterfeit Heptodin^{TM} tablets by near infrared chemical imaging. **Analytica Chimica Acta**, v. 633, n. 1, p. 149–155, 2009.

MARUCHECK, A.; GREIS, N.; MENA, C.; CAI, L. Product safety and security in the global supply chain: Issues, challenges and research opportunities. **Journal of Operations Management**, v. 29, n. 7–8, p. 707–720, 2011.

MORGANO, M. A.; DE FARIA, C. G.; FERRÃO, M. F.; FERREIRA, M. M. C. Determinação de açúcar total em café cru por espectroscopia no infravermelho próximo e regressão por mínimos quadrados parciais. **Quimica Nova**, v. 30, n. 2, p. 346–350, 2007.

OKAMOTO, R. T. **Desenvolvimento de método analítico rápido de cefalexina**, 2014. Dissertação de mestrado. Universidade de São Paulo.

PAIVA, F. D.; CARDOSO, R. T. N.; HANAOKA, G. P.; DUARTE, W. M. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. **Expert Systems with Applications**, v. 115, p. 635–655, 2019. Elsevier Ltd.

REID, L. M.; O'DONNELL, C. P.; DOWNEY, G. Recent technological advances for the determination of food authenticity. **Trends in Food Science and Technology**, v. 17, n. 7, p. 344–353, 2006.

ROLDÁN, J.; BOUBETA-PUIG, J.; LUIS MARTÍNEZ, J.; ORTIZ, G. Integrating complex event processing and machine learning: An intelligent architecture for detecting IoT security attacks. **Expert Systems with Applications**, v. 149, p. 113251, 2020.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics Review**, v. 23, p. 2507–2517, 2007.

SINGH, B. K. Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning

paradigm. **Biocybernetics and Biomedical Engineering**, v. 39, n. 2, p. 393–409, 2019.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 171–178, 2017.

VAZ JUNIOR, S. Análise Química Instrumental e sua Aplicação em Controle de Qualidade de Biocombustíveis. **Embrapa Agroenergia-Circular Técnica (INFOTECA-E)**, 2010. Brasília.

WHO. **Guidelines for the development of measures to combat counterfeit drugs**. 1999.

ZIMMER, J.; ANZANELLO, M. J. A new framework for predictive variable selection based on variable importance indices. **Producao**, v. 24, n. 1, p. 84–93, 2014.

2. Primeiro artigo:Uma revisão dos princípios e mecanismos básicos dos métodos de seleção de variáveis

Resumo

Métodos de seleção de variáveis têm atraído crescente atenção no campo do aprendizado de máquina, desempenhando um papel significativo na melhoria da performance e precisão dos modelos. Dado o elevado número de abordagens e métodos voltados ao problema de seleção de variáveis, não há um consenso na literatura sobre quais os melhores métodos de seleção, sendorecomendável avaliar cuidadosamente os princípios de cada abordagem de seleção e sua aderência aos objetivos de cada aplicação. Por esta razão, seleção de variáveis continua sendo uma das linhas de pesquisa ativamente perseguidas por conta de suas implicações teóricas e práticas. À medida que estratégias de mineração de dados se expandem para novas áreas, a seleção de variáveis também enfrenta novos desafios. Em vista disso, esteartigo fornece uma revisão de vários aspectos associados ao problema da seleção de variáveis, apresentando a estrutura geral para processos de seleção e categorização de diversas abordagens. Tem como objetivos facilitar o entendimento das diversas abordagens existentes, bem como orientar o desenvolvimento de novos algoritmos de seleção de variáveis.

Palavras-chave: Seleção de variáveis, aprendizado de máquina, dados de alta dimensionalidade.

2.1 Introdução

Os avanços tecnológicos têm favorecido progressos substanciais no que diz respeito à administração e armazenamento de variadas e massivas quantidades de dados, as quais se caracterizam como a base para a informação. Porém, dados por si só não transmitem nenhuma mensagem que possibilite o entendimento dedeterminada situação, tornando-se necessário o uso de técnicas de análise de dados para extração de informações relevantes e aplicáveis. Em vista disso, algoritmos de aprendizado de máquina têm ganhado popularidade no meio acadêmico nos últimos anos. Estes referem-se a métodos de análise de dados que automatizam a construção de modelos analíticos e baseiam-se na ideia de que sistemas podem aprender com

dados, identificar padrões e tomar decisões com o mínimo de intervenção humana (HAMET; TREMBLAY, 2017).

Contudo, lidar com dados massivos é uma tarefa desafiadora. De acordo com Salimi et al. (2018), excessivos volumes de dados trazem consigo variáveis irrelevantes ou redundantes que contribuem apenas para aumentar o tamanho e a complexidade do espaço de variáveis. Além disso, nota-se um exponencial aumento na quantidade de problemas de alta dimensionalidade – bancos de dados que possuem um número de variáveis significativamente superior em relação à quantidade de observações – o que tende a prejudicar o desempenho dos algoritmos de aprendizagem em termos de velocidade e precisão (KIRA; RENDELL, 1992; LEE; LOH; CHIN, 2017; SAEYS; INZA; LARRAÑAGA, 2007; URBANOWICZ et al., 2018). Espaços de alta dimensão podem apresentar, ainda, insuficiência de graus de liberdade para estimação de métodos frequentistas e probabilidade significativa de haver multicolinearidade elevada entre esses preditores (VASCONCELOS, 2017).

Em vista disso, métodos de seleção de variáveis têm atraído considerável atenção no campo do aprendizado de máquina com respeito a abordagens supervisionadas, semi-supervisionadas e não supervisionadas. Seleção de variáveis, também conhecida como seleção de *features* ou seleção de atributos, é uma das principais estratégias de redução da dimensionalidade; seu principal objetivo é identificar o subconjunto de variáveis úteis que preservam as principais informações e estrutura dos dados, eliminando, assim, as variáveis irrelevantes e redundantes ((BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015;YAN et al., 2020). Dessa forma, espera-se que abordagens de seleção de variáveis aprimorem a interpretabilidade do modelo proposto, a precisão e o desempenho dos métodos analíticos, além de fornecer modelos mais rápidos, confiáveis e com custo menorcomputacional (COCCHI; BIANCOLILLO; MARINI, 2018; NADLER; COIFMAN, 2005; SAEYS; AKI INZA; LARRAÑ AGA, 2007; THOMAS, 1994; XIAOBO et al., 2010).

A partir de uma revisão bibliográfica, este artigo traz uma compilação das informações referentes a métodos de seleção de variáveis – incluindo elementos, principais características e estudos relevantes – visando contribuir na disseminação dos conhecimentos necessários para o desenvolvimento de novos métodos de seleção. Este artigo tem como objetivo fornecer informações básicas de metodologia de seleção de variáveis, afim de facilitar o entendimento

das diversas abordagens existentes, bem como orientar o desenvolvimento de novos algoritmos de seleção de variáveis.

2.2 Problema da dimensionalidade dos dados

O número de variáveis significativamente superiorao número de observações de uma amostra gerado por algumas técnicas (por exemplo, técnicas espectrais) configura em um problema de alta dimensionalidade dos dados. Matematicamente, isto é entendido como uma matriz que apresenta um pequeno número de linhas (instâncias) em relação ao número de colunas (variáveis ou dimensões) (ZAKI; ZULKURNAIN, 2018). Com as variáveis selecionadas, um método de modelagem é usado para construir o relacionamento entre variáveis e a propriedade de interesse.

No aprendizado supervisionado, um algoritmo de indução de classificação – usado para construir o relacionamento entre variáveis e a propriedade de interesse – trata cada instância do conjunto de treinamento como um vetor de características(variáveis ou atributos) e um rótulo de classe (KOHAVI; JOHN, 1997).A partir das instâncias, é aplicado o algoritmo de aprendizado de máquina para geração do classificador, o qual trata-se de uma função (DE OLIVEIRA JUNIOR, 2010).

Espaços de alta dimensão (geralmente com centenas ou milhares de dimensões) tendem a prejudicar o desempenho dos algoritmos de aprendizagem no que se refere à velocidade e taxa de acerto (KIRA; RENDELL, 1992; LEE; LOH; CHIN, 2017; URBANOWICZ et al., 2018). Isso ocorre pelo seguinte motivo: cada atributo *d* pode ser visto como uma coordenada do espaço *d*-dimensional (FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, 2011). Logo, conforme o número de dimensões cresce, as instâncias tornam-se mais dispersas no espaço, havendo menos instâncias por regiões e tornando a tarefa de aprendizado mais difícil, uma vez que algoritmos de aprendizado de máquina constroem preditores com base nas proporções de instâncias estimadas em cada classe por regiões do espaço. A figura2.1 ilustra tal situação, na qual gráficos com dimensões distintas (unidimensional, bidimensional e tridimensional), para um mesmo número de amostras, foram gerados. A partir destes é possível verificar a dispersão das instâncias em função do aumento da dimensionalidade dos dados. Na figura 2.1, cada eixo (x, y e z) representa uma variável preditiva.

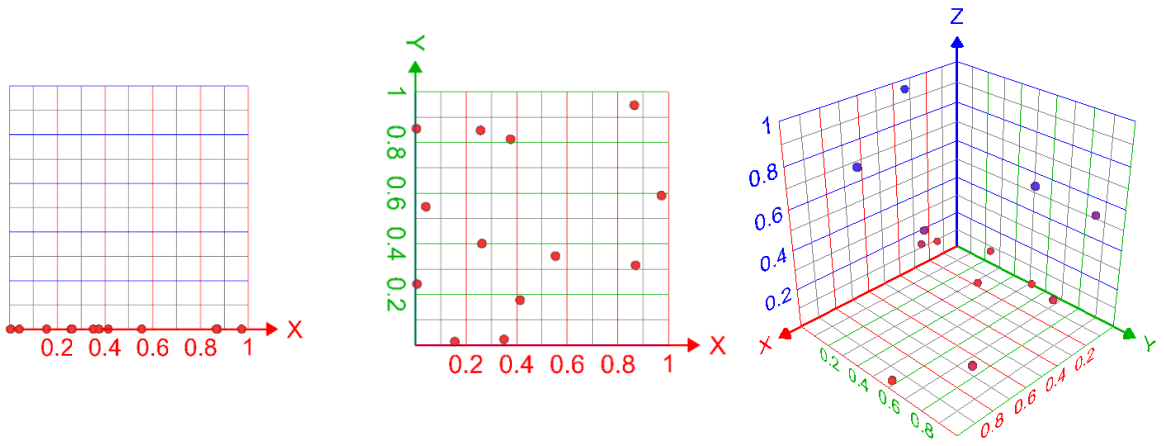

Figura 2.1– Dispersão das instâncias em função do aumento da dimensionalidade dos dados
Fonte: Elaborado pela autora (2018).

Problemas causados em espaços de alta dimensão são explicados pelo fenômeno denominado “maldição da dimensionalidade” introduzido por Bellman (1961). O autor o define como o crescimento exponencial do número de amostras que são necessárias para estimar uma função arbitrária de acordo com o número de variáveis de entrada da função. Na prática, a maldição da dimensionalidade implica que, para um dado tamanho de amostra, existe um número máximo de características, a partir do qual o desempenho dos algoritmos irá degradar ao invés de melhorar. Para contornar este problema, pesquisadores têm se dedicado ao desenvolvimento e aprimoramento de técnicas que envolvem redução de dimensionalidade dos dados, que do ponto de vista computacional diminuem o espaço de busca de hipóteses, melhoram o desempenho dos modelos e simplificam sua interpretação (CAMARGO, 2010).

Existem duas abordagens principais para reduzir o número de dimensões de um conjunto de dados: Extração de variáveis (ou agregação) e Seleção de variáveis (FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, 2011). A primeira realiza uma combinação dos atributos originais no conjunto de dados afim de reduzir a quantidade de atributos de um conjunto de dados bruto. Já os métodos de seleção de variáveis, foco deste artigo, atuam como filtros eliminando variáveis irrelevantes ou redundantes (KOTU; DESHPANDE, 2015).

2.3 Procedimento geral de seleção de variáveis

Para desenvolver métodos de seleção de variáveis, faz-se necessário definir uma estratégia de busca do subconjunto ótimo, uma medida de importância para identificar variáveis

relevantes e um critério de avaliação dos subconjuntos gerados. Afigura 2.2 apresenta uma visão geral da categorização de métodos de seleção de variáveis, os quais serão abordados nas seções subsequentes.

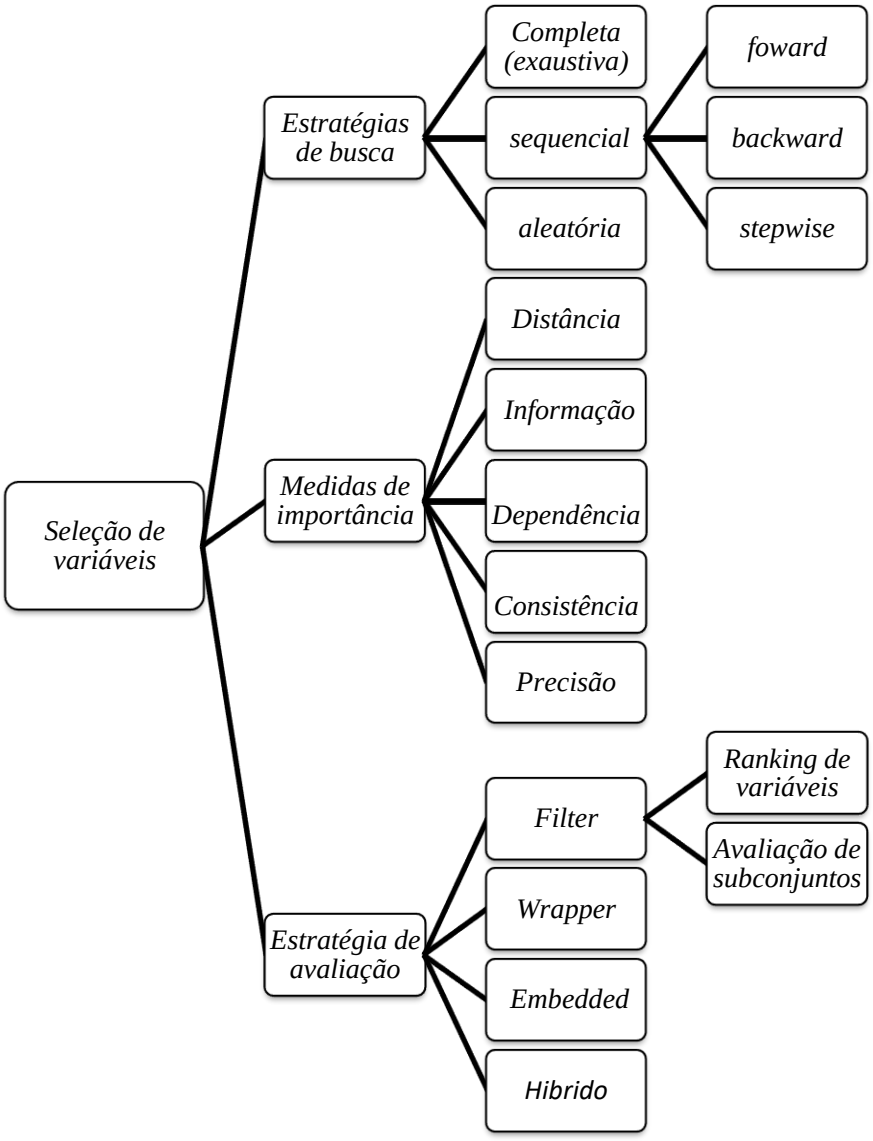

Figura 2.2– Categorização dos métodos de seleção de variáveis
Fonte: Elaborado pelos autores (2019)

Métodos de seleção de variáveis, em geral, apoiam-seem quatro etapas básicas: *(i)* geração de um subconjunto de variáveis a partir do conjunto original dos dados, *(ii)* avaliação do subconjunto gerado, *(iii)* definição do critério de parada e *(iv)* validação do resultado (LIU; YU, 2005). A figura 2.3 apresenta a sequência genérica do processo de seleção de variáveis, as quais são detalhadas na sequência.

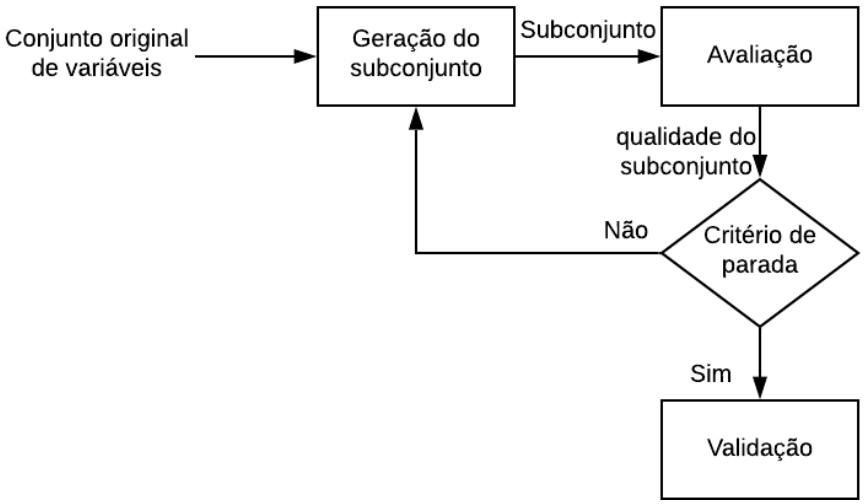

Figura 2.3 – Etapas chave para a seleção de variáveis
Fonte: Kumar e Rath (2016)

2.3.1 Geração do subconjunto

A seleção de um subconjunto de variáveis pode ser vista como um problema de busca (FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, 2011). Nesse caso, cada ponto no espaço de busca pode ser visto como um possível subconjunto de variáveis. O primeiro passo para se gerar um subconjunto de variáveis é definir o ponto de partida da busca, para o qual uma estratégia de busca deve ser definida.

Estratégias de busca podem ser do tipo completa (exaustiva), sequencial ou aleatória. As estratégias de busca completa garantem encontrar o subconjunto ótimo, pois estas testam todos os possíveis subconjuntos de variáveis. Porém, uma busca exaustiva, onde todos os subconjuntos são testados, pode tornar-se impraticável, uma vez que existem 2^d subconjuntos possíveis para as *d* variáveis existentes (ARAUZO-AZOFRA et al., 2017; BLUM; LANGLEY, 1997; (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015). No caso de estratégias sequenciais, diferentes heurísticas são aplicadas para reduzir o espaço de busca sem comprometer a chances de encontrar o resultado ideal. Estas incluem *i)* remoção de variáveis do conjunto original de dados, conhecido como abordagem*backward*; *ii)* inserção de variáveis em um conjunto inicialmente vazio (*forward*); *iii)* adição ou remoção simultânea de variáveis (*stepwise*) (BLUM; LANGLEY, 1997; CILIA et al., 2019). Porém, como nem todos os subconjuntos possíveis são testados, estratégias desse tipo podem omitir algumas variáveis relevantes, levando à perda de subconjuntos ótimos (SETIONO, 1997). Por fim, existe a

estratégia aleatória, a qual busca subconjuntos com algum tipo de aleatoriedade(LIU; YU, 2005). Alguns métodos de busca aleatória incluem GARGS (*Genetic Algorithm – Based Region Selection*)(HASEGAWA; KIMURA; FUNATSU, 1997), MC-UVE (Monte Carlo – *Uninformative Variable Elimination*) (CAI; LI; SHAO, 2008), e MCS-RPLS (Monte Carlo – *Partial least squares regression*) (ZHANG; ZHANG; IQBAL, 2013).

2.3.2 Avaliação do subconjunto gerado

Cada subconjunto de variáveis gerado precisa ser analisado por um critério de avaliação. As estratégias de avaliação dos subconjuntos gerados podem ser dependentes ou independentes do algoritmo de indução, sendo divididas em três abordagens: *filter* (filtro), *wrapper* (empacotamento) e *embedded* (embutido) ((BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015; CILIA et al., 2019). Há também uma tendência em combinar algoritmos de diferentes origens conceituais em um processo sequencial; métodos que utilizam esta abordagem são denominados híbridos (REMESEIRO; BOLON-CANEDO, 2019).

2.3.2.1 Abordagem filter

Normalmente, um critério independente é usado em modelos do tipo *filter* (figura 2.4). Aqui, o processo de seleção de variáveis é conduzido por meio da avaliação da qualidade de uma variável ou de um subconjunto de variáveis usando uma medida de qualidade independente do algoritmo de indução que será aplicado às variáveis selecionadas (WAN, 2018). Neste caso, a seleção é utilizada como uma etapa de pre-processamento, onde se procura por variáveis significantes estimando e classificando-as de acordo com sua importância. Por fim, o subconjunto de atributos selecionados é apresentado como entrada para o algoritmo de classificação. As vantagens de modelos baseados na abordagem *filter* são que estes possuem baixo custo computacional e boa capacidade de generalização (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015)

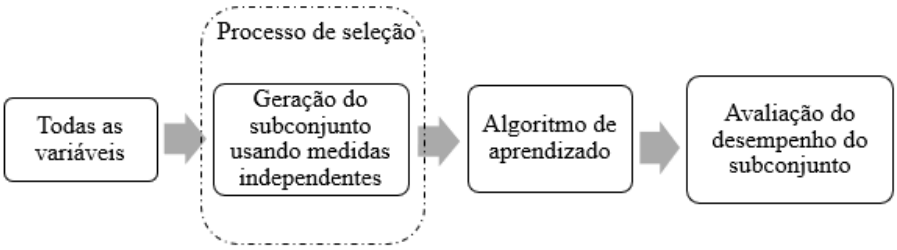

Figura 2.4– Seleção de variáveis pela abordagem filter
Fonte: Elaborado pela autora (2018).

Os métodos baseados na abordagem *filter* podem compreender tanto técnicas univariadas (avaliação individual de variáveis) como multivariadas (avaliação de subconjuntos de variáveis) (YU; LIU, 2004). As técnicas univariadas avaliam todas as variáveis individualmente e, posteriormente, algum critério de corte é aplicado para decidir quais variáveis serão selecionadas (ARAUZO-AZOFRA et al., 2017). Já as técnicas multivariadas realizam a análise das variáveis de forma conjunta (VICINI, 2005) e, por isso, são recomendadas para cenários em que as variáveis estejam correlacionadas, uma vez que seus efeitos não podem ser interpretados de maneira eficiente separadamente. Dessa forma, técnicas multivariadas desembaraçam a sobreposição de informações fornecidas pela correlação e esclarecem informações relevantes (YAMASHITA, 2015).

Métodos *filter* baseados na análise univariada referem-se aos métodos de ranqueamento de variáveis. Uma maneira de ordenar um conjunto de variáveis em um *ranking* é utilizar métricas para avaliar a relevância de cada variável para um dado critério. Tal *ranking* pode então ser utilizado para selecionar variáveis cujo valor da métrica tenha resultado em um valor igual ou superior a um limite estabelecido (PEREIRA, 2009). Dependendo da aplicação, as medidas podem ou não considerar a informação sobre a classe. No caso da ordenação dependente de classe, a primeira variável é aquela que melhor discrimina as instâncias das diferentes classes, a seguinte é a segunda melhor variável para essa discriminação e assim por diante (FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, 2011).

Algumas métricas usadas pelos métodos do tipo *filter* incluem medidas de distância, medidas de informação, medidas de dependência e medidas de consistência (LIU; YU, 2005).

- **Medidas de distância** – também conhecidas como separabilidade, medidas de divergência ou discriminação. Nesse caso, procura-se a variável que conduz à máxima separação das classes existentes. Exemplos incluem PCA (*Partial Least*

Squares) (BENIN et al., 2003) e *Laplacian Score*, que utiliza a distância Euclidiana como medida de dissimilaridade(ZHAO; LIU, 2011), e relief-F (*RELevance In Estimating Features*) que procura pelas instâncias mais próximas de mesma classe e de classes distintas por meio da distância de Manhattan (LEE, 2005; SHARMA et al., 2017).

- **Medidas de informação** – buscam variáveis que resultam em ganho de informação, determinando a diferença entre a incerteza a priori e a posteriori associada à variável investigada (YU; LIU, 2004). Exemplos são Informação mútua (VERGARA; ESTÉVEZ, 2014), ganho de informação (SHARMA et al., 2017), e razão de ganho (SHARMA et al., 2017).
- **Medidas de dependência** – também são conhecidas como medidas de correlação ou medidas de similaridade, avaliam a capacidade de prever o valor de uma variável a partir do valor de outra (HALL, 1999). Na seleção de variáveis para classificação, procura-se o quão fortemente uma variável está associada à classe (por exemplo, CFS (seleção de atributos baseada em correlação)) (HORTA et al., 2010; TERUYA, 2008).
- **Medidas de consistência** – tentam encontrar um número mínimo de variáveis que separam classes tão consistentemente quanto o conjunto completo de variáveis. A inconsistência é definida como duas instâncias com o mesmo valor de variável, porém rótulos de classe diferentes (LIU; YU, 2005). Pode-se citar o algoritmo CBF (*Consistency-Based Filter*) (TERUYA, 2008) como exemplo de técnicas baseadas em medidas de consistência.

2.3.2.2 Abordagem wrapper

Assim como as abordagens do tipo *filter* (que avaliam subconjuntos de variáveis), as abordagens *wrapper* (figura 2.5) realizam uma busca entre os possíveis subconjuntos a serem avaliados e, em vez de usar um teste independente como nas abordagens *filter*,utilizao próprio algoritmo de indução adotado como uma caixa preta para avaliar os subconjuntos de variáveis de acordo com a sua capacidade preditiva (KOHAVI; JOHN, 1997). Métodos desta natureza utilizam estratégias de busca do tipo sequencial e a importância da variável é baseada em medidas de precisão. Estas medidas são dependentes do algoritmo de aprendizado considerado,

pois a importância dos subconjuntos de variáveis selecionadas está diretamente relacionada ao desempenho preditivo do modelo induzido por um determinado algoritmo (PARMEZAN et al., 2012). No entanto, a complexidade de tempo na abordagem *wrapper* é relativamente maior que em abordagens *filter* e *embedded*. A razão disto é que, na abordagem *wrapper*, o algoritmo de indução necessita ser executado diversas vezes (WAN, 2018); exemplos de métodos *wrapper*incluemSFS(*Sequential Forward Selection*) (SFS) e SBE (*Sequential Backward Elimination*)(KITTLER, 1978).

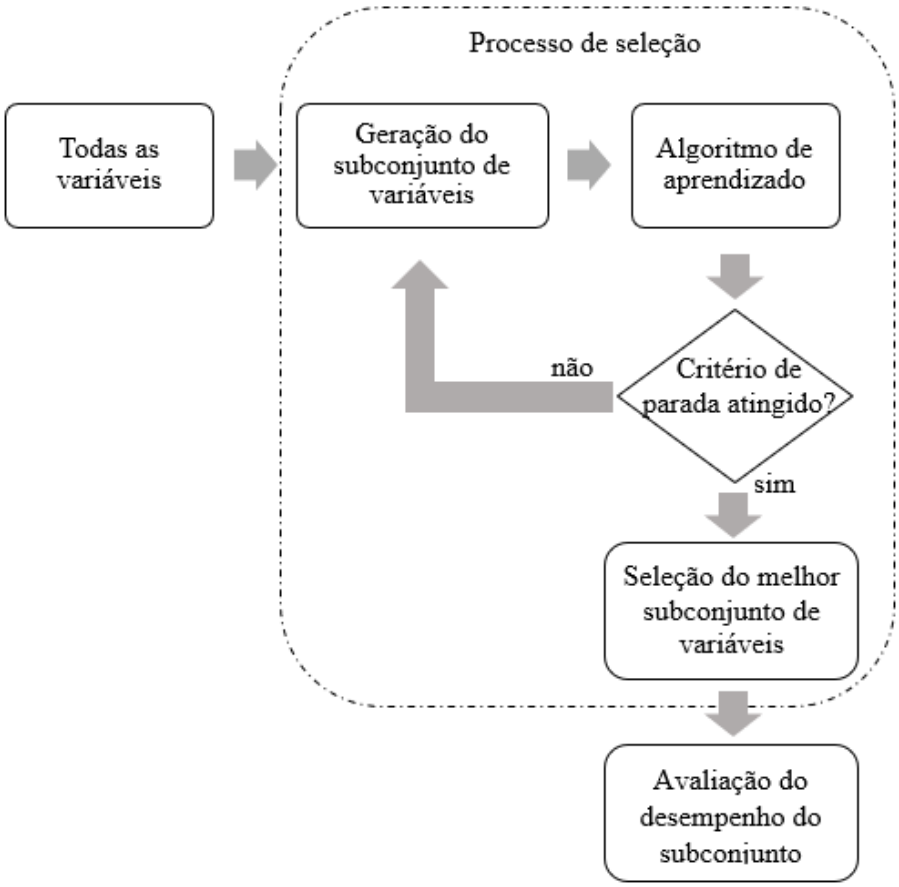

Figura 2.5– Seleção de variáveis pela abordagem wrapper
Fonte: Elaborado pela autora (2018).

2.3.2.3 Abordagem embedded

Já em abordagens do tipo *embedded* (figura2.6) ,a seleção do subconjunto é integrada no próprio algoritmo de indução. Em outras palavras, as variáveis são selecionadas durante o processo de aprendizado. Alguns exemplos de métodos *embedded* são regressão L1 (ou LASSO) para modelos lineares generalizados (GLM – *Generalized Linear Models*)(TIBSHIRANI, 1996), árvores de decisão (*decision tree*), florestas aleatórias (*random*

forest) (GEURTS et al., 2005; WU et al., 2003), Vetor de pesos de algoritmos de máquinas de vetor de suporte (SVM – *Suport Vector Machine*) (WESTON; ELISSEEFF; SCHÖLKOPF, 2003; ZHANG et al., 2006) e redes neurais artificiais (ANN – *Artificial Neural Network*)(SABANDO; PONZONI; SOTO, 2019).

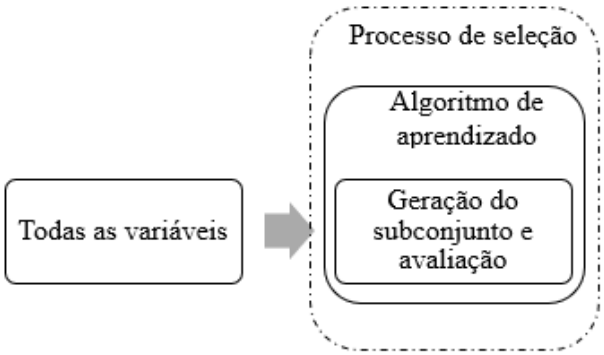

Figura 2.6– Seleção de variáveis pela abordagem *embedded*
Fonte: Elaborado pela autora (2018).

2.3.2.4 Abordagem híbrida

Os métodos híbridos são uma combinação de abordagens baseadas em *filter* e *wrapper*, de forma a explorar seus princípios distintos em diferentes estágios de pesquisa (LIU; YU, 2005). Em geral, o processamento dos dados de alta dimensão é uma tarefa difícil para o método *wrapper*, pois este precisa treinar o modelo sempre que um novo subconjunto for selecionado. Na abordagem híbrida, as variáveis são ranqueadas com base em sua relevância e, então, aquelas que apresentam *scores* mais altos são fornecidas ao método *wrapper*, de modo que o número de avaliações necessárias para o método *wrapper* seja menor (reduzindo, assim,a complexidade computacional).

Além disso, observa-se que os métodos híbridos são computacionalmente mais complexos que os métodos *filter*: métodos híbridos combinam *wrapper* e *filter* e têm menos generalidade em comparação com os métodos *filter*, uma vez que utilizam o algoritmo de aprendizado supervisionado no processo de seleção de variáveis.

2.3.3 Critério de parada

Um critério de parada deve ser adotado em processos de seleção de variáveis, definindo-se quando terminar a busca pelo melhor subconjunto de variáveis. Tal critério pode

ser, por exemplo, um número de variáveis a serem selecionadas ou um número máximo de alternativas testadas de forma que o desempenho do classificador ou do tempo de processamento não seja degradado (FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, 2011). Outros critérios podem considerar a conclusão da busca, quando a adição (ou exclusão) subsequente de qualquer variável não produz um subconjunto melhor, ou quando um subconjunto suficientemente bom é selecionado (ou seja, se sua taxa de erro de classificação/predição é menor que a taxa de erro permitida para uma determinada tarefa) (LIU; YU, 2005).

2.3.4 Validação

A validação do modelo reduzido gerado pode se dar pela comparação da medida de erro (em casos de problemas de regressão) ou de acerto (para problemas de classificação) obtida pelo modelo com as variáveis selecionadas em relação à medida obtida quando nenhuma variável é retirada, ou seja, a utilização da totalidade das variáveis disponíveis (LIU; YU, 2005). Dentre as medidas mais comumente usadas para aferir a qualidade de um modeloestão a raiz do erro médio quadrático (*Root Mean Square Error – RMSE*) e erro médio quadrático (*Mean Square Error – MSE*) para problemas de regressão. Já a acurácia (*Acc*), fração de amostras corretamente classificadas,é tradicionalmente usada em modelos de classificação. No entanto, em vários cenários práticos faz-se necessário a definição de critérios mais específicos como, por exemplo, em análises forenses para identificação de autenticidade de medicamentos. Deve-se levar em consideração os impactos de classificar erroneamente uma amostra. Neste caso, avaliar as consequências de deixar chegar ao mercado amostras de medicamentos não autênticas ou descartar amostras autênticas em razão de erros de classificação. Neste sentido, medidas como sensibilidade (*S*) e especificidade (*E*) podem ser utilizadas (ANZANELLO et al., 2015). Entende-se por sensibilidade a fração de casos positivos classificados como positivos, ou seja, quão sensível o classificador é para detectar instâncias da classe positiva (no caso de identificação de autenticidade de medicamentos, a capacidade em classificar corretamente amostras adulteradas) e especificidade trata-seda fração de casos negativos que são classificados como negativos (para o mesmo exemplo, a capacidade em classificar corretamente amostras não adulteradas). Dependendo do estudo é possível atribuir maior importância à sensibilidade ou à especificidade (no exemplo, a consequência mais grave está em classificar uma amostra adulterada em autêntica). Tais medidas são calculadas de acordo com as equações

2.1 à 2.5 respectivamente.Entendemos que vários cenários práticos, incluindo os farmacêuticos e forenses, são favorecidos por critérios mais específicos.Estas medidas são calculadas de acordo com as equações 2.1 a 2.5, respectivamente.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(f_i - y_i)^2} \tag{2.1}$$

$$MSE = \frac{1}{N}\sum_{I=1}^{N}(f_i - y_i)^2 \tag{2.2}$$

$$Acc = \frac{n_{cc}}{N} \tag{2.3}$$

$$\mathrm{S} = \frac{VP}{VP + FN} \tag{2.4}$$

$$\mathrm{E} = \frac{VN}{VN + FP} \tag{2.5}$$

onde Né o número de observações do conjunto de dados, f_i se refere aos valores previstos pelo modelo, y_i aos valores observados, n_{cc} é o número de amostras classificadas corretamente, VP é o número de amostras da classe positiva classificadas como positivas, VN é o número de amostras da classe negativa classificadas como negativa, FN é o número de amostras da classe positiva classificadas como negativa, FP é o número de amostras da classe negativa classificadas como positivas.

2.4 Visão geral dos métodos de seleção de variáveis

A tabela 2.1 apresenta algumas técnicas de seleção de variáveis abordadas na literatura identificando a estratégia de avaliação do método, a estratégia de busca e a medida de importância que o método se baseia.

Tabela 2.1– Métodos de seleção de variáveis apresentados na literatura

Estratégia de avaliação	Estratégia de busca	Medidas de importância	Ref
Filter	*ranking*	dependência	(KIRA; RENDELL, 1992)
	ranking	dependência	(WU et al., 2013)
	backward	distância	(PUDIL; NOVOVIČOVÁ; KITTLER, 1994)
	forward	distância	(PUDIL; NOVOVIČOVÁ; KITTLER, 1994)
Wrapper	*forward*	Informação/precisão	(KAHMANN et al., 2018)
	forward	informação/precisão	(HUANG; LUO; XIA, 2019)
	forward	dependência/precisão	(NØRGAARD et al., 2000)
	backward	dependência/precisão	(LEARDI; NØRGAARD, 2004)
	completa	dependência/precisão	(FERRÃO et al., 2011)
Embedded	*backward*	consistência	(WESTON; ELISSEEFF; SCHÖLKOPF, 2003; ZHANG et al., 2006)
	backward	dependência	(TIBSHIRANI, 1996)
	forward	informação	(KUSWANTO; MUBAROK, 2019)
Híbrido	*backward*	correlação /informação	(ASDAGHI; SOLEIMANI, 2019)
	aleatória	consistência	(ZHANG; XIONG; MIN, 2019)
	aleatória	distância	(ZHANG; XIONG; MIN, 2019)

2.5 Conclusão

O problema da seleção de variáveis está presente no reconhecimento de padrões e tem sido objeto crescente de interesse. Com o aprimoramento das técnicas de aquisição de dados e avanços computacionais, a dimensionalidade dos dados tem aumentado cada vez mais. Assim, o desenvolvimento de técnicas de redução da dimensionalidade torna-se crucial para que os métodos de reconhecimento de padrões possam lidar com amostras compostas por milhares de variáveis extraindo, assim, informações úteis e precisas.

Embora métodos de seleção de variáveis tenham sido amplamente aplicados em diferentes campos e com diversas finalidades, não há um consenso na literatura sobre quais os melhores métodos. Portanto, é aconselhável encontrar um algoritmo adequado para uma determinada aplicação. Por esta razão, seleção de variáveis continua sendo uma das linhas de pesquisa ativamente perseguidas devido à sua importância e influência sobre diversas áreas.

Métodos de seleção de variáveis são categorizados com base na maneira em que as variáveis são combinadas no processo de seleção (subconjuntos de variáveis*ranking* de

variáveis) e com base no grau de dependência do algoritmo de aprendizado (*filter*, *wrapper*, *embedded*, híbrido). Métodos baseados em subconjuntos geram subconjuntos de variáveis usando alguma estratégia de busca (completa, sequencial ou aleatória). Dentre tais estratégias, a busca completa ou exaustiva conduz a uma alta complexidade computacional, uma vez que todos os subconjuntos são testados, resultando em 2^d possibilidades de combinação de subconjuntos para as *d* variáveis existentes. Por esta razão, a estratégia de busca completa não é adequada para espaços de alta dimensão, tornando-se muitas vezes impraticável. No entanto, há maneiras de percorrer o espaço de busca de forma reduzida, seguindo uma sequência (sequencial) ou não (aleatória), neste caso não garantindo a solução ótima, mas ganhando em velocidade de processamento e até viabilidade do processamento. Quando a estratégia de busca é aleatória tem-se ainda como vantagem a maior abrangência em relação aos pontos acessados, permitindo que o algoritmo "escape" de ótimos locais e encontre o ótimo global. Os métodos de seleção de variáveis baseados em subconjuntos que usam a busca sequencial seguem uma abordagem baseada em *wrapper* e utilizam heurísticas (*forward*, *backward*, *stepwise*) onde cada subconjunto gerado precisa desenvolver um modelo de classificação para avaliá-los, o que tende a aumentar a complexidade computacional. Métodos *wrapper* produzem alta precisão de classificação, porém não possuem alta generalidade. Já os métodos baseados em *ranking* de variáveis levam menos tempo de computação e atingem alta generalidade, pois não usam o algoritmo de aprendizado supervisionado. No entanto, métodos baseados em *ranking* não são capazes de remover as variáveis redundantes, uma vez que apenas avaliam a relevância de cada variável para a variável de resposta (classe). Portanto, eles podem ser uma escolha adequada para espaço de alta dimensão com um mecanismo de análise de redundância adequado. Os métodos *embedded* levam vantagem sobre os *wrapper* porque a seleção funciona como parte do processo de aprendizagem, reduzindo o custo computacional. Já abordagem híbrida, visa se beneficiar dos pontos positivos das abordagens *filter* e *wrapper*, ou seja, a eficiência computacional da abordagem *filter* e a precisão preditiva da abordagem *wrapper*.

À medida que estratégias de mineração de dados se expandem para novas áreas, a seleção de variáveis também enfrenta novos desafios. Em vista disso, esta pesquisa forneceu uma visão abrangente de vários aspectos da seleção de variáveis, apresentando a estrutura geral para processos de seleção de variáveis, bem como a categorização de diversas abordagens, a fim de facilitar o entendimento das diversas abordagens existentes e orientar o desenvolvimento de novas abordagens de seleção de variáveis.

2.6 Referências

ANZANELLO, M. J.; KAHMANN, A.; MARCELO, M. C. A.; et al. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, v. 115, p. 562–569, 2015. Elsevier.

ARAUZO-AZOFRA, A.; MOLINA-BAENA, J.; JIMÉNEZ-VÍLCHEZ, A.; LUQUE-RODRIGUEZ, M. Using Individual Feature Evaluation to Start Feature Subset Selection Methods for Classification. **ICAART - 9th International Conference on Agents and Artificial Intelligence**, p. 607–614, 2017.

ASDAGHI, F.; SOLEIMANI, A. An effective feature selection method for web spam detection. **Knowledge-Based Systems**, v. 166, p. 198–206, 2019.

BELLMAN, R. **Adaptive Control Processes: A Guided Tour**. Princeton ed. 1961.

BENIN, G.; DE CARVALHO, F. I. F.; DE OLIVEIRA, A. C.; et al. Comparações entre medidas de dissimilaridade e estatísticas multivariadas como critérios no direcionamento de hibridações em aveia. **Ciência Rural**, v. 33, n. 4, p. 657–662, 2003.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v. 97, p. 245–271, 1997.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. **Feature Selection for High-Dimensional Data.** Springer ed. 2015.

CAI, W.; LI, Y.; SHAO, X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 90, n. 2, p. 188–194, 2008.

CAMARGO, S. D. S. **Um modelo neural de aprimoramento progressivo para redução de dimensionalidade**, 2010. Porto Alegre: Universidade Federal do Rio Grande do Sul (UFRGS).

CILIA, N. D.; DE STEFANO, C.; FONTANELLA, F.; SCOTTO DI FRECA, A. A ranking-based feature selection approach for handwritten character recognition. **Pattern Recognition Letters**, v. 121, p. 77–86, 2019.

COCCHI, M.; BIANCOLILLO, A.; MARINI, F. Chemometric Methods for Classification and Feature Selection. **Comprehensive Analytical Chemistry**. v. 82, p.265–299, 2018.

DE OLIVEIRA JUNIOR, Gilson Medeiros. **Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado**. 2010. - Universidade Federal de Pernambuco (UFPE), Recife, 2010.

FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, A.

C. P. DE L. F. DE. **Inteligência artificial: uma abordagem de aprendizado de máquina**. LTC ed. Rio de Janeiro, 2011.

FERRÃO, M. F.; VIERA, M. D. S.; PAZOS, R. E. P.; 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.

GEURTS, P.; FILLET, M.; DE SENY, D.; et al. Proteomic mass spectra classification using decision tree based ensemble methods. **Bioinformatics Original Paper**, v. 21, n. 15, p. 3138–3145, 2005.

HALL, M. A. **Correlation-based Feature Selection for Machine Learning**, 1999. University of Waikato.

HAMET, P.; TREMBLAY, J. Artificial intelligence in medicine. **Metabolism: Clinical and Experimental**, v. 69, p. S36–S40, 2017.

HASEGAWA, K.; KIMURA, T.; FUNATSU, K. GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. **Journal of Chemical Information and Computer Sciences**, v. 39, n. 1, p. 112–120, 1997.

HORTA, R. A. M.; CARVALHO, F. A. DE; ALVES, F. J. DOS S.; JORGE, M. J. Comparação de Técnicas de Seleção de Atributos para Previsão de Insolvência de Empresas Brasileiras no Período 2005-2007. EnANPAD - XXXIV Encontro da ANPAD. **Anais...** , 2010. Rio de Janeiro.

HUANG, X.; LUO, Y. P.; XIA, L. An efficient wavelength selection method based on the maximal information coefficient for multivariate spectral calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 194, 2019.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018.

KIRA, K.; RENDELL, L. A. The feature selection problem: traditional methods and a new algorithm. **Aaai**, v. 2, p. 129–134, 1992.

KITTLER, J. Feature set search algorithms. **Pattern Recognition and Signal Processing**, 1978.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1–2, p. 273–324, 1997.

KOTU, V.; DESHPANDE, B. Chapter 12 – Feature Selection. **Predictive Analytics and Data Mining**. p.347–370, 2015.

KUCHERYAVSKIY, S. Package "mdatools": Multivariate Data Analysis for Chemometrics. , 2019.

KUHN, M.; WING, J.; WESTON, S.; et al. Package ‘caret’: Classification and Regression Training. , 2020.

KUSWANTO, H.; MUBAROK, R. Classification of Cancer Drug Compounds for Radiation Protection Optimization Using CART. **Procedia Computer Science**, v. 161, p. 458–465, 2019.

LEARDI, R.; NØRGAARD, L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. **Journal of Chemometrics**, v. 18, n. 11, p. 486–497, 2004.

LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**, 2005. São Paulo: Universidade de São Paulo (USP).

LEE, P. Y.; LOH, W. P.; CHIN, J. F. Feature selection in multimedia: The state-of-the-art review. **Image and Vision Computing**, v. 67, p. 29–42, 2017.

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.

NADLER, B.; COIFMAN, R. R. The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration. **Journal of Chemometrics**, v. 19, n. 2, p. 107–118, 2005.

NØRGAARD, L.; SAUDLAND, A.; WAGNER, J.; et al. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. **Applied Spectroscopy**, v. 54, n. 3, p. 413–419, 2000.

PARMEZAN, A. R. S.; LEE, H. D.; SPOLAÔR, N.; CHUNG, W. F. **Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados**. Foz do Iguaçu, 2012.

PEREIRA, R. B. **Seleção Lazy de atributos para a tarefa de classificação**, 2009. Niterói: Universidade Federal Fluminense.

PUDIL, P.; NOVOVIČOVÁ, J.; KITTLER, J. Floating search methods in feature selection. **Pattern Recognition Letters**, v. 15, n. 11, p. 1119–1125, 1994.

R CORE TEAM. R: A Language and Environment for Statistical Computing. , 2018. Vienna, Austria.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in Biology and Medicine**, v. 112, p. 103375, 2019.

ROMANSKI, P.; KOTTHOFF, L. Package “FSelector”: Selecting Attributes. , 2018.

SABANDO, M. V.; PONZONI, I.; SOTO, A. J. Neural-based approaches to overcome feature selection and applicability domain in drug-related property prediction. **Applied Soft Computing Journal**, v. 85, 2019.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics Review**, v. 23, p. 2507–2517, 2007.

SALIMI, A.; ZIAII, M.; AMIRI, A.; et al. Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification. **Egyptian Journal of Remote Sensing and Space Science**, v. 21, n. 1, p. 27–36, 2018.

SETIONO, H. L. AND R. Feature Selection And Classification -A Probabilistic Wrapper Approach. **IEA/AIE**, p. 419- 424., 1997.

SHARMA, A.; AMARNATH, · M; PAVAN, ·; KANKAR, K. Novel ensemble techniques for classification of rolling element bearing faults. **Journal of the Brazilian Society of Mechanical Sciences and Engineering**, v. 39, p. 709–724, 2017.

SOARES, F.; ANZANELLO, M. J.; FOGLIATTO, F. S.; et al. Element selection and concentration analysis for classifying South America wine samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 150, p. 33–40, 2018.

TERUYA, A. **Uma nova metodologia para seleção de atributos no processo de extração de conhecimento de base de dados baseada na teoria de rough sets**, 2008. Campo Grande: Universidade Federal do Mato Grosso do Sul.

THOMAS, E. V. A Primer on Multivariate Calibration. **Analytical Chemistry**, v. 66, n. 15, p. 795–804, 1994.

TIBSHIRANI, R. Regression Shrinkage and Selection Via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

URBANOWICZ, R. J.; MEEKER, M.; LA CAVA, W.; OLSON, R. S.; MOORE, J. H. Relief-based feature selection: Introduction and review. **Journal of Biomedical Informatics**, v. 85, p. 189–203, 2018.

VASCONCELOS, B. F. B. DE. **Poder preditivo de métodos de Machine Learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países.**, 2017. Brasília: Universidade de Brasília (UnB).

VERGARA, J. R.; ESTÉVEZ, P. A. A review of feature selection methods based on mutual information. **Neural computing and applications**, v. 24, n. 1, p. 175–186, 2014.

VICINI, L. **Análise multivariada da teoria à prática**, 2005. Santa Maria: Universidade Federal de Santa Maria.

WAN, C. **Hierarchical Feature Selection for Knowledge Discovery: Application of Data mining to the biology of ageing**. Springer ed. 2018.

WESTON, J.; ELISSEEFF, A.; SCHÖLKOPF, B. Use of the Zero-Norm with Linear Models and Kernel Methods. **Journal of Machine Learning Research**, v. 3, p. 1439–1461, 2003.

WU, B.; ABBOTT, T.; FISHMAN, D.; et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. **Bioinformatics**, v. 19, n. 13, p. 1636–1643, 2003.

WU, S.-D.; WU, C.-W.; WU, T.-Y.; WANG, C.-C. Multi-Scale Analysis Based Ball Bearing Defect Diagnostics Using Mahalanobis Distance and Support Vector Machine. **Entropy**, v. 15, n. 2, p. 416–433, 2013.

XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J. W.; HOLMES, M.; HANPIN, M. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, v. 667, n. 1–2, p. 14–32, 2010.

YAMASHITA, G. H. **Abordagens multivariadas para a seleção de variáveis com vistas à caracterização de medicamentos**, 2015. Porto Alegre: Universidade Federal do Rio Grande do Sul.

YAN, X.; NAZMI, S.; EROL, B. A.; et al. An Efficient Unsupervised Feature Selection Procedure Through Feature Clustering. **Pattern Recognition Letters**, v. 131, p. 277–284, 2020.

YU, L.; LIU, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. **Journal of Machine Learning Research**, v. 5, p. 1205–1224, 2004.

ZAKI, F. A. M.; ZULKURNAIN, N. F. RARE: Mining colossal closed itemset in high dimensional data. **Knowledge-Based Systems**, v. 161, p. 1–11, 2018.

ZHANG, J.; XIONG, Y.; MIN, S. A new hybrid filter/wrapper algorithm for feature selection in classification. **Analytica Chimica Acta**, v. 1080, p. 43–54, 2019.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, v. 128, p. 17–24, 2013.

ZHANG, X.; LU, X.; SHI, Q.; et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. **BMC Bioinformatics**, v. 7, n. 1, p. 197, 2006.

ZHAO, Z. A.; LIU, H. **Spectral feature selection for data mining**. Taylor & F ed. 2011.

3. Segundo artigo:Seleção de comprimentos de onda para predição de propriedades do biodiesel/diesel apoiada na lei de Lambert-Beer

Resumo

Dados espectrais descrevendo propriedades de amostras de produtos tipicamente são compostos por grande número de comprimentos de onda (COs) ruidosos e altamente correlacionados, o que tende a reduzir o desempenho de técnicas preditivas aplicadas a tais dados. O objetivo deste artigo é propor um método de seleção de COs com vistas à predição de propriedades de biodiesel baseado em conceitos da lei de Lambert-Beer, a qual é utilizada para ranquear os COs de acordo com sua relevância. A ferramenta preditiva utilizada é a rede neural artificial (*Artificial Neural Network* – ANN). Para este fim, foram analisados três bancos de dados espectrais de diesel descritos por espectroscopia no infravermelho próximo (NIRS). Tais dados foram usados para fazer predições de propriedades do combustível que impactam na determinação da sua qualidade. O desempenho de predição do método foi avaliado por intermédio do erro quadrático médio (*Root Mean Square Error* – RMSE) no conjunto de teste. O método proposto reduziu significativamente o número de COs e melhorou o desempenho preditivo, quando comparado à utilização da totalidade de COs disponíveis. O método proposto ainda superou abordagens da literatura voltadas à seleção de COs.

Palavras-Chave: ANN; dados espectrais; diesel; seleção de variáveis; comprimentos de onda.

3.1 Introdução

No Brasil, o óleo diesel é o derivado de petróleo mais consumido por conta de sua predominância no transporte rodoviário de passageiros e de carga. A grande demanda do mercado tem instigado esforços por lucros ilegais por meio de adulterações deste produto. Os efeitos da falta de conformidade deste combustível estão fortemente ligados ao desempenho dos motores e ao aumento nas emissões de poluentes (CÂMARA et al., 2017; CUNHA et al., 2019; KRISHNA et al., 2019; LI et al., 2005). Cunha et al. (2019) enumeram diversos problemas oriundos da adulteração de combustíveis: falha repentina do motor, dificuldade de partida, aumento do consumo de combustível, baixa taxa de pulverização de combustível na câmara de combustão e aumento das emissões de material particulado.

Ponto de ebulição, número de cetano e ponto de fulgor são algumas das propriedades físico-químicas utilizadas pela indústria para monitorar a qualidade do diesel e garantir a queima adequada nos motores (ALEME; BARBEIRA, 2012; LI et al., 2005). Nos últimos anos, técnicas analíticas como espectroscopia no infravermelho próximo (*Near-Infrared Spectroscopy*– NIR) e transformada de Fourier (*Fourier-transform infrared*– FTIR) foram amplamente adotadas como ferramentas analíticas em diferentes campos e com diversas finalidades (SOARES et al., 2017). Estas técnicas podem ser usadas para identificar um composto ou investigar a composição de uma amostra, a fim de identificar possíveis adulterações (CÂMARA et al., 2017; GAYDOU; KISTER; DUPUY, 2011). Palou et al. (2017) descrevem a técnica como um atrativo recurso para a indústria do petróleo, uma vez que contribui para o alcance da qualidade exigida em seus produtos.

Entretanto, técnicas analíticas apoiadas em espectroscopia usualmente geram conjuntos de dados compostos por centenas ou milhares de comprimentos de onda (COs) altamente correlacionados. Quando tais dados são usados como preditores em modelos matemáticos, percebe-se redução no desempenho preditivo de técnicas multivariadas no que diz respeito à velocidade e taxa de acerto (KAHMANN et al., 2018). Cada variável passa a ser vista como uma coordenada *d*-dimensional no espaço; à medida que o número de dimensões aumenta, as instâncias ficam mais dispersas no espaço. Quando se tem menos instâncias (amostras) por região, a tarefa de aprendizado se torna mais difícil. Isso ocorre porque os algoritmos de aprendizado de máquina constroem preditores com base nas proporções estimadas de instância em cada classe por regiões do espaço. Diante disso, ocorre um aumento da demanda por técnicas de seleção de COs que permitam executar a análise dos dados sem a perda de informações relevantes, diminuindo o risco de inferências não confiáveis e do custo computacional de análises (GAUCHI; CHAGNON, 2001).

Abordagens baseadas em diferentes técnicas estatísticas têm sido propostas para identificar o melhor subconjunto de variáveis com o objetivo de classificar ou prever propriedades de uma amostra. Yun et al. (2019) generalizaram os métodos de seleção de variáveis de uma maneira simples para apresentar suas classificações, vantagens e desvantagens. Os autores apresentaram tanto métodos de seleção de pontos de COs como métodos de seleção de intervalos de COs. Lin et al. (2018) propuseram um novo método de seleção de variáveis para PLS (Mínimos Quadrados Parciais) com a contribuição variável

ponderada (PLS-WVC) para o primeiro valor singular da matriz de covariância para cada componente do PLS. O método PLS-WVC proposto é integrado ao *i*PLS para selecionar os intervalos de COs informativos para modelagem espectroscópica. A utilidade do método proposto foi demonstrada com dois conjuntos de dados espectrais reais para amostras de cerveja e milho, gerando resultados superiores quando comparado a outros métodos de seleção de variáveis. Lavoie et al. (2019)apresentaram uma nova metodologia PLS não linear (NL) implementando um processo iterativo. Neste processo, as relações NL são modeladas iterativamente com *splines* cúbicos por partes restritas e os pesos das variáveis são iterativamente calculados através de uma modificação da PLS. Em seguida, os autores compararam as variantes mais comuns de NL-PLS com a metodologia proposta de NL-PLS usando três estudos de caso, cada um com características específicas. Já Kahmann et al. (2018) propuseram um método para selecionar os COs mais relevantes a serem incluídos nos modelos de regressão usando a Programação Quadrática (QP) para prever a concentração de cocaína e adulterantes em amostras apreendidas.

Alinhado com os objetivos acima, este artigo apresenta um novo método para determinação das regiões mais relevantes do espectro com vistas à predição de propriedades de amostras. O método propõe o ranqueamento dos COs baseado na lei de *Lambert-Beer*, o que reduz significativamente o número de possibilidades de combinações de variáveis a serem testadas em cada subconjunto (uma vez que, para n variáveis, tem-se $2^n - 1$ combinações possíveis). A Lei de *Lambert-Beer* estabelece uma relação linear entre a absorbância e a concentração de substâncias químicas presentes na amostra (GOBRECHT et al., 2015). A absorbância é uma medida da "quantidade" de luz absorvida pela amostra, sendo influenciada por variáveis como natureza do solvente, pH da solução, temperatura, concentração de eletrólitos e presença de substâncias interferentes. Por esse motivo, quando dados de origem espectral com o objetivo de prever características como concentração e temperatura são analisados, buscam-se COs com maior variabilidade para explicar a variável dependente. De tal forma, em vez de analisar os dados (amostras) em sua totalidade, este artigo propõe considerar apenas duas faixas das amostras: o intervalo que contém as observações cuja variável dependente apresenta os menores valores, e aquele cujas observações apresentam variável dependente com os maiores valores. As diferenças entre tais faixas são usadas na geração de um índice de importância dos COs. Na sequência, subconjuntos de variáveis são gerados usando a inserção *forward*, uma de cada vez, iniciando a partir de um conjunto vazio, de acordo com

uma estratégia de seleção de busca sequencial. Nesta estratégia, as variáveis são adicionadas sequencialmente a um conjunto de candidatas de acordo com a ordem sugerida pelo índice de importância até que a adição de outras variáveis não melhore o critério de avaliação (YUN et al., 2019). Nas proposições deste artigo, ANN são usadas como ferramenta de predição. Finalmente, o método seleciona as regiões mais apropriadas do espectro para predição de três importantes propriedades do diesel: ponto de ebulição, número de cetano e ponto de fulgor.

Embora a literatura ofereça várias estruturas para identificar as variáveis mais informativas, muitas delas normalmente não consideram a não linearidade das relações entre as variáveis. De acordo com Lee et al. (2015), as métricas comumente usadas em quimiometria são baseadas principalmente em PLS. Embora o PLS seja um método robusto usado normalmente para propósitos preditivos ou monitoramento de controle, apresenta uma desvantagem importante em relação à forma da equação preditiva resultante. O PLS tradicional produz apenas modelos preditivos lineares, que podem produzir desempenhos preditivos ruins na presença de uma forte relação não linear (NL) entre variáveis independentes e dependentes (LAVOIE; MUTEKI; GOSSELIN, 2019). Por outro lado, as ANNs podem modelar várias variáveis e suas relações não lineares, além de resolver problemas complexos (HAYKIN, 2001). Além disso, também encontram-se na literatura métodos que se baseiam em técnicas de otimização, como Kahmann et al. (2018). Em geral, os métodos de otimização apresentam dificuldades relacionadas à não convergência, existência de vários pontos locais ótimos da função objetivo e aumento do tempo computacional (TELES, MATEUS LEMBI; GOMES, 2010). O método proposto inova ao tomar como base conceitos advindos da química. Desse modo, desconsidera as amostras com valores intermediários na composição do índice de importância de COs. Isto torna a abordagem proposta mais eficiente e direta, além de gerar um índice com maior poder de separação. Além disso, utiliza ANN para a geração dos subconjuntos tornando o método capaz de se adaptar bem a presença de não-linearidades nos dados.

3.2 Material e métodos

3.2.1 Bases de dados espectrais

Três bancos de dados foram utilizados a fim de avaliar o desempenho da abordagem proposta. Todos os bancos utilizados consistem em espectros NIR de combustível diesel,

incluindo várias propriedades do combustível que impactam na determinação da sua qualidade. As amostras coletadas advêm de diferentes origens.

O número de cetano foi determinado de acordo com a norma ASTM D613, o ponto de ebulição determinado de acordo com a norma ASTM D86 e o ponto de fulgor determinado pelo método padrão brasileiro ABNT NBR 14598, definido pela ANP (Agência Nacional Brasileira de Petróleo, Gás Natural e Biocombustíveis), com valores entre 47,0 e 79,5 °C. Os bancos que tratam do número de cetano e do ponto de ebulição são de domínio público e provêm de um estudo financiado pelo exército americano sobre propriedades de diesel conduzido pelo Southwest Research Institute (SOARES; ANZANELLO, 2018). O banco de dados referente ao ponto de fulgor vem de um estudo de Ferrão et al. (2011), que testa modelos de calibração PLS e seus variantes para a seleção de COsHATR – FTIR para prever parâmetros de qualidade das misturas de biodiesel / diesel. Os pontos de fulgor foram determinados de acordo com o método padrão brasileiro ABNT NBR 14598,definido pela ANP, utilizando o equipamento (Pensky Martens - ASTM D93) Petrotest, modelo PMA4. Os valores estão entre 47,0 e 79,5 °C. A tabela 3.1 apresenta uma síntese de cada banco, assim como uma breve descrição.

Tabela 3.1– Bancos de dados de espectroscopia de infravermelho próximo usados neste estudo.

Banco	COs	Nº de amostras	Descrição da VR	Refs	Fonte
Ponto de ebulição (ºC)	401	113	Temperatura na qual o diesel muda do estado líquido para o gasoso em toda a sua massa	(SOARES; ANZANELLO, 2018)	http://www.eigenvector.com/data/
Número de cetano	401	113	Indicador da velocidade de combustão do combustível diesel e da compressão necessária para a ignição.	(SOARES; ANZANELLO, 2018)	http://www.eigenvector.com/data/
Ponto de fulgor (ºC)	1738	85	Temperatura na qual o vapor de combustível se inflamará em forma de flash quando entrar em contato com o fogo.	(FERRÃO et al., 2011)	-

3.2.2 Fundamentos de Redes Neurais Artificiais (ANN)

Uma estrutura típica de um modelo ANN é composta por uma camada de entrada, que apresenta os padrões à rede (variáveis que alimentam o modelo), *n* camadas escondidas consideradas como extratoras de características, e uma camada de saída onde o resultado final é apresentado. Cada camada consiste de um número de neurônios conectados a todos os neurônios da camada seguinte e cada conexão tem um peso numérico w_{ij} atribuído, onde *i* identifica o neurônio da camada seguinte ao qual o neurônio *j* está ligado (ANTONOPOULOS et al., 2019; WU; DANDY; MAIER, 2014).

O número de camadas e a maneira como os neurônios artificiais são agrupados em cada uma destas são características que definem a arquitetura de uma ANN. Não existe uma metodologia estabelecida para a definição da arquitetura (ANTONOPOULOS et al., 2019; JAIN; NAYAK; SUDHEER, 2008). Em geral, experimentos com ANN são realizados através de repetidos testes com diferentes topologias até serem obtidos resultados satisfatórios. Após estabelecer a arquitetura da rede, uma constante, denominada Bias, é inserida no modelo, sendo considerada pesos de uma entrada extra (IGEL; HÜSKEN, 2000), a fim de ajudá-lo a se adaptar melhor aos dados fornecidos.

O processamento em cada neurônio se dá pela função de ativação, definida como a transformação linear ou não linear feita ao longo do sinal de entrada. Ela basicamente decide se a informação que o neurônio está recebendo é relevante ou deve ser ignorada, determinando assim, a saída dos neurônios (JIMENEZ-MARTINEZ; ALFARO-PONCE, 2019). Neste estudo, a função de ativação escolhida foi a função sigmoide ou (logística) devido ao seu amplo uso em redes multicamadas e em redes com sinais contínuos. A inicialização dos pesos ocorre de forma aleatória e com valores diferentes de zero. Dentre os algoritmos de aprendizado existentes, o resiliente *backpropagation* (RPROP) foi o utilizado neste estudo por conta de sua rapidez computacional (PRASAD; SINGH; LAL, 2013). Este algoritmo funciona de maneira similar ao *backpropagation* tradicional, o qual se baseia na retro propagação dos erros para realizar os ajustes de pesos das camadas intermediárias. Objetiva otimizar os pesos sinápticos para que a rede neural possa aprender associações ente as entradas e as saídas (KRISHNAKUMAR, 1993).

A diferença entre esses dois algoritmos é que, no RPROP, a atualização de pesos é feita de forma independente por meio de uma taxa de aprendizado dinâmico. A taxa de aprendizado é atualizada para cada conexão de neurônio, reduzindo o erro independentemente para cada neurônio. Em outras palavras, cada peso tem seu próprio valor de atualização, que muda com o tempo (DE ALMEIDA, 2003; OLIVEIRA; BARBAR; SOARES, 2015). Os novos valores de atualização individuais Δij são determinados pelo algoritmo RPROP de acordo com a equação (3.1).

$$\Delta_{ij}{}^{(t)} = \begin{cases} \eta^{+} \times \Delta_{ij}{}^{(t-1)}, \text{if } \dfrac{\partial \mathrm{E}^{(t-1)}}{\partial w_{ij}} \times \dfrac{\partial \mathrm{E}^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} \times \Delta_{ij}{}^{(t-1)}, \text{if } \dfrac{\partial \mathrm{E}^{(t-1)}}{\partial w_{ij}} \times \dfrac{\partial \mathrm{E}^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}{}^{(t-1)}, \quad if\ else \end{cases} \qquad 0 < \eta^{-} < 1 < \eta^{+} \qquad (3.1)$$

O algoritmo estabelece que toda vez que a derivada parcial atual ($\partial E^{(t)}/\partial \mathrm{w}_{ij}$) e a anterior ($\partial E^{(t-1)}/\partial \mathrm{w}_{ij}$) do peso correspondente $\partial \mathrm{w}_{ij}$obtiverem o mesmo sinal, o valor de atualização é ligeiramente aumentado, por meio de um fator de aumento η^{+} , para acelerar a convergência em regiões superficiais. Caso estas derivadas obtenham sinais diferentes, indicando que a última atualização foi muito grande e o algoritmo saltou sobre um mínimo local, o valor de atualização $\Delta_{ij}{}^{(t)}$ é diminuído pelo fator de diminuição η^{-}. Neste último caso, não deve haver adaptação na etapa de aprendizagem subsequente, fazendo com que o valor de atualização retorne ao estado anterior. Porém, ao reverter o valor do peso, a derivada anterior também precisa ser alterada, caso contrário, quando o peso for atualizado novamente, ele aplicará as mesmas alterações. Por esta razão, a derivada anterior é definida como zero ($\partial E^{(t-1)}\ \partial \mathrm{w}_{ij} = 0$) (ANASTASIADIS; MAGOULAS; VRAHATIS, 2005; IGEL; HÜSKEN, 2003; RIEDMILLER; RPROP, 1994).

3.3 Método para seleção dos COs mais informativos para predição das propriedades de amostras de diesel

O método proposto compreende quatro passos principais: (*i*) pré-processamento dos dados com o ajuste de uma escala única para todas as variáveis, (*ii*) divisão do banco de dados em conjuntos de treino e teste, (*iii*) ranqueamento dos COs e (*iv*) Seleção iterativa dosCOs mais relevantes. Estes passos são ilustrados na figura3.1e detalhados em seguida.

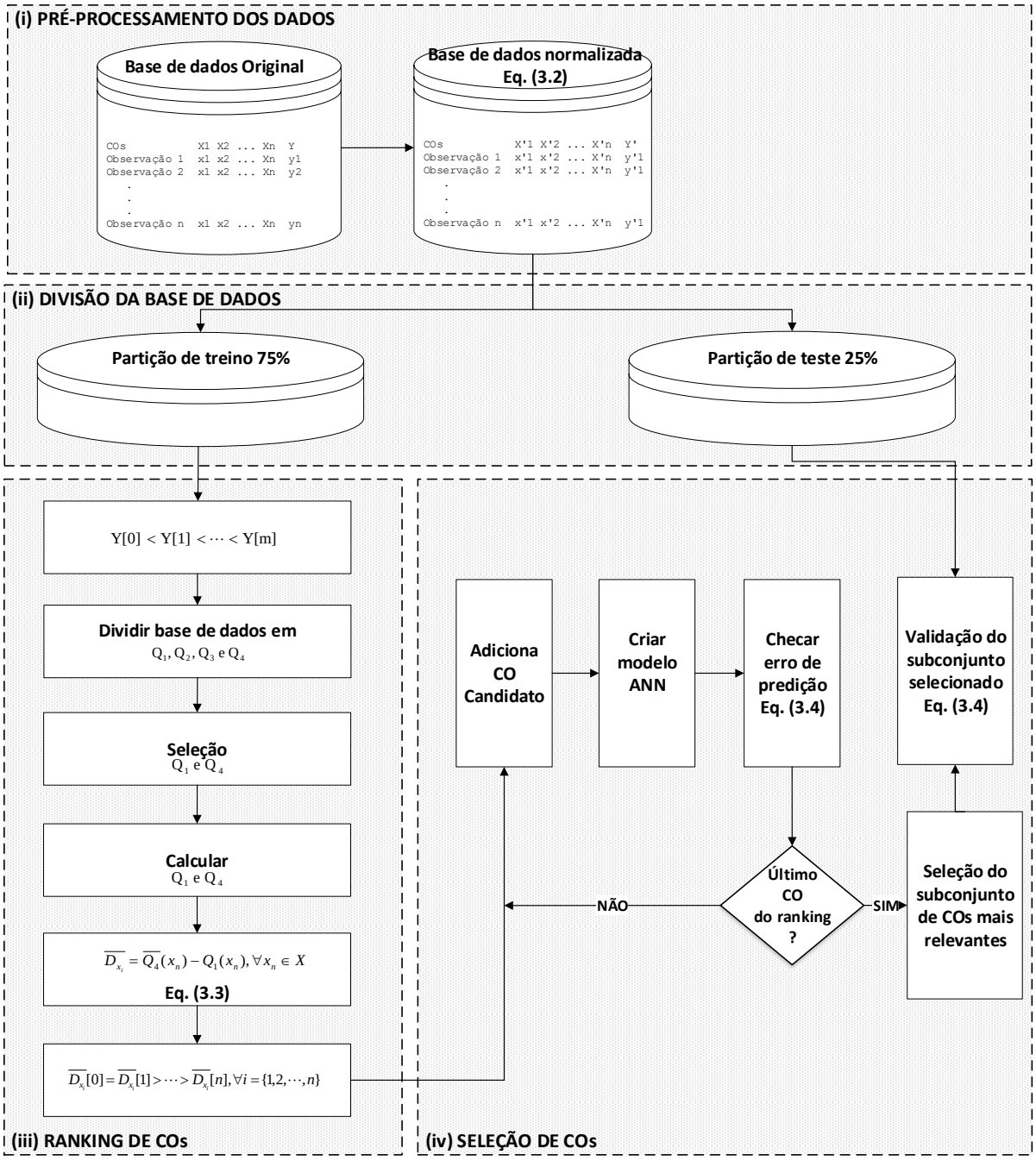

Figura 3.1– Método proposto

3.3.1 Pré-processamento dos dados

Todos os bancos foram previamente normalizadosutilizando o método do mínimo-máximo [0,1], conforme a equação (3.2).

$$x_{norm} = \frac{(x - x_{\min})}{(x_{max} - x_{\min})} \tag{3.2}$$

onde x_{norm} é o valor escalonado correspondente ao valor original x, e $x_{\min}$ e x_{max} são os valores mínimo e máximo de cada CO. A normalização dos dados serve para evitar que o algoritmo interprete níveis de importância errôneos para cada variável em razão de sua magnitude.

3.3.2 Divisão do banco de dados em conjuntos de treino e teste

O conjunto de dados original é particionado em dois conjuntos: treino e teste, onde 75% das amostras são inseridas aleatoriamente no conjunto de treino e os 25% restantes no conjunto de teste. Tal particionamento é realizado 200 vezes, gerando subconjuntos aleatórios de treino e teste.

O conjunto de treino é utilizado para ranquear os COs e selecionar os subconjuntos de COs de maior relevância, enquanto que o conjunto de teste avalia o desempenho do modelo ANN do subconjunto selecionado.

3.3.3 Ranqueamento dos COs de acordo com a sua relevância

De acordo com os conceitos da lei de *Lambert Beer*, infere-se que os COs que apresentam maiores variações podem auxiliar na explicação da variabilidade da variável dependente (propriedade em análise). Para identificar os COs com maior variabilidade, ordenam-seas observações (amostras) em ordem crescente em relação à variável dependente y; os dados são então divididos em quatro quartis (Q1, Q2, Q3 e Q4). Na sequência, seleciona-se o primeiro e o quarto quartil, uma vez que estes apresentam os valores extremos da variável dependente (e, de acordo com a lei de *Lambert-Beer*, as diferenças desses valores estão associadas a diferenças nas variáveis independentes – os COs que descrevem cada amostra). Calcula-seentão a média de cada CO em Q1 e Q4 e, em seguida, calcula-se a diferença entre tais médias, conforme a equação (3.3).

$$\bar{D}_{x_n} = \bar{Q}_4(x_n) - \bar{Q}_1(x_n)\ , \forall\ x_n\ \in X \tag{3.3}$$

onde $\overline{D}_{x_i}$ é a diferença entre médias de quartis da variável x_i e $\bar{Q}_4(x_i)$ e $\bar{Q}_1(x_i)$ são as médias do quartil 4 e do quartil 1 para a mesma variável x_i. Dessa maneira, gera-se um vetor $\overline{D}$ que contém as diferenças médias entre os quartis extremos de cada CO. Os elementos do vetor são então colocados em ordem decrescente (dando origem a um índice de importância); variáveis com maiores distâncias são entendidas como mais relevantes na explicação da variabilidade da variável dependente.

3.3.4 Seleção dosCOsmais relevantes para predição da variável dependente

A partir do *ranking* gerado anteriormente, COs são inseridos um a um em modelo ANN. Para isso alguns parâmetros associados ao ANN devem ser estabelecidos: arquitetura da rede, tipo de função de ativação (φ) e algoritmo de aprendizado. Após a escolha do algoritmo de aprendizado, deve-se ainda ajustar os parâmetros referentes a este. No caso desta pesquisa, tais parâmetros são fator de diminuição η^- e fator de aumento η^+.

O desempenho de predição do modelo ANN é avaliado no conjunto de teste por meio da métrica raiz do erro quadrático médio (RMSE), medida de erro amplamente reportada na literatura, conforme a equação (3.4). A cada inserção de uma nova variável no subconjunto, o RMSE é reavaliado.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(Y_i - D_i)^2} \tag{3.4}$$

onde N é o número de observações da amostra, Y_i se refere aos valores estimados e D_i aos valores reais.

Por fim, o subconjunto que apresentar o menor erro médio (RMSE) será o selecionado. Porém, quando a diferença entre as médias dos erros de dois ou mais subconjuntos for significativamente pequena, verificam-se os valores do desvio-padrão; o subconjunto que obtiver um erro reduzido com menor variabilidade (sugerindo maior robustez) será selecionado.

3.4 Resultados numéricos e discussão

Os experimentos computacionaisforam realizados no software RStudio versão Microsoft R Open 3.5.3 (R CORE TEAM, 2018) e os códigos desenvolvidos pelos autores, fazendo uso do pacote neuralnet (FRITSCH; GUENTHER, 2016) do R.

Repetidos ensaios com topologias distintas foram realizados para geração do modelo de ANN. Verificou-se que o erro do modelo sofria pequenas mudanças, porém com tempos de execução bem distintos, levando a altos custos computacionais dependendo da topologia escolhida. Porém, para o objetivo deste estudo (seleção de variáveis), o resultado se mantinha o mesmo independente das arquiteturas testadas. Sendo assim, a construção do modelo contou com 2 camadas escondidas, a primeira contendo 15 neurônios e a segunda 3 neurônios. Para a escolha da função de ativação, testou-se a função tangente hiperbólica e a função sigmoide (logística); os melhores resultados foram obtidos via função sigmoide. Quanto ao algoritmo de aprendizado, utilizou-se o *resilient backpropagation* (RPROP), com um fator de diminuição $\eta^- = 0.5$ e um fator de aumento $\eta^+ = 1.2$. Isso significa que, se o sinal do gradiente não mudar, aumenta-se o valor de atualização de peso em 20%. Caso contrário, se o sinal do gradiente mudar (a atualização passa por um mínimo), diminui-se o valor de atualização de peso pela metade. Riedmiller (1994) afirma que a escolha de $\eta^- = 0.5$ e $\eta^+ = 1.2$ possui potencial de geração de bons resultados. Para avaliar a robustez do método, foram realizadas 200 replicações para cada subconjunto analisado. Os parâmetros utilizados para a aplicação da ANN estão listados na tabela 3.2.

Tabela 3.2– Parâmetros do modelo ANN

Arquitetura (*i* = input; *o* = output)	$i - 15 - 3 - o$
Função de ativação (φ)	Sigmoide
Algoritmo de aprendizagem	RPROP
Fator de diminuição (η^-)	0.5
Fator de aumento (η^+)	1.2
Inicialização dos pesos (w_{ij})	Aleatória
Replicações	200

3.4.1 Resultados da seleção dos COs

Calculou-se o RMSE médio das 200 replicações, gerando uma medida de precisão preditiva mais confiável do que o resultado de uma única partição do banco de dados em treino e teste. O aumento da confiabilidade do resultado é alcançado com a amostragem aleatória pois, além de representar a diversidade de uma população avaliando os modelos gerados por cada um dos *n* conjuntos de treino e teste, este é um método que permite a aplicação de procedimentos de inferência estatística, o que gera maior segurança nas análises (OLIVEIRA; GRÁCIO, 2005).

As figuras 3.2 a 3.4 apresentam os gráficos gerados para cada banco de dados associando o RMSE médio do conjunto de treinode cada subconjunto de variáveis e a quantidade de COs retidos. Nestes gráficos, o RMSE é representado pelo eixo vertical e o número de COs pelo eixo horizontal. Percebe-se comportamento similar dos perfis para os 3 bancos analisados:o erro de predição diminui significativamente quando COs de menor relevância são retirados do modelo. Porém, a eliminação demasiada de variáveis reduz a habilidade preditiva do modelo, aumentando, consequentemente, o seu erro (como visto nas figuras 3.2 e 3.4).

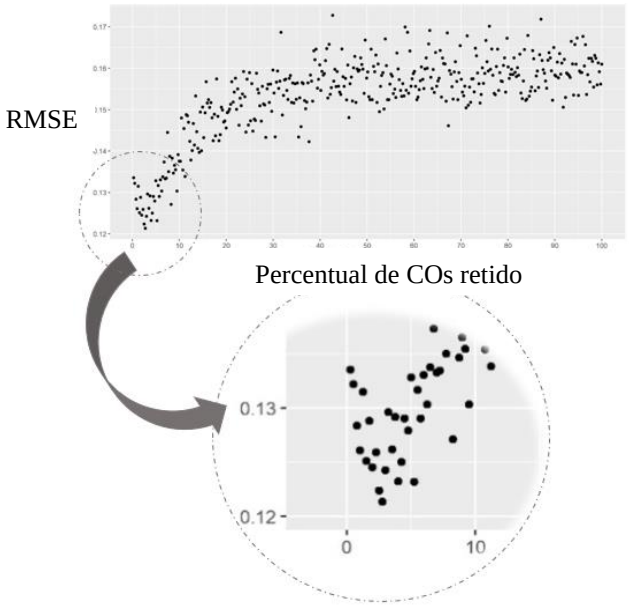

Figura 3.2– Perfil do erro por quantidade de COs retidos para o Ponto de ebulição

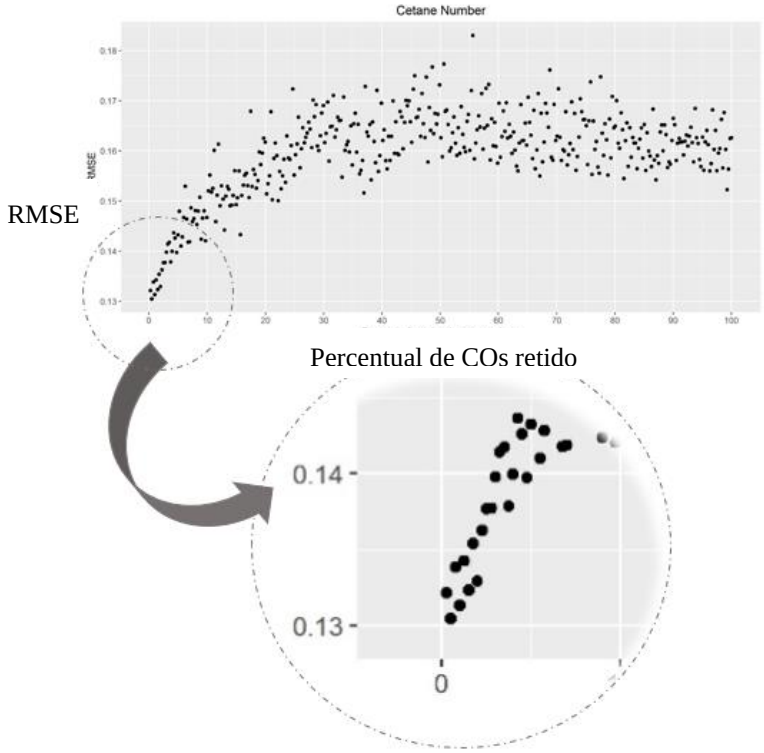

Figura 3.3– Perfil do erro por quantidade de COs retidos para o Cetane Number

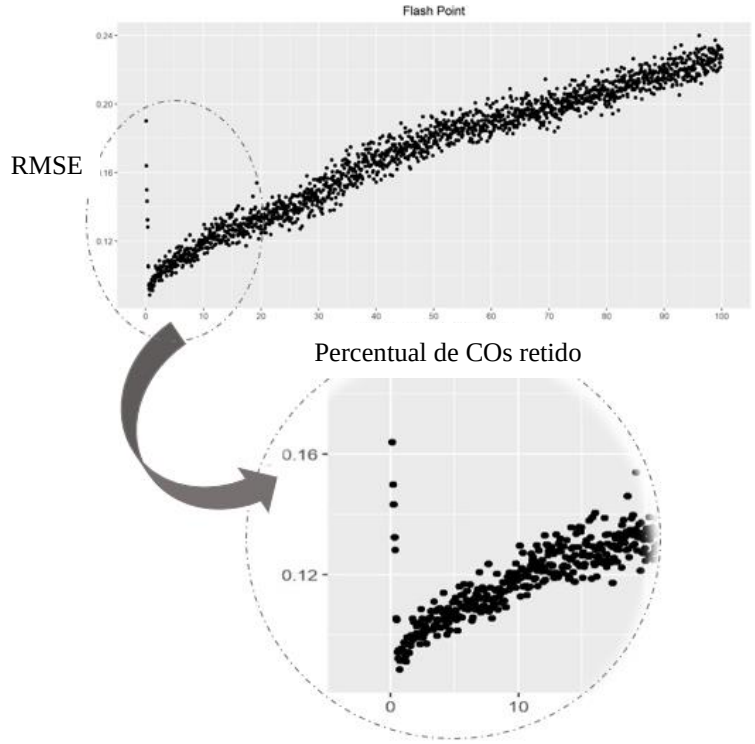

Figura 3.4– Perfil do erro por quantidade de COs retidos para o Flash Point

A tabela 3.3 apresenta os três melhores resultados obtidos para o conjunto de teste em cada banco de dados analisado com destaque para o número e o percentual de COs retidos, o RMSE médio e o desvio-padrão. O melhor resultado do banco ponto de ebulição reteve cerca de 2,74% dos COs. Para o número de cetano, embora o menor RMSE médio seja obtido quando 0,5% das variáveis são retidas, o conjunto composto por cerca de 1% destas apresentou incremento de 0,68% no erro médio, porém um desvio-padrão do erro menor; por esta razão, este último foi o conjunto selecionado como o melhor resultado deste banco de dados. Já para o ponto de fulgor, reteve-se 0,69% das variáveis.

Tabela 3.3– Desempenho do modelo para conjunto de teste

Descrição	Nº de COs retidos	% COs retidos	RMSE Médio	Sd
Ponto de ebulição	**11**	**2.7431**	**0.1214**	**0.0306**
	10	2.4938	0.1224	0.0329
	21	5.2369	0.1232	0.0409
Número de cetano	2	0.4988	0.1305	0.0208
	4	**0.9975**	**0.1314**	**0.0176**
	1	0.2494	0.1321	0.0190
Ponto de fulgor	**12**	**0.6904**	**0.0886**	**0.0399**
	21	1.2083	0.0912	0.0343
	15	0.8631	0.0915	0.0354

3.4.2 Comparação com outros métodos de seleção de COs da literatura

Os resultados da abordagem proposta foram comparados com os resultados das pesquisas de Anzanello et al. (2015) e Soares e Anzanello (2018). Esses autores utilizaram as bases de dados analisadas neste trabalho e a mesma medida para avaliação do desempenho dos modelos, RMSE.

Anzanello et al. (2015) propuseram um método que aplica *Partial Least Squares* (PLS) *regression* no espectro e quatro índices de importância derivados de parâmetros do PLS para eliminar iterativamente COsruidosos e irrelevantes na base de dados Ponto de fulgor. Os autores apresentaram duas abordagens para a seleção do subconjunto de COs após o procedimento de eliminação iterativa das variáveis. Uma delas é o Mínimo RMSE (MR), em que o subconjunto selecionado é aquele que apresentar o menor RMSE, e a outra é a Distância Euclidiana Mínima (MED), onde um gráfico associando o RMSE calculado após a remoção de cada CO e o

percentual destas variáveis retidas é construído. Cada ponto (Percentual de COsretidos; RMSE) é compreendido como uma coordenada no espaço. Além disso, os autores estabeleceram um ponto denominado ótimo, com coordenadas próximas à zero (0.01, 0.01), uma vez que, o ideal seria obter um erro próximo à zero retendo-se o menor percentual possível de variáveis. Os autores então selecionaram a coordenada que gerou a menor distância Euclidiana entre do ponto ótimo.

Já Soares e Anzanello (2018) analisam doze bancos, dentre os quais Número de cetano e Ponto de fulgor, analisados neste artigo. Os autores propõem o uso do método *Support Vector Regression – Recursive Feature Elimination* (SVR-RFE), um algoritmo para selecionar os COs mais informativos para utilizar em modelos SVR. Além disso, os autores construíram outros modelos SVR e PLS utilizando outros métodos de seleção de COs, tais como *interval* PLS (iPLS), *backward interval* PLS (biPLS), *synergy interval* PLS (siPLS) e *Successive Projection Algorithm* PLS (SPA-PLS). O estudo revelou SVR como uma robusta alternativa ao PLS, especialmente quando SVR-RFE é empregado para seleção de COs. Para as bases analisadas no presente artigo (Ponto de ebulição e Número de cetano) SVR-RFE apresentou os melhores resultados.

A tabela 3.4 exibe os melhores resultados encontrados em cada trabalho, assim como os melhores resultados do método proposto. O método proposto neste artigo apresentou resultados substancialmente melhores do que os métodos propostos por Anzanello et al. (2015) e Soares e Anzanello (2018).Tal superioridade pode estar associada a dois aspectos: (*i*) o índice de importância de CO aqui proposto apoia-se somente nas amostras com valores extremos da variável dependente (amostras inseridas em Q1 e Q4), dando origem a um índice com maior poder de separação ao desconsiderar as amostras intermediárias (amostras inseridas em Q2 e Q3) na composição do índice; e (*ii*) a ferramenta de predição ANN mostrou-se mais robusta do que PLS e SVM nos bancos analisados, o que pode sugerir a presença de não-linearidades nos dados (condição em que a ANN apresenta melhor desempenho quando comparada a outras técnicas de regresão).

Tabela 3.4– Comparação dos resultados entre a abordagem proposta e métodos propostos por outros autores.

Base	Autor	Abordagem	Nº de COs	% COs retidos	RMSE Médio
Ponto de Ebulição	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**11**	**2.7431**	**0.1214**
	Soares e Anzanello (2018)	SVR-RFE	15	3.7406	2.6462
Número de Cetano	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**4**	**0.9975**	**0.1314**
	Soares e Anzanello (2018)	SVR-RFE	4	0.9975	1.7906
Ponto de Fulgor	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**12**	**0.6904**	**0.0886**
	Anzanello, Fogliatto e Ferrão (2015)	Índice de importância de COs baseado em parâmetros PLS + Regressão PLS	89	5.13	1.021

3.5 Conclusão

Tendo em vista que os dados gerados por técnicas de espectrometria podem reduzir o poder preditivo de modelos matemáticos em virtude de conjuntos de dados compostos por milhares de COs altamente correlacionados, este artigo propôs um método de seleção de COscom vistas à redução de dimensionalidade. Tal método integra os conceitos da lei de Lambert-Beerà técnica de classificação de redes neurais. As quatro etapas desse método são: (i) pré-processar os dados, ajustando uma escala única para todas as variáveis, (ii) dividir o conjunto de dados em conjuntos de treinamento e teste, (iii) ranquear os comprimentos de onda e (iv) selecionar comprimentos de onda. Foram preditas propriedades de três conjuntos de dados NIR referentes a diesel para validação do método:ponto de ebulição, número de cetano e ponto de fulgor.

O melhor resultado do conjunto de dados Ponto de ebulição reteve média de 2,74% dos COs iniciais, pouco menos de 1% para Número de cetano, e 0,69% para Ponto de fulgor. O RMSE médio para todos os conjuntos de dados permaneceu pequeno, ficando abaixo de 0,14, com desvio-padrão variando de 1% a 4%. Os resultados obtidos neste artigo demonstram a robustez do método e sugerem que os métodos de seleção de variáveis podem ser simplificados usando uma estratégia de análise de parte de um banco de dados (somente quartis extremos), ao invés de utilizar todas as amostras na construção de um índice de importância de COs.

A abordagem aqui proposta superou outros métodos para seleção de COs com propósitos de predição de propriedade de produtos. Para a base Ponto de ebulição,a abordagem proposta reteve 1% a menos de COs do que a deSoares e Anzanello (2018), com um erro 22 vezes menor. Para Número de cetano, o percentual de COs retido foi equivalente nas duas abordagens – proposta deste artigo e de Soares e Anzanello (2018) – porém a abordagem proposta obteve um erro médio 13 vezes menor que a abordagem concorrente. Da mesma maneira, a abordagem proposta apresentou os melhores resultados para a base Ponto de fulgor, retendo cerca de 13,5% a menos de COs que a abordagem de Anzanello et al. (2015)e erro 11 vezes menor.

Para pesquisas futuras, recomenda-se um estudo do grau de relação/iteração existente entre as variáveis independentes a fim de otimizar a geração dos subconjuntos. Embora um *ranking* referente às variáveis que mais explicam a variabilidade da variável dependente tenha sido estabelecido, sabe-se que, quando estas variáveis são combinadas, subconjuntos com altos níveis de importância não necessariamente precisam ser constituídos pela sequência das variáveis que mais explicam a variabilidade da variável dependente. Sendo assim, além do ranqueamento, cada variável deveria ser ponderada de acordo com suas inter-relações.

3.6 Referências

ALEME, H. G.; BARBEIRA, P. J. S. Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration. **Fuel**, v. 102, p. 129–134, 2012.

ANASTASIADIS, A. D.; MAGOULAS, G. D.; VRAHATIS, M. N. Sign-based learning schemes for pattern classification. **Pattern Recognition Letters**, v. 26, p. 1926–1936, 2005.

ANTONOPOULOS, V. Z.; PAPAMICHAIL, D. M.; ASCHONITIS, V. G.; ANTONOPOULOS, A. V. Solar radiation estimation methods using ANN and empirical models. **Computers and Electronics in Agriculture**, v. 160, p. 160–167, 2019.

ANZANELLO, M. J.; FU, K.; FOGLIATTO, F. F.; FERRÃO, M. F. HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point. **Chemometrics and Intelligent Laboratory Systems**, v. 145, p. 1–6, 2015.

CÂMARA, A. B. F.; DE CARVALHO, L. S.; DE MORAIS, C. L. M.; et al. MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends. **Fuel**, v. 210, p. 497–506, 2017.

CUNHA, D. A.; NETO, Á. C.; COLNAGO, L. A.; CASTRO, E. V. R.; BARBOSA, L. L. Application of time-domain NMR as a methodology to quantify adulteration of diesel fuel with soybean oil and frying oil. **Fuel**, v. 252, p. 567–573, 2019.

DE ALMEIDA, Marcelo Barbosa. **Um estudo comparativo de técnicas conexionistas na implementação de um sistema de reconhecimento de padrões para um nariz artificial**. 2003. Universidade Federal de Pernambuco (UFPE), Recife, 2003.

FERRÃO, M. F.; VIERA, M. D. S.; PAZOS, R. E. P.; 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.

FRITSCH, S.; GUENTHER, F. neuralnet: Training of Neural Networks. , 2016. Disponível em: <https://cran.r-project.org/package=neuralnet>. .

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, p. 171–193, 2001.

GAYDOU, V.; KISTER, J.; DUPUY, N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. **Chemometrics and Intelligent Laboratory Systems**, v. 106, p. 190–197, 2011.

GOBRECHT, A.; BENDOULA, R.; ROGER, J. M.; BELLON-MAUREL, V. Combining linear polarization spectroscopy and the Representative Layer Theory to measure the Beer-Lambert law absorbance of highly scattering materials. **Analytica Chimica Acta**, v. 853, p. 486–494, 2015.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman ed. Porto Alegre, 2001.

IGEL, C.; HÜSKEN, M. Improving the Rprop Learning Algorithm. **Proceedings of the second international ICSC symposium on neural computation (NC 2000)**, p. 115–121, 2000. ICSC Academic Press.

IGEL, C.; HÜSKEN, M. Empirical evaluation of the improved Rprop learning algorithms. **Neurocomputing**, v. 50, p. 105–123, 2003.

JAIN, S. K.; NAYAK, P. C.; SUDHEER, K. P. Models for estimating evapotranspiration using artificial neural networks, and their physical interpretation. **Hydrological Processes**, v. 22, p. 2225–2234, 2008.

JIMENEZ-MARTINEZ, M.; ALFARO-PONCE, M. Fatigue damage effect approach

by artificial neural network. **International Journal of Fatigue**, v. 124, p. 42–47, 2019.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018.

KRISHNA, S. M.; ABDUL SALAM, P.; TONGROON, M.; CHOLLACOOP, N. Performance and emission assessment of optimally blended biodiesel-diesel-ethanol in diesel engine generator. **Applied Thermal Engineering**, v. 155, p. 525–533, 2019.

KRISHNAKUMAR, K. Optimization of the neural net connectivity pattern using a backpropagation algorithm. **Neurocomputing**, v. 5, p. 273–286, 1993.

LAVOIE, F. B.; MUTEKI, K.; GOSSELIN, R. A novel robust NL-PLS regression methodology. **Chemometrics and Intelligent Laboratory Systems**, v. 184, p. 71–81, 2019.

LEE, J.; CHANG, K.; JUN, C.; et al. Kernel-based calibration methods combined with multivariate feature selection to improve accuracy of near-infrared spectroscopic analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 147, p. 139–146, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.08.009>. .

LI, D. G.; ZHEN, H.; XINGCAI, L.; WU-GAO, Z.; JIAN-GUANG, Y. Physico-chemical properties of ethanol-diesel blend fuel and its effect on performance and emissions of diesel engines. **Renewable Energy**, v. 30, n. 6, p. 967–976, 2005.

LIN, W.; HANG, H.; ZHUANG, Y.; ZHANG, S. Variable selection in partial least squares with the weighted variable contribution to the first singular value of the covariance matrix. **Chemometrics and Intelligent Laboratory Systems**, v. 183, p. 113–121, 2018.

OLIVEIRA, E. F. T. DE; GRÁCIO, M. C. C. Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação. **Revista de Ciência da Informação**, v. 6, n. 3, p. 1–11, 2005.

OLIVEIRA, T. P.; BARBAR, J. S.; SOARES, A. S. Predição do tráfego de rede de computadores usando redes neurais tradicionais e de aprendizagem profunda. **Revista de Informática Teórica e Aplicada**, v. 22, n. 1, p. 10–28, 2015.

PALOU, A.; MIRÓ, A.; BLANCO, M.; et al. Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 180, p. 119–126, 2017.

PRASAD, N.; SINGH, R.; LAL, S. P. Comparison of back propagation and resilient propagation algorithm for spam classification. Fifth International Conference on Computational Intelligence, Modelling and Simulation. **Anais...** . p.29–34, 2013.

R CORE TEAM. R: A Language and Environment for Statistical Computing. , 2018. Vienna, Austria. Disponível em: <https://www.r-project.org/>. .

RIEDMILLER, M.; RPROP, I. Rprop-Description and Implementation Details. ,

1994.

SOARES, F.; ANZANELLO, M. J. Support vector regression coupled with wavelength selection as a robust analytical method. **Chemometrics and Intelligent Laboratory Systems**, v. 172, p. 167–173, 2018.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 171–178, 2017.

TELES, M. L.; GOMES, H. M. Comparação de algoritmos genéticos e programação quadrática seqüencial para otimização de problemas em engenharia. **Teoria e Prática na Engenharia Civil**, v. 10, n. 15, p. 29–39, 2010.

WU, W.; DANDY, G. C.; MAIER, H. R. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. **Environmental Modelling and Software**, v. 54, p. 108–27, 2014.

YUN, Y. H.; LI, H. D.; DENG, B. C.; CAO, D. S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. **TrAC - Trends in Analytical Chemistry**, v. 113, p. 102–115, 2019.

4. Terceiro artigo: Uso de florestas aleatórias como alternativa robusta para calibração em dados espectroscópicos

Resumo

No presente artigo,a ferramenta de aprendizado de máquina Floresta aleatória (*Random Forest* – RF) foi estudada como alternativa para a calibração multivariada em seis bancos de dados de espectroscopia de natureza binária e multiclasse. Seis diferentes técnicas voltadas ao ranqueamento de importância de comprimentos de onda (CO) foram testadas para orientar a seleção dos COs na ferramenta classificadora RF usando uma abordagem do tipo *forward* (inclusão dos COs um a um no subconjunto utilizado pela RF). Os resultados recomendam a estatística Qui-Quadrado (χ^2) para geração do índice de importância de COs. Os modelosχ^2– RF construídos foram comparados com modelos de diferentes naturezas: índices de importância baseados no teste Kolmogorov-Smirnovintegrado à ferramenta máquina de vetores de suporte (*Suport Vector Machine* – SVM); importância deCOs com base na distância de *Bhattacharyya*(BD) combinado com a técnica de rede neural probabilística (*Probabilistic Neural Network* –PNN); relevância baseada na otimização da programação quadrática(*Quadratic Program* – QP) aliada a SVM; regressão PLS; e algoritmo genético (*Genetic Algoritm* – GA) empregado na procura do subconjunto de COs com utilização da análise discriminante linear (*Linear Discriminant Analysis* – LDA) para classificação. Na comparação entre osalgoritmos de seleção de COs, o χ^2– RF reteve um número substancialmente menor de COsquando comparado aos demais métodos.

Palavras-chave: Florestas Aleatórias, Qui-quadrado, Espectroscopia, Seleção de COs.

4.1 Introdução

A adulteração de produtos é um problema que afeta diversos setores industriais (DE SANTANA; BORGES NETO; POPPI, 2019). Normalmente, adulterações são realizadas adicionando substâncias de menor valor ou em quantidades menores ao produto autêntico, gerando uma mistura de qualidade inferior, de forma que o produto seja rotulado de forma deliberada e fraudulenta com relação à sua identidade e/ou origem (WHO, 1999). Produtos

alimentícios, farmacêuticos e até mesmo drogas ilícitas têm ganhado destaque neste tipo de contravenção (ANZANELLO et al., 2013; FERNANDEZ et al., 2011). A adulteração destes produtos pode acarretar diversos problemas, dentre os quais destacam-se aqueles relacionados à saúde pública e financeiros. O consumidor tem se conscientizado, cada vez mais, sobre questões referentes à segurança e autenticidade dos produtos. Além da autenticação de produtos alimentícios e farmacêuticos serem motivo de preocupação dos consumidores, este também é motivo de preocupação para os fabricantes que não desejam ser submetidos à concorrência desleal de fabricantes que obteriam vantagem econômica com a deturpação dos produtos que estão vendendo (REID; O'DONNELL; DOWNEY, 2006; WHO, 1999).

A menção à adulteração de drogas ilícitas pode causar certa estranheza, porém, determinar a concentração, por exemplo, de cocaína e adulterantes em amostras de drogas apreendidas é importante tanto do ponto de vista da saúde quanto forense (KAHMANN et al., 2018a). Neste caso, a vantagem da identificação de adulterantes em drogas (lícitas ou ilícitas), do ponto de vista clínico, é vista no tratamento de dependentes químicos, que é realizado de maneira mais apropriada quando a forma da droga utilizada é conhecida (BERTOL et al., 2011; BRUNT et al., 2009). Já na perspectiva forense, a vantagem da determinação da concentração de drogas ilícitas e seus adulterantes em amostras apreendidas reside naoferta de informações relevantes aos foros de investigação que interrompem o tráfico de drogas (RODRIGUES et al., 2013). A necessidade de assegurar o uso de ingredientes seguros e em conformidade legal, aliada às razões econômicas do comércio e da indústria, tem instigado pesquisas envolvendo o desenvolvimento de diversas técnicas para detecção de adulteração de produtos. Além disso, os avanços tecnológicos, proporcionados pela modernização através da globalização, têm favorecido expressivos progressos no que diz respeito à administração e armazenamento de variadas e massivas quantidades de dados (ZGUROVSKY; ZAYCHENKO, 2020). Nos últimos anos, técnicas analíticas como espectroscopia no infravermelho próximo (*Near-Infrared Spectroscopy*– NIR) e transformada de Fourier (*Fourier-transform infrared* – FTIR) têm sido utilizadas em diferentes campos e com diversas finalidades (SOARES et al., 2017). Estas técnicas podem ser usadas para identificar um composto ou investigar a composição de uma amostra, a fim de identificar possíveis adulterações (CÂMARA et al., 2017; GAYDOU; KISTER; DUPUY, 2011). No entanto, são responsáveis por gerar dados com alta dimensionalidade. À vista disso, é preciso lançar mão de técnicas de análise de dados, tais como

aprendizagem de máquina, associadas à seleção de variáveis(ANZANELLO; FOGLIATTO, 2014).

Tais conjuntos de dados, gerados por técnicas de espectroscopia, podem ser compostos por milhares de características, denominadas comprimentos de onda (COs), variáveis ou atributos. No entanto, algoritmos de aprendizagem de máquina tendem a funcionar melhor ao serem aplicados em bancos de dimensionalidade reduzida. Entende-se por dimensionalidade a quantidade de variáveis presente nos dados (TAN; STEINBACH; KUMAR, 2009). Em vista disso, faz-se necessário a utilização de técnicas de redução de dimensionalidade que permitam executar a análise dos dados sem a perda de informações relevantes, diminuindo o risco de inferências não confiáveis, assim como a redução do custo computacional (GAUCHI; CHAGNON, 2001).

Uma das principais estratégias de redução da dimensionalidade é o processo de seleção de variáveis (COCCHI; BIANCOLILLO; MARINI, 2018; MLADENIÉ, 2006; SAEYS; INZA; LARRAÑAGA, 2007). Métodos de seleção de variáveis são utilizados para identificar o "melhor" subconjunto de variáveis preditoras; em outras palavras, um subconjunto contendo as variáveis que melhor definem a classe da amostra ou melhor correlacionamvariáveis de entrada (independentes) a propriedades de interesse (variável dependente). Dentre as principais razões para se utilizar técnicas de seleção de variáveis, destaca-se redução da complexidade dos modelos, geração de preditores mais rápidos, melhora do desempenho dos modelos, melhor entendimento do processo e redução nos custos experimentais e de coleta de dados (COCCHI; BIANCOLILLO; MARINI, 2018; SAEYS; INZA; LARRAÑAGA, 2007).

Existe uma vasta quantidade de métodos de seleção de variáveis na literatura. Tais métodos incluem três abordagens mais comumente usadas: *filter*, *wrapper* e *embedded*. Métodos *filter* operam diretamente no banco de dados e fornecem pesos, *ranking* ou subconjuntos de variáveis (SAEYS; INZA; LARRAÑAGA, 2007). Métodos *wrapper* realizam uma busca entre os possíveis subconjuntos a serem avaliados e, conforme Kohavi e John (1997), em vez de usar um teste independente como abordagens *filter*, utiliza o próprio algoritmo de indução para avaliar os subconjuntos de acordo com a acurácia de um dado preditor. Já em abordagens do tipo *embedded*, a seleção do subconjunto é embutida ou integrada no próprio algoritmo de aprendizado, que funciona como uma caixa preta. Faceli et al.(2011) citam árvores de decisão como algoritmos que realizam esse tipo de seleção interna de atributos.

O presente artigo propõe a utilização da ferramenta de aprendizagem de máquina, denominada floresta aleatória (*Random Forest* – RF) com vistas à calibração multivariada em seis bancos de dados de espectroscopia de natureza binária e multiclasse. Tais classes podem dizer respeito à autenticidade ou categoria de qualidade das amostras. Seis diferentes técnicas de ranqueamento foram testadas para orientar a seleção das variáveis na ferramenta classificadora RF, de acordo com a abordagem *forward inclusion stepwise*. Esta abordagem é baseada na ordem das variáveis estabelecida por cadaíndice de importância anteriormente gerado. A partir de um subconjunto inicialmente vazio, as variáveis são inseridas uma a uma, da mais importante para a menos importante, na modelagem. Se houver aumento na acurácia, a variável candidata é mantida no subconjunto de variáveis selecionadas; caso contrário, é eliminada e a próxima variável é adicionada ao subconjunto. Quando todas as variáveis forem testadas ou o subconjunto em análise atingir acurácia de 100%, o processo é encerrado. Os modelos construídos pelo índice de importância recomendado foram comparados com modelos de diferentes naturezas reportados nas pesquisas de Kahmann et al. (2018b), Anzanello et al. (2015), Kahmann et al. (2017), Holland, Kemsley e Wilson (1998) e Tapp, Defernez e Kemsley(2003). Sob perspectiva prática, o método visa atender à ciência forense na identificação de amostras adulteradas em produtos alimentícios e em drogas (lícitas e ilícitas), oferecendo subsídios para ações que interrompam falsificações.

4.2 Materiais e método

4.2.1 Bancos de dados espectrais

Seis bancos de dados espectrais relacionados à identificação de adulteração de alimentos e drogas (lícitas e ilícitas)foram usados para avaliação do desempenho do método proposto (ver tabela 4.1). As bases de dados referentes ao Cialis®, cocaína, Viagra® e purê de fruta têm como objetivo classificar as amostras em duas classes (autêntica ou adulterada). Já as bases referentes a erva mate e azeite de oliva objetivam identificar a origem geográfica de cada amostra (4 classes para erva mate:Brasil, Paraguai, Argentina e Uruguai; e quatro classes para azeite de oliva:Grécia, Itália, Portugal e Espanha).

Na base Cialis®, comprimidos autênticos contendo 20mg de tadalafila e tadalafila padrão (99,8%) foram fornecidos pelo laboratório Eli Lilly do Brasil. 20 comprimidos

autênticos de Cialis® (TAD, 20mg), de 8 lotes distintos, foram comprados em farmácias locais. Quanto às amostras falsificadas, 104 comprimidos foram enviados à polícia federal brasileira (Porto Alegre, RS) para análise forense por intermédio da *reflexão total atenuada* no infravermelho com transformada de Fourier (ATR-FTIR). A base Cocaina contém dados espectrais obtidos por meio de técnicas de infravermelho com transformada de Fourier (FTIR) de 513 amostras de cocaína (277 amostras de sal e 236 amostras base) apreendidas entre 2011 e 2012 pela Polícia Federal Brasileira do Rio Grande do Sul. Na base Viagra®, os comprimidos autênticos da droga contendo 50 mg de citrato de sildenafilae citrato de sildenafilapadrão (99,9%) foram fornecidos pelo laboratório Pfizer Ltda. Comprimidos de Viagra® (SLD, 50 mg) de 6 lotes distintos foram comprados em farmácias locais. Da mesma forma que na base Cialis®, amostras falsificadas foram enviadas à polícia federal brasileira (Porto Alegre, RS) para análise forense por meio da técnica ATR-FTIR. Quanto aos dados referentes ao purê de frutas, as amostras de purês foram preparadas em laboratório com frutas inteiras, frescas ou descongeladas. Misturas foram feitas pegando o purê puro e adicionando uma quantidade apropriada de adulterante (maça, ameixa, soluções de açúcar, suco de uva tinto e compota de ruibarbo). Os dados espectrais foram obtidos utilizando ATR-FTIR. Maiores detalhes sobre a obtenção dos dados desta base estão disponíveis em Holland, Kemsley Wilson(1998).

Já na base erva mate, 54 pacotes de diferentes marcas foram comprados em mercados de 4 países sul americanos: 19 do Brasil, 14 do Paraguai, 14 da Argentina e 7 do Uruguai. O número diferente de marcas disponíveis em cada país justifica o diferente número de amostras derivadas de cada país. A origem geográfica e informações adicionais estavam disponíveis nos rótulos das embalagens. Os dados foram obtidos por meio de análise de espectroscopia no infravermelho próximo (NIRS). Na base Óleo de Oliva, 60 amostras autenticadas de azeite virgem, originárias de quatro países produtores europeus, foram obtidas no Conselho Internacional do Azeite em Madri. Os dados foram obtidos por meio de análises ATR-FTIR em dois períodos distintos, gerando um espectro diferente de absorbância para cada período em cada amostra. Juntando os dados dos dois períodos de aquisição forneceu um conjunto de dados de 120 espectros de absorbância.Detalhes sobre os dados são fornecidosem Tapp, Defernez e Kemsley (2003).

Tabela 4.1– Bases de dados espectrais usados nesse estudo.

Base	Nº classes	COs	Nº de amostras	Descrição	Refs	Fonte
Cialis®	2	661	300	Indentificação de autenticidade ou adulteração de Cialis®	(DOS SANTOS et al., 2019; ORTIZ et al., 2013)	-
Cocaina	2	665	513	Indentificação de autenticidade ou adulteração da cocaína	(ANZANELLO et al., 2015)	-
Erva Mate	4	3801	54	Identificação do país de origem da erva mate	(KAHMANN et al., 2017)	-
Purê de Frutas	2	235	984	Indentificação de autenticidade ou adulteração do purê de fruta de morango	(HOLLAND; KEMSLEY; WILSON, 1998)	http://www.timeseriesclassification.com/description.php?Dataset=Strawberry
Azeite de Oliva	4	570	120	Identificação do país de origem do azeite de oliva	(TAPP; DEFERNEZ; KEMSLEY, 2003)	https://csr.quadram.ac.uk/example-datasets-for-download/
Viagra®	2	661	177	Indentificação de autenticidade ou adulteração da Viagra®	(DOS SANTOS et al., 2019; ORTIZ et al., 2013)	-

4.2.2 Métodos de ranqueamento

Rankings de importância de variáveis têm sido consistentemente utilizados em processos de seleção das variáveis mais informativas com vistas à predição e classificação de amostras. Abaixo são detalhados os métodos de ranqueamento que serão testados neste artigo.

4.2.2.1 Qui-Quadrado (χ^2)

O grau de importância da variável é avaliado pelo valor da estatística qui-quadrado em relação à classe.A hipótese inicial (H_0) é a suposição de que as duas variáveis (variável independente e variável resposta) não são relacionadas. Esta hipótese é testada pela equação 4.1; caso rejeitada, assume-se a hipótese alternativa (H_1), a qual supõe que as duas variáveis são dependentes (ASDAGHI; SOLEIMANI, 2019; NOVAKOVIĆ; STRBAC; BULATOVIĆ, 2011).

$$\chi_j^2 = \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4.1)$$

Onde χ_j^2 é o valor do teste qui-quadrado para a variável j,Né o número total de observações, k é o número total de variáveis independentes, O_{ij} é a frequência observada da variável j na observação i e E_{ij} é a frequência esperada (teórica) da variável j na observação i, afirmada pela hipótese nula. Quanto maior o valor de χ^2, maior é a evidência contra a hipótese H_0, ou seja, a variável tem maior impacto na variável resposta; sendo assim, pode ser selecionada para o treinamento do modelo. O valor da hipótese nula (H_0) é aceito desde que o valor de qui-quadrado seja menor do que o valor do qui-quadrado crítico.

Os valores esperados para E_{ij} são gerados pela equação 4.2 onde,para cada valor $x_i$da variável $X(1 \leq i \geq k)$ e para cada valor c_j da classe $C(1 \leq j \geq m)$, existe uma frequência esperada quando $(X = x_i))$ e $\left(C = c_j\right)$ (PEREIRA, 2009).

$$E_{ij} = \frac{count(X = x_i) * count\left(C = c_j\right)}{N} \quad (4.2)$$

onde $count(X = x_i)$é o número de observações em que ocorre o valor $x_i$do variável$X$e $count\left(C = c_j\right)$é o número de observações que pertencem à classe c_j. A partir da frequência esperada de todas as combinações de valores i e j, pode-se calcular a métrica χ^2.

4.2.2.2 Relief-F

O método *Relief-F* (ReF) seleciona aleatoriamente n instâncias do conjunto de treinamento. Para cada instância selecionadar_i, o *Relief-F* calcula os k vizinhos mais próximos da mesma classe de r_i, chamados acertos mais próximos, e os k vizinhos mais próximos para cada classe diferente, chamados de erros mais próximos. Os valores da amostra selecionada são comparados com os acertos e os erros e, em seguida, a pontuação de relevância para cada variável é atualizada. A importância de uma variável é dada pelo fato desta possuir valores semelhantes às instâncias da mesma classe e diferentes das instâncias de outras classes. Sendo assim, uma variável tem uma propriedade indesejada se diferencia das instâncias próximas que pertencem à mesma classe. Do contrário, tem uma propriedade desejada diferenciando as

instâncias próximas que pertencem à classes diferentes (REMESEIRO; BOLON-CANEDO, 2019; REYES; MORELL; VENTURA, 2015). A ideia deste algoritmo é estimar a capacidade de distinção de classes apresentada por cada variável levando em consideração instâncias vizinhas. O critério de avaliação *Relief-F* é definido na equação 4.3.

$$ReF(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{1}{n_{i,C(r_i)}}\sum_{r_t \in NH(r_i)}\|r_{i,j} - r_{t,j}\| + \sum_{C \neq C(r_i)}\left(\frac{P(C)}{1-P(C(r_i))} * \frac{1}{n_{i,C}} * \sum_{x_t \in NM(r_i,C)}\|r_{i,j} - r_{t,j}\|\right)\right\} \tag{4.3}$$

onde $C(r_i)$ retorna o rótulo da classe da instância r_i e $P(C)$é a probabilidade de instâncias pertencentes à classe C;$r_{i,j}$ é o valor da variável x_j na instância r_i. $NH(r_i)$ indica o conjunto de amostras mais próximo de r_ie com a mesma classe de r_i. Uma amostra no $NH(r_i)$ é chamada de "*nearest hit*" (acerto mais próximo) de r_i; $NM(r_i, C)$ denota o conjunto de amostras mais próximas de r_i e com rótulo de classe$C(C \neq C(r_i))$. E uma amostra no $NM(r_i)$ é chamada de "*nearest miss*" (erro mais próximo) de r_i; $n_{i,C(r_i)}$ é o tamanho de $NH(r_i)$ e $n_{i,C}$ é o tamanho de $NM(r_i, C)$. Geralmente,os tamanhos de $NH(r_i)$ e $NM(r_i)$ são ajustados para uma constante pré-especificada. Quanto maior o valor de $ReF(x_j)$, mais significativa é a variável.

4.2.2.3 Laplacian Score

Para cada variável, o escore de *Laplacian* é calculado para refletir seu poder de preservação da localidade. *Laplacian score* (LS) baseia-se na hipótese de que as instâncias de mesma classe estão próximas umas das outras (HINDAWI, 2013). Para tanto, o LS constrói um gráfico de vizinho mais próximo para modelar a estrutura geométrica local e buscar as variáveis que respeitam essa estrutura (HE; CAI; NIYOGI, 2006);escores elevados sugerem variáveis relevantes. Considere LS_r o escore da *j-th* variável, $x_{j,i}$ a *i-th* amostra da *j-th* variável, com *i= 1, ..., m*. O método segue os seguintes passos:

1. Para a j-th variável define-se:

$$x_j = [x_{j1}, x_{j2}, \dots, x_{jm}]^T \tag{4.4}$$

2. Construção do gráfico de vizinhos mais próximos G com *m* nós.

a. O *i-th* nó corresponde a r_i.

b. Liga os nós *i* e *j* por uma aresta se r_ie r_j estão "próximos" (se r_i está entre os *k* vizinhos mais próximos de r_j ou vice-versa.

3. Calcular a matriz de pesos S a partir do gráfico G.

a. A matriz *S* é definida conforme a equação 4.5.

$$S(i,j) = \begin{cases} S_{ij} = e^{-\frac{\|r_i - r_j\|^2}{t}}, & i\ e\ j\ estão\ conectados \\ 0, & caso\ contrário \end{cases} \tag{4.5}$$

4. Calcular a matriz diagonal *D*

$$D = diag(S\mathbf{1}) \tag{4.6}$$

a. A soma de cada linha da matriz *S* é o valor da diagonal principal da matriz D (equação 4.7)

$$d_{ij} = \sum_{k=1}^{n} S_{ik} \tag{4.7}$$

a. A matriz *D* é definida conforme a equação 4.8.

$$D(i,j) = \begin{cases} d_{ij}, & i = j \\ 0, & caso\ contrário \end{cases} \tag{4.8}$$

5. Calcular a matriz laplacian (*L*)

$$L = D - S \tag{4.9}$$

6. Computar o *LS* das *j-ths* variáveis conforme a equação 4.10.

$$LS_j = \frac{\tilde{x}_j^T * L\tilde{x}_j}{\tilde{x}_j^T * D\tilde{x}_j} \tag{4.10}$$

onde $\tilde{x}_j$ é dado pela equação 4.11.

$$\tilde{x}_j = \frac{x_j - x_j^T * D\mathbf{1}}{\mathbf{1}^T * D\mathbf{1}} \tag{4.11}$$

$$\mathbf{1} = [1, \dots, 1]^T \tag{4.12}$$

4.2.2.4 Ganho de informação

O ganho de informação (GI) é amplamente usado em dados de alta dimensão para medir a relevância das variáveis em tarefas de classificação (JADHAV; HE; JENKINS, 2018). Este é um método de partição muito utilizado em algoritmos de árvores de decisão baseado em medidas de impureza, a qual é medida através da entropia. Para determinar a qualidade de uma condição de teste, é necessário comparar o grau de entropia do conjunto todo com o grau de entropia (após a divisão).Quando uma nova subdivisão dos dados acarreta na redução da entropia ocorre o ganho de informação. Sendo assim,a variável que gerar uma maior diferença é escolhida. Em outras palavras, quanto maior GI, maior a relevância da variável(ALMUALLIM; KANEDA; AKIBA, 2002).

O primeiro passo para obter o ganho de informação de uma variável é calcular o grau de entropia do conjunto de treinamento, definido pela equação 4.13 (PEREIRA, 2009).

$$Entropia(N) = -\sum_{j=1}^{m} p_j * log_2 p_j \tag{4.13}$$

onde *N* é qualquer conjunto de amostras, podendorepresentar a base completa (no caso do nó raiz) ou partições da base de dados, *m* é o número de classes e p_j é a proporção de amostras pertencendo à classe C*j* que ocorre em N.

Em seguida, considerando-se uma das variáveis do conjunto de dados N, deve-se dividi-lo em subconjuntos $N_1, N_2, \dots, N_n$, que representam os possíveis valores de uma variável de teste x.Então, a entropia da variável x é calculada de acordo com a equação 4.14.

$$Entropia_x(N) = \sum_{i=1}^{n} \left[\frac{|N_j|}{|N|} * Entropia(N_j) \right] \tag{4.14}$$

onde $|N|$é o número total de instâncias presentes no conjunto de treinamento, $|N_i|$ é o número de instâncias da classe C_j na partição N.

O ganho de informação é então obtido segundo a equação 4.15.

$$GI(x) = Entropia(N) - \sum_{i=1}^{n} \left[\frac{|N_j|}{|N|} * Entropia(N_j) \right] \quad (4.15)$$

4.2.2.5 Razão de ganho

O cálculo do ganho de informação privilegia atributos com um grande número de possíveis valores. Uma forma de minimizar esse problema é utilizar a razão de ganho (RG), que aplica uma espécie de normalização do ganho (FERRARI; SILVA, 2016; LOPES, 2016). Essa normalização faz com que os valores de RG caiam no intervalo [0, 1], onde 1 indica que a informação advinda da variável independente (preditor) prediz completamente a variável resposta; para GR = 0 denota que não há relação entre o preditor e a variável de resposta (logo, quanto maior o valor de RG, maior a relevância da variável) (ASDAGHI; SOLEIMANI, 2019; NOVAKOVIĆ; STRBAC; BULATOVIĆ, 2011). A razão de ganho é definida pela equação 4.16.

$$RG(x) = \frac{GI(x)}{Entropia_x(N)} \quad (4.16)$$

A razão de ganho expressa a proporção de informação gerada pela partição que aparenta ser mais útil para a classificação (GARCIA, 2015).

4.2.2.6 Índice Gini

O índice gini (IG) é um número compreendido entre 0 e 1, onde 0 corresponde ao fato de todas as instâncias pertencerem a uma única classe (melhor cenário) e 1 quando as instâncias relativas a uma determinada partição de valores estão igualmente distribuídas entre todas as classes (pior cenário) (ELSALAMONY, 2015; PEREIRA, 2009; SOMAN; DIWAKAR; AJAY, 2006). Para *m* classes, o índice Gini num determinado nó é definido pela equação 4.17.

$$Gini(x) = 1 - \sum_{j=1}^{m} P(C_j)^2 \tag{4.17}$$

onde $P(C_j)$ refere-se à proporção de instâncias da classe *j* no nó *t*, com $j = 1, \dots, m$.

4.2.3 Random Forest

O algoritmo floresta aleatória (*Random Forest* – RF) é usado neste artigo como classificador e como ferramenta de seleção do melhor conjunto de atributos. Trata-se de algoritmo baseado em *ensemble* de árvores de decisão (BREIMAN, 2001; STAŃCZYK; ZIELOSKO; JAIN, 2018), o qual cria árvores considerando subconjuntos aleatórios das variáveis. Dessa forma,busca a melhor variável em um subconjunto aleatório devariáveis ao fazer a partição de nós ao invés de procurar pela melhor variáveldentre todo o conjunto. Este processo de criação de árvores adiciona aleatoriedade ao modelo, criando grande diversidade e gerando modelos mais robustos. O RF é operacionalizado da seguinte forma (LIU; MOTODA, 2007; STAŃCZYK; ZIELOSKO; JAIN, 2018):

1. Para cada árvore a ser construída, uma amostra diferente com reposição é retirada dos dados de treinamento (amostra *bootstrap*).
    a. Aproximadamente 1/3 das instâncias não é usada para construir uma árvore. Estas instâncias são denominadas *out-of-bag* (OOB)
    b. O tamanho da amostra é o mesmo da base de dados original.
2. Em cada passo da construção da árvore, um subconjunto diferente de *n* variáveis é aleatoriamente selecionado
    a. Um número *n* menor que o total de *N* variáveis é definido
        i. O valor padrão para tarefas de classificação é, normalmente, $n \sim \sqrt{N}$
3. A melhor divisão entre estas *n* variáveis é escolhida para o nó atual, em contraste à construção da típica árvore de decisão que seleciona a melhor divisão entre todas as variáveis.

4.3 Procedimento Experimental

O presente estudo está dividido em quatro passos principais: (*i*) padronização da escala dos dados, (*ii*) divisão da base de dados em conjuntos de calibração e validação, (*iii*)

ranqueamento dos COs e (*iv*) seleção iterativa de COs. Estes passos são ilustrados na figura 4.1.

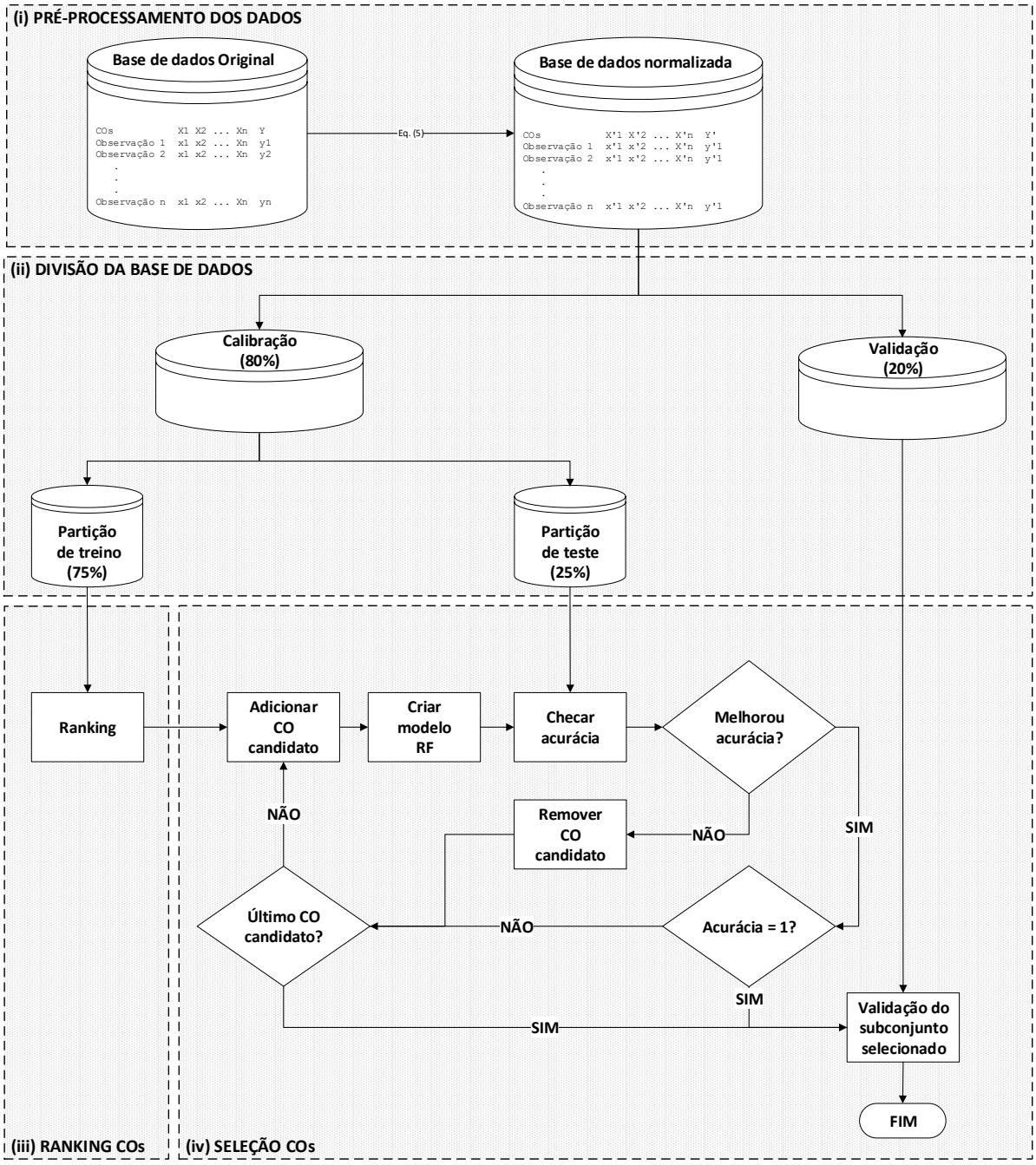

Figura 4.1 – Etapas metodológicas

4.3.1 Padronização da escala dos dados

Todas as bases de dados foram previamente normalizadasutilizando o método do mínimo-máximo[0,1], conforme a equação 19.

$$x_{norm} = \frac{(x - x_{\min})}{(x_{max} - x_{\min})} \tag{19}$$

onde x_{norm} é o valor escalonado correspondente ao valor original x, e $x_{\min}$ e x_{max} são os valores mínimo e máximo de cada CO. A normalizaçãodos dados serve para evitar que o algoritmo interprete níveis de importância em razão da escala dos valores.

4.3.2 Divisão da base de dados em conjuntos de calibração e validação

Dentro de cada base de dados, os espectros foram separados em conjuntos de validação e calibração. O conjunto de calibração, contendo 80% das amostras, é usado para ajustar o modelo e selecionar o subconjunto de COs de maior relevância, enquanto o conjunto de validação, contendo 20% das amostras restantes, é um conjunto independente usado paravalidar o modelo de calibração. O conjunto de calibração é subdividido entre o subconjunto de treinamento, contendo 75% das amostras da calibração, e o subconjunto de teste, contendo 25% das amostras da calibração.Tal particionamento é realizado 100 vezes, gerando diferentes subconjuntos aleatórios.

4.3.3 Ranqueamento dos COs

O ranqueamento dos COs é realizado utilizando os métodos qui-quadrado, *relief-F*, *laplacian score*, ganho de informação, razão de ganho e índice gini, separadamente. Os fundamentos dos métodos de ranqueamento foram apresentados na seção 4.2.2.

4.3.4 Seleção de variáveis

Cada um dos métodos de ranqueamentoorienta um processo iterativo de inserção de COs no subconjunto utilizado pela ferramenta de classificação, do mais até o menos importante. O subconjunto inicia com a variável do topo do *ranking*; uma vez que a acurácia do conjunto

vazio é zero, a primeira variável sempre irá melhorar a acurácia inicial do modelo RF. Então, a segunda variável do topo do *ranking* é inserida ao subconjunto de variáveis candidatas. Caso a acurácia do modelo não aumente, a última variável adicionada é removida do subconjunto; caso contrário, a mesma pertencerá ao subconjunto de variáveis selecionadas.O processo segue até a última variável ou até atingir acurácia 1.

4.4 Resultados e discussão

Os experimentos computacionais foram desenvolvidos usando a linguagem de programação R, versão Microsoft R Open 3.5.3 (R CORE TEAM, 2018). Foram usados os pacotes Rdimtools (SUH; YOU, 2018), randomForest (LIAW; WIENER, 2018) e CORElearn (ROBNIK-SIKONJA; SAVICKY, 2018). Foram realizadas 100 replicações aleatórias de forma estratificada e proporcional por classe. Dessa forma, calculou-se a média da acurácia dessas 100 repetições, produzindo uma medida preditiva mais confiável do que o resultado de uma única execução do modelo.

4.4.1 Desempenho dos métodos de ranqueamento na seleção dos COs

Os resultados dos experimentos trazem a média e o desvio padrão da acurácia, assim como o percentual médio de variáveis retidas nas 100 simulações. O aumento da confiabilidade do resultado é alcançado com a amostragem aleatória, pois assegura que os resultados não sejam distorcidos por conta de uma divisão favorável das amostras em calibração e validação. A tabela 4.2 apresenta a média e o desvio padrão da acurácia e do percentual de variáveis eleitas para cada método*de ranqueamento*aplicado em cada um dos bancos de dados.

Tabela 4.2– média e o desvio padrão da acurácia e do percentual de variáveis eleitas para cada método de ranqueamentoaplicado em cada um dos bancos de dados

Base	Número de COs	Métricas de desempenho	χ^2	*ReF*	*LS*	*GI*	*RG*	*IG*
Cialis® *(binário)*	**661**	**Acc_media**	0,988	**0,990**	0,948	0,983	0,982	0,988
		Acc_desvio	0,014	0,019	0,04	0,022	0,022	0,017
		PercVar_ media (%)	0,711	1,012	1,926	0,728	**0,648**	0,682
		PercVar_desvio (%)	0,195	0,316	0,347	0,220	0,195	0,184

Base	Número de COs	Métricas de desempenho	χ^2	*ReF*	*LS*	*GI*	*RG*	*IG*
Cocaina (binário)	**665**	**Acc_media**	0,990	**0,994**	0,976	0,991	0,992	0,991
		Acc_desvio	0,008	0,012	0,017	0,009	0,008	0,01
		PercVar_ media (%)	0,343	0,567	2,040	0,330	**0,311**	0,326
		PercVar_desvio (%)	0,114	0,157	0,337	0,117	0,093	0,107
ErvaMate (multiclasse)	**3801**	**Acc_media**	**0,842**	0,784	0,765	0,754	0,757	0,785
		Acc_desvio	0,123	0,138	0,147	0,16	0,164	0,135
		PercVar_ media (%)	0,109	0,110	0,114	0,109	**0,106**	0,110
		PercVar_desvio (%)	0,028	0,028	0,028	0,032	0,030	0,026
Purê de Frutas (binário)	**235**	**Acc_media**	0,949	0,947	**0,959**	0,945	0,947	0,947
		Acc_desvio	0,015	0,014	0,018	0,016	0,014	0,015
		PercVar_ media (%)	4,853	**4,412**	6,965	4,843	4,610	4,794
		PercVar_desvio (%)	1,038	0,844	0,878	1,005	0,916	0,787
Azeite de Oliva (multiclasse)	**570**	**Acc_media**	0,872	**0,875**	0,82	0,872	**0,875**	0,852
		Acc_desvio	0,081	0,076	0,088	0,063	0,079	0,071
		PercVar_ media (%)	0,938	0,999	1,718	1,037	0,981	**0,964**
		PercVar_desvio (%)	0,234	0,234	0,300	0,284	0,268	0,258
***Viagra®* (binário)**	**661**	**Acc_media**	0,900	0,907	0,875	**0,915**	**0,915**	0,905
		Acc_desvio	0,068	0,052	0,056	0,059	0,057	0,052
		PercVar_ media (%)	0,842	0,868	1,159	0,845	0,916	**0,835**
		PercVar_desvio (%)	0,227	0,239	0,313	0,232	0,274	0,230

É possível notar que todos os métodos obtiveram resultados próximos quanto à acurácia e o percentual de variáveis selecionadas. Para osbancos com duas classes, todos os métodos apresentaram desempenho superior a 90%, com exceção do método composto por *laplacian score* que atingiu a menor acurácia média (87,5%), retendo 1,16% dos COs no banco Viagra®. Todos os métodos retiveram em média menos que 5% das variáveis. Já para os problemas multiclasse, as acurácias médias mantiveram-se acima de 75%, sendo a base ErvaMate a que apresentou maior dificuldade preditiva (com as menores acurácias apontadas) e um baixo índice de variáveis selecionadas, menos de 0,2%. Isso pode ser explicado pelo fato desta ser a base de maior dimensionalidade (3801 variáveis e 54 observações). Para o banco Azeite de Oliva, o

percentual médio de variáveis selecionadas foi superior ao observado no banco Erva Mate, resultando em valores inferiores a 0,4% e acurácia média acima de 82%.

A tabela 4.3 apresenta a média geral de desempenhode cada técnica de ranqueamento combinada com a ferramenta classificadora RF.

Tabela 4.3– Média geral de desempenho de cada método de ranqueamento para todas as bases testadas

Categoria	Métricas de desempenho	χ^2	*ReF*	*LS*	*GI*	*RG*	*IG*
Binário	**Acc_media**	0,95675	**0,9595**	0,9395	0,9585	0,959	0,95775
	Acc_desvio	0,02625	0,02425	0,03275	0,0265	0,02525	0,0235
	PercVar_ media (%)	1,68725	1,71475	3,0225	1,6865	1,62125	1,65925
	PercVar_desvio (%)	0,3935	0,389	0,46875	0,3935	0,3695	0,327
	Maior_Acc	0,99	0,994	0,976	0,991	0,992	0,991
	Menor_Acc	0,9	0,907	0,875	0,915	0,915	0,905
Multiclasse	**Acc_media**	**0,857**	0,8295	0,7925	0,813	0,816	0,8185
	Acc_desvio	0,102	0,107	0,1175	0,1115	0,1215	0,103
	PercVar_ media (%)	0,5235	0,5545	0,916	0,573	0,5435	0,537
	PercVar_desvio (%)	0,131	0,131	0,164	0,158	0,149	0,142
	Maior_Acc	0,872	0,875	0,82	0,872	0,875	0,852
	Menor_Acc	0,842	0,784	0,765	0,754	0,757	0,785
Geral	**Acc_media**	**0,924**	0,916	0,891	0,910	0,911	0,911
	Acc_desvio	0,052	0,052	0,061	0,055	0,057	0,050
	PercVar_ media (%)	1,299	1,328	2,320	1,315	1,262	1,285
	PercVar_desvio (%)	0,306	0,303	0,367	0,315	0,296	0,265
	Maior_Acc	0,99	0,994	0,976	0,991	0,992	0,991
	Menor_Acc	0,842	0,784	0,765	0,754	0,757	0,785

Nota-se que o método *wrapper* composto pela combinação de chi-quadrado com a ferramenta classificadora *Random Forest* (χ^2 – RF)é o quedemonstra melhor desempenho, com uma acurácia média geral de 92,4% (tabela 4.3) em todos os bancos testados. Percebe-se que esse método apresentou os melhores resultados no que tange a problemas de classificação multiclasse, enquanto que o desempenho geral dos métodos para problemas binários não resultou em diferenças significativas, ondepara as bases Cialis® e Cocaina o método χ^2 – RF apresentou uma acurácia média acima de 98% e para Purê de Frutas e Viagra® acima de 90% (tabela 4.2). Por outro lado, o método composto por *laplacian score* (LS – RF)obteve os piores

resultados. Sendo assim, recomenda-se χ^2 – RF para aplicações futuras de categorização de amostras.

A tabela 4.4 apresenta os COs selecionados em cada base de dados e suas respectivas frequências percentuais. As figuras 4.2 a 4.7 apresentam os gráficos de frequência de retenção dos COs nas 100 simulações para cada base de dados. Um limite de 5% de frequência foi estabelecido para a identificação das regiões mais representativas do espectro para χ^2 – RF.

Tabela 4.4– Frequências percentuais dos COs retidos para cada base de dados

Cialis®		**Cocaína**		**Erva Mate**		**Purê de Frutas**		**Azeite de Oliva**		**Viagra®**	
CO(cm^{-1})	F %	CO(cm^{-1})	F %	CO(cm^{-1})	F %	CO(cm^{-1})	F %	CO(cm^{-1})	F %	CO(cm^{-1})	F %
1618	95	1176	62	7548	37	1443,585	99	1136,5595	43	1259	43
1620	48	1747	18	4580	16	1181,106	76	962,9005	41	1257	28
769	37	1174	17	4588	16	1150,226	73	1134,63	28	1261	22
1614	27	1751	15	4546	13	1177,246	62	966,7595	28	866	22
1151	26	1147	14	4558	9	1184,966	55	1256,1925	20	868	22
771	25	1741	14	4603	9	1439,725	50	964,83	19	883	17
1153	23	1749	9	4604	8	1173,386	39	968,689	19	1263	16
1612	18	1178	6	7547	7	1119,347	30	1732,795	18	1255	13
752	15	1745	6	4316	6	1516,925	27	970,6185	18	1290	12
1622	12	731	6	4567	6	1447,445	23	960,971	17	870	12
775	11	850	5	4573	6	1435,865	22	1198,3055	15	860	11
773	10	-	-	4322	5	1146,366	21	1734,7245	15	872	9
1616	9	-	-	7549	5	1327,786	20	959,0415	9	864	8
754	8	-	-	-	-	1698,344	19	974,4775	8	1298	7
1155	7	-	-	-	-	1188,826	18	978,3365	8	829	7
777	7	-	-	-	-	1115,487	17	1200,235	7	833	7
1610	5	-	-	-	-	1706,064	17	1258,122	7	862	7
-	-	-	-	-	-	1713,784	16	1717,358	7	1115	6
-	-	-	-	-	-	1694,484	13	1725,077	7	1147	6
-	-	-	-	-	-	1215,846	11	1730,8655	7	1265	6
-	-	-	-	-	-	1331,646	11	1130,77	6	835	6
-	-	-	-	-	-	1702,204	11	1132,7005	6	879	6
-	-	-	-	-	-	1709,924	11	1182,8685	6	881	6
-	-	-	-	-	-	1323,926	10	1213,7415	6	1093	5
-	-	-	-	-	-	1482,185	10	1254,263	6	1144	5
-	-	-	-	-	-	1520,785	10	1261,981	6	831	5
-	-	-	-	-	-	1103,907	9	957,112	6	-	-
-	-	-	-	-	-	1320,066	9	1088,32	5	-	-
-	-	-	-	-	-	1486,045	9	1138,489	5	-	-
-	-	-	-	-	-	1513,065	9	1215,671	5	-	-

-	-	-	-	-	-	1756,244	9	1236,8965	5	-	-
-	-	-	-	-	-	1142,506	8	1721,217	5	-	-
-	-	-	-	-	-	1211,986	8	-	-	-	-
-	-	-	-	-	-	1432,005	7	-	-	-	-
-	-	-	-	-	-	1501,485	7	-	-	-	-
-	-	-	-	-	-	1069,167	6	-	-	-	-
-	-	-	-	-	-	1204,266	6	-	-	-	-
-	-	-	-	-	-	1335,506	6	-	-	-	-
-	-	-	-	-	-	1524,645	6	-	-	-	-
-	-	-	-	-	-	1065,307	5	-	-	-	-
-	-	-	-	-	-	1096,187	5	-	-	-	-
-	-	-	-	-	-	1123,207	5	-	-	-	-
-	-	-	-	-	-	1208,126	5	-	-	-	-
-	-	-	-	-	-	1428,145	5	-	-	-	-
-	-	-	-	-	-	1493,765	5	-	-	-	-
-	-	-	-	-	-	1578,685	5	-	-	-	-
-	-	-	-	-	-	1717,644	5	-	-	-	-
-	-	-	-	-	-	1721,504	5	-	-	-	-
-	-	-	-	-	-	1729,224	5	-	-	-	-
-	-	-	-	-	-	1763,964	5	-	-	-	-

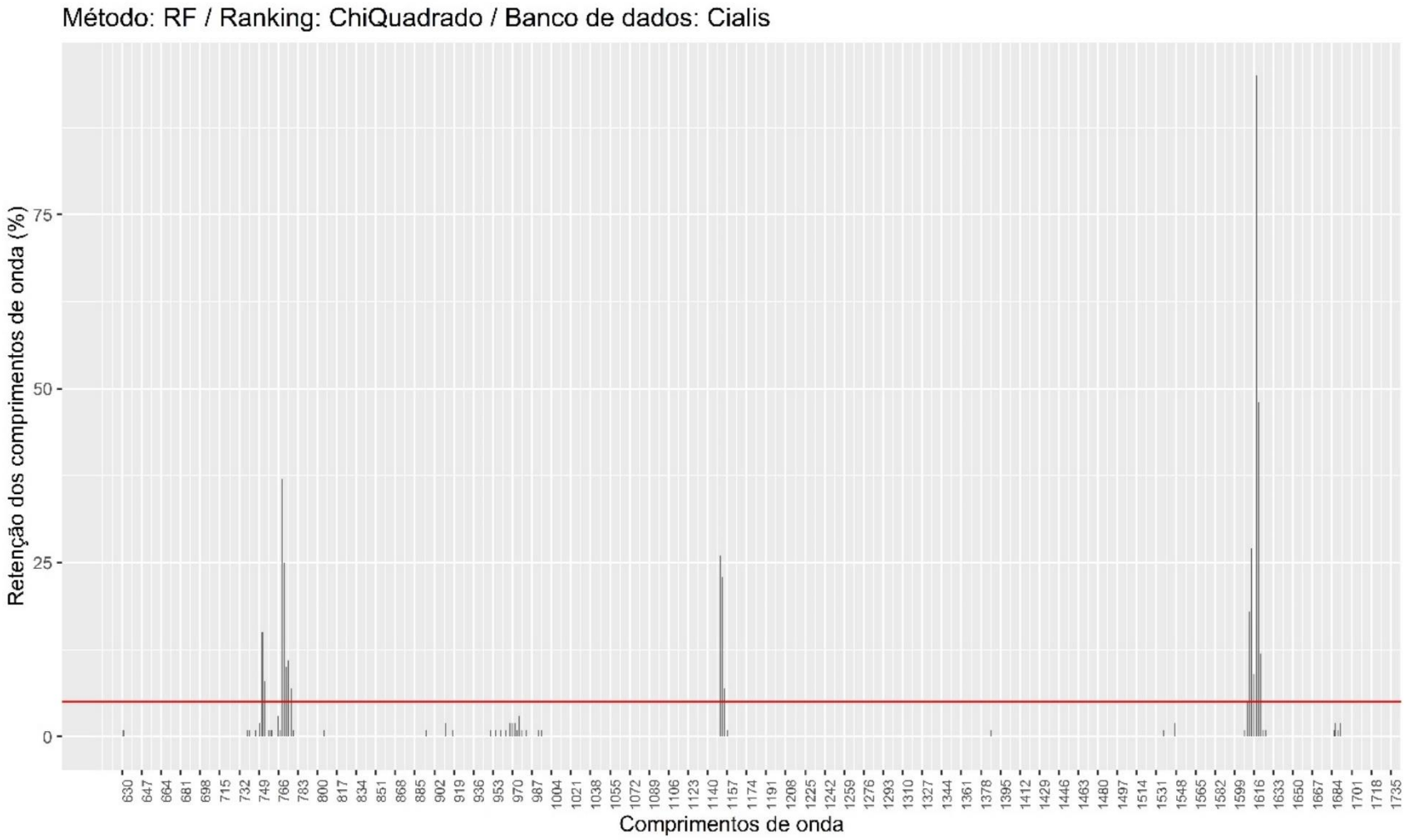

Figura 4.2– Regiões representativas do espectro da base cialis® para χ^2 – RF

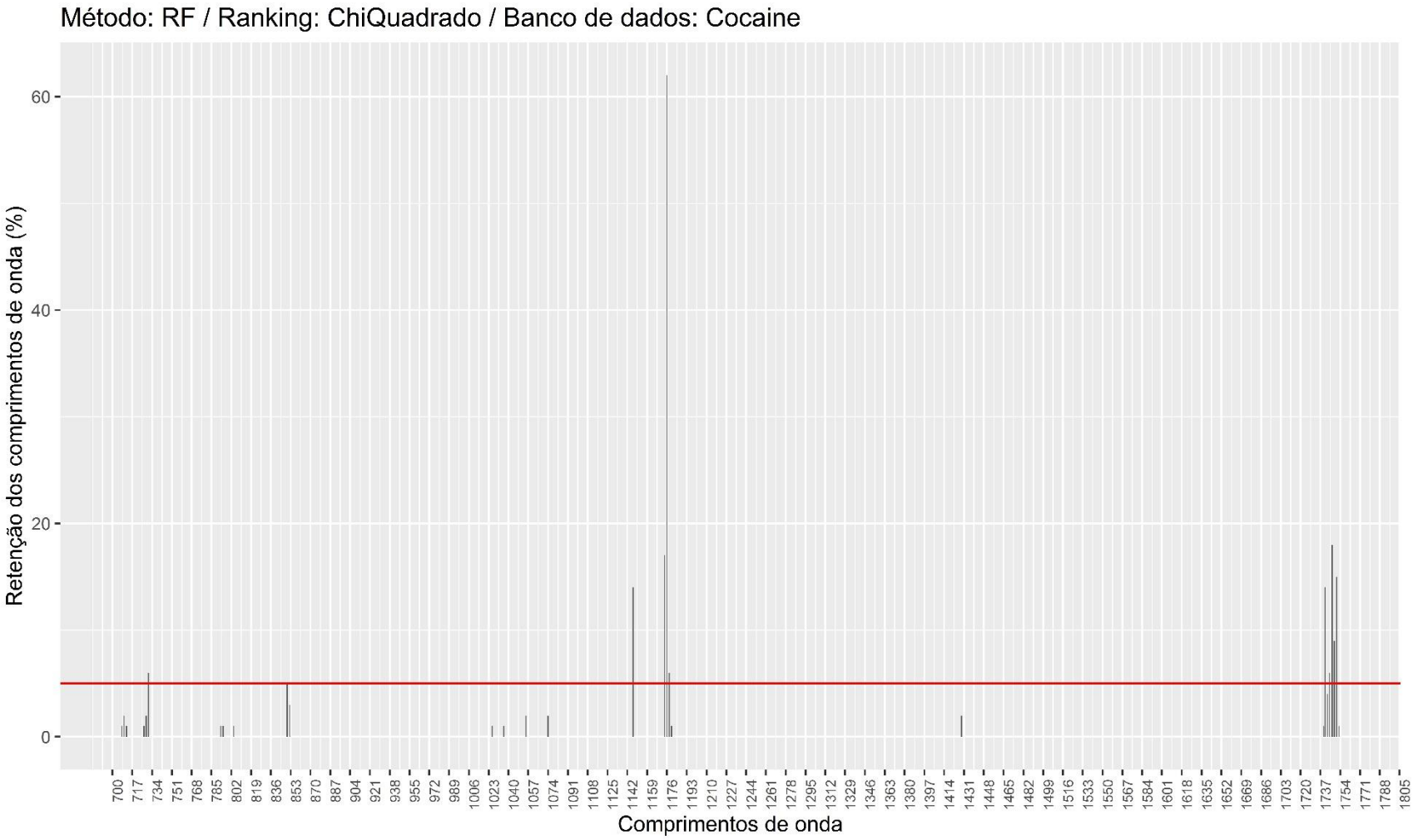

Figura 4.3– Regiões representativas do espectro da base cocaina para χ^2 – RF

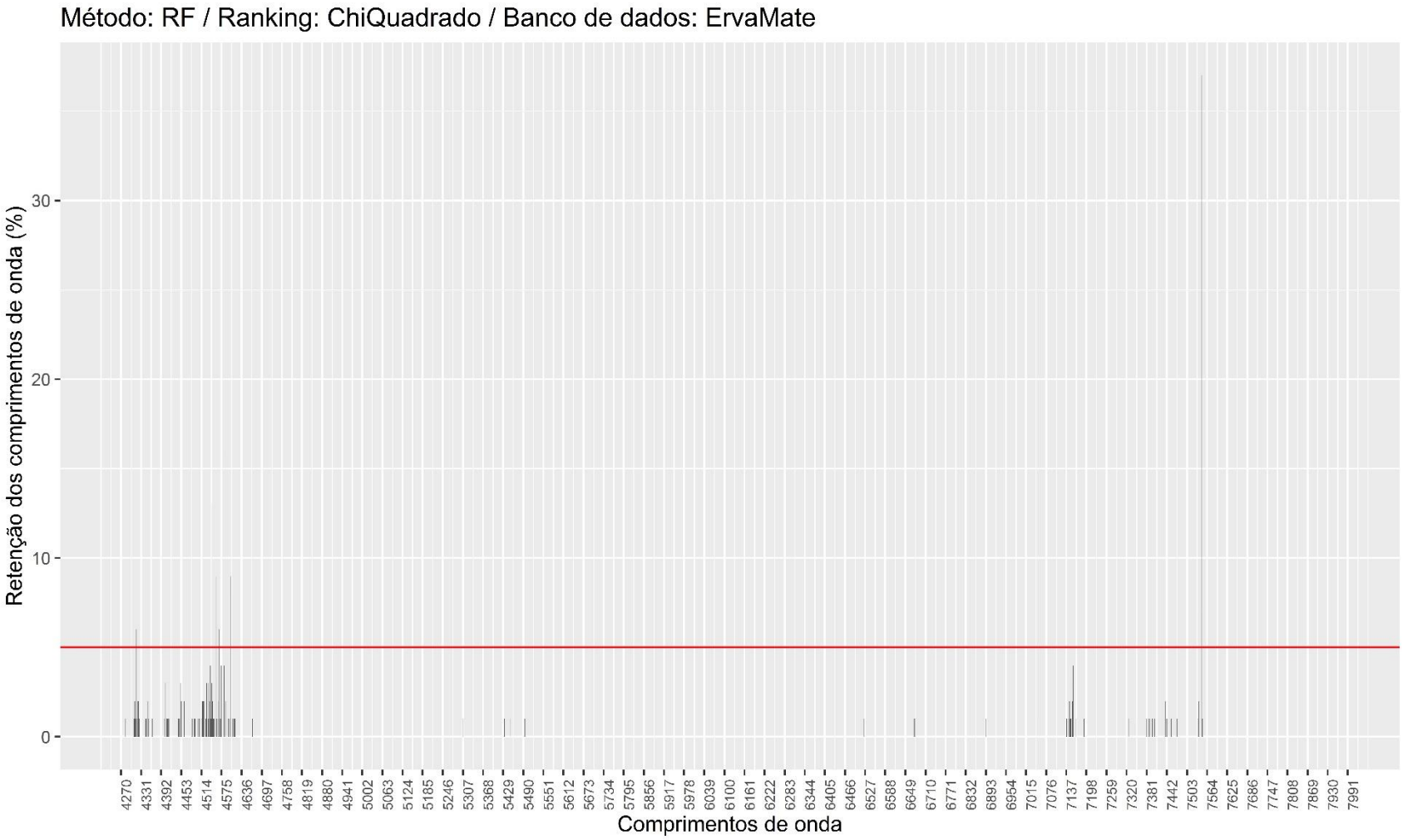

Figura 4.4– Regiões representativas do espectro da base erva mate para χ^2 – RF

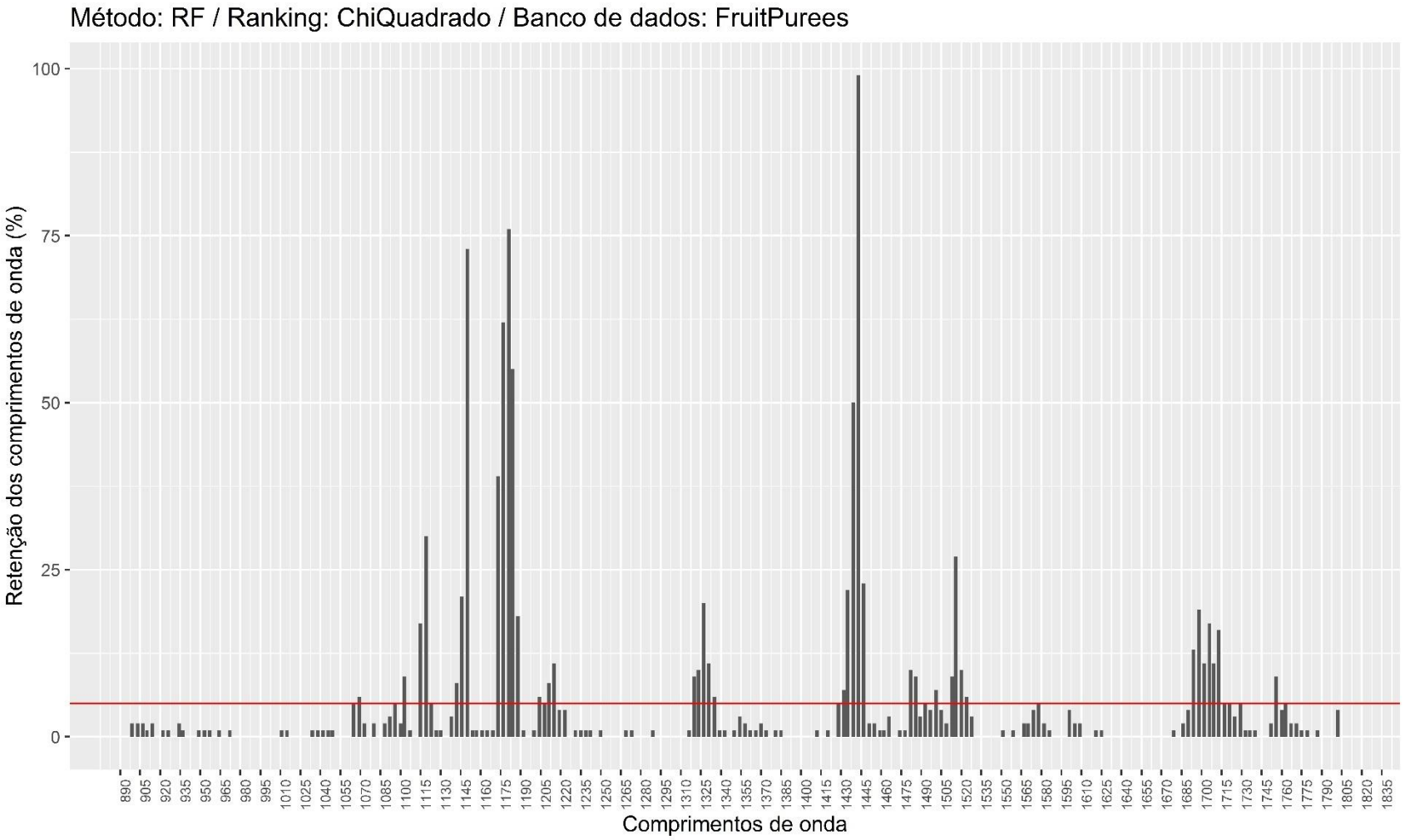

Figura 4.5– Regiões representativas do espectro da base purê de frutas para χ^2 – RF

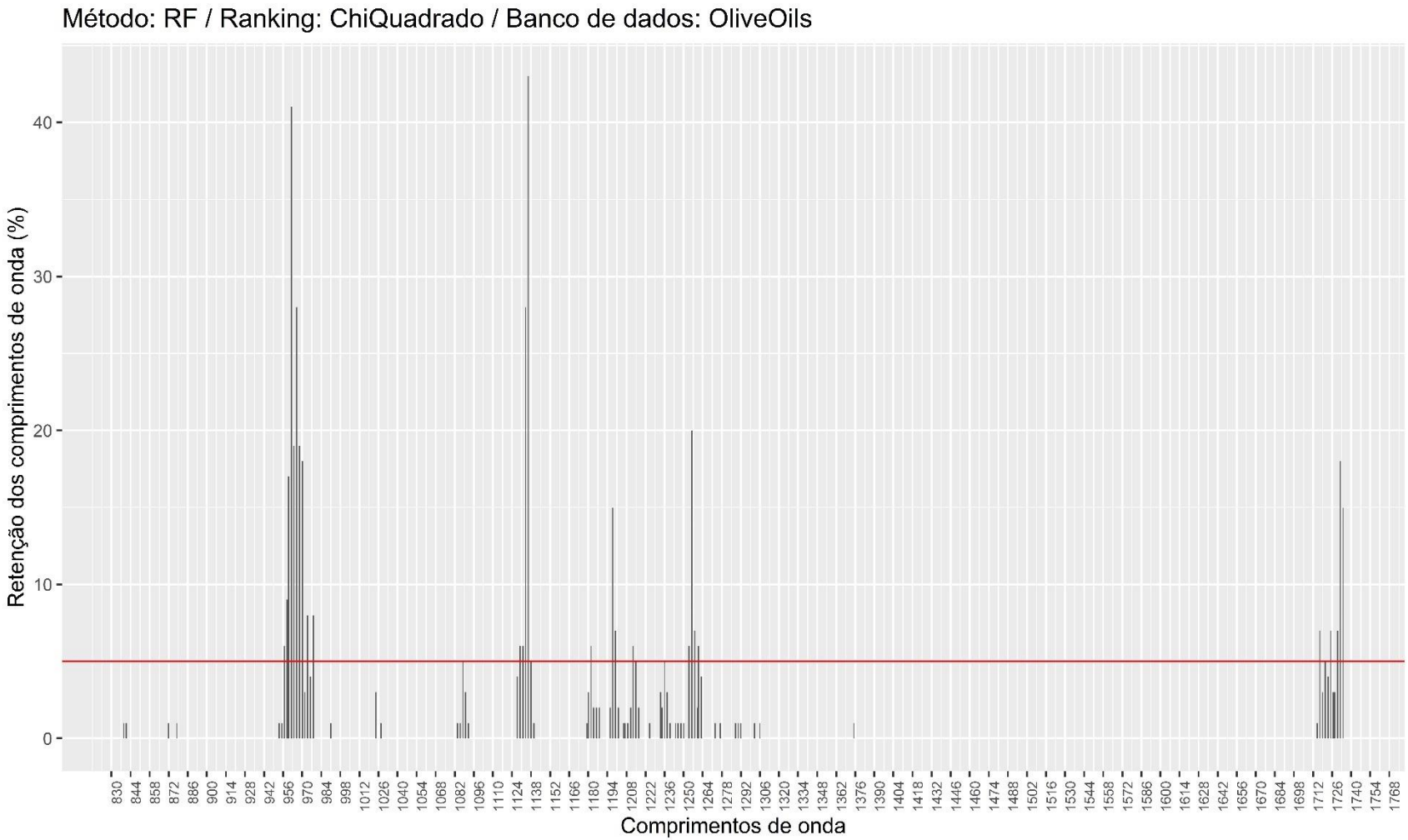

Figura 4.6– Regiões representativas do espectro da base azeite de oliva para χ^2 – RF

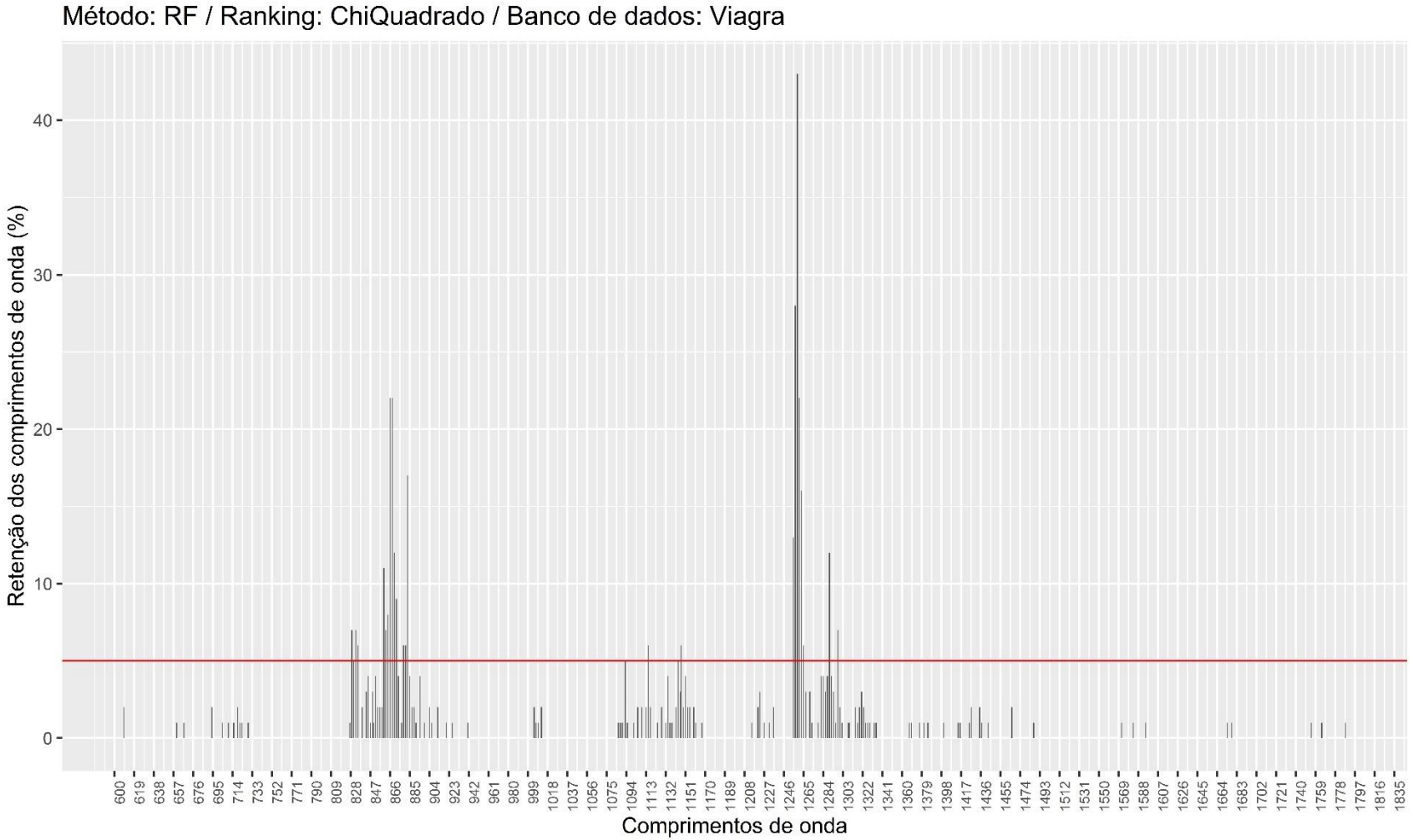

Figura 4.7– Regiões representativas do espectro da base Viagra®para χ^2 – RF

A Figura 4.8 apresenta a matriz de confusão média para cada banco de dados, com valores percentuais, construída à partir do conjunto de teste referente ao método χ^2 – RF. Para as bases de natureza binária, a classe 0 (zero) refere-se à amostras adulteradas e a classe 1 à amostras autênticas. Já para as bases de natureza multiclasse, cada classe representa um país. Para erva mate: classe 1 = Brasil, classe 2 = Paraguai, Classe 3 = Argentina e classe 4 = Uruguai; e para azeite de oliva: classe 1 = Grécia, classe 2 = Itália, classe 3 = Portugal e classe 4 = Espanha).

Figura 4.8 – Matriz de confusão média referente ao método χ^2 – RF para cada banco de dados.

Cialis®

Real \ Predição	0	1
0	27,76	0,24
1	0,77	71

Cocaina

Real \ Predição	0	1
0	45,66	0,43
1	0,25	53,66

Viagra®

Real \ Predição	0	1
0	54,58	3,56
1	5,16	36,70

Purê de Frutas

Real \ Predição	0	1
0	62,17	2,32
1	2,26	33,25

Erva Mate

Real \ Predição	1	2	3	4
1	23,73	1,36	0,09	2,09
2	1,55	31,36	3,36	0,09
3	0	2,73	6,36	0
4	2,36	0	0	24,90

Azeite de Oliva

Real \ Predição	1	2	3	4
1	14,41	2,03	0,55	0,24
2	1,07	26,14	0,03	0,34
3	0,59	0,069	12,41	0,72
4	0,79	2,597	0,97	37,03

A Tabela 4.5 apresenta a proporção de amostras por classe em cada banco de dados.

Tabela 4.5 – Percentual de amostras por classe para cada banco de dados.

Base de dados	***Cialis*®**	***Cocaina***	***Erva Mate***	***Purê de Frutas***	***Azeite de Oliva***	***Viagra*®**
Classe 0	28%	46%	-	64%	-	58%
Classe 1	72%	54%	26%	36%	17%	42%
***C**lasse 2*	-	-	35%	-	28%	-
Classe 3	-	-	13%	-	13%	-
Classe 4	-	-	26%	-	42%	-

4.4.2 Comparação com outros métodos da literatura para seleção de COs

Os resultados da abordagem proposta foram comparados com os resultados das pesquisas de Kahmann et al. (2018b) para as bases cialis® e viagra®, Anzanello et al. (2015) para Cocaina, Kahmann et al. (2017) para erva mate, Holland, Kemsley e Wilson (1998) para purê de frutas e Tapp, Defernez e Kemsley(2003) para azeite de oliva. Os autores citados utilizaram as bases de dados analisadas neste trabalho e a mesma medida para avaliação do desempenho dos modelos (acurácia). A tabela 4.6 apresenta uma síntese dos resultados do método proposto e dos métodos comparados.

Kahmann et al. (2018b) propuseram o uso de um Índice de Importância de Intervalos (III) baseado no teste de duas amostras Kolmogorov-Smirnov (TSKS), para orientar a inclusão de intervalos de número de onda relevantes e informativos no subconjunto analisado. TSKS é um teste de hipóteses frequentistas não paramétrico concebido originalmentepara comparar duas distribuições empíricas de frequência cumulativa. A estatística do teste TSKS é usada como uma medida da capacidade discriminante dos números de onda para separar amostras em duas classes. Depois de avaliado para cada número de onda, o III é calculado como a média dos valores estatísticos do teste TSKS dos números de onda em um determinado intervalo. Os intervalos de número de onda são então classificados de acordo com seus IIIs e inseridos no conjunto usado para classificação. A classificação das amostras é realizada por meio da ferramenta SVM.

Anzanello et al. (2015)criam um índice de importância do número de onda com base na distância de *Bhattacharyya* (BD). As amostras são então inseridas em duas classes por meio da técnica *Probabilistic Neural Network* (PNN) usando todos os COs e o desempenho da classificação é calculado. O número de onda com o menor índice é então removido do conjunto de dados e uma nova classificação é realizada. Esse procedimento iterativo, baseado na abordagem *backward,* é realizado até que um único CO reste, de acordo com a ordem sugerida pelo índice de BD. Os autores avaliam ainda o desempenho da classificação das técnicas *k - Nearest Neighbor* (KNN) e Análise Discriminante Linear (LDA).

Kahmann et al. (2017) propuseram um método de seleção de variáveis que computa as informações mútuas entre variáveis independentes e a variável de resposta, criando um Índice de Importância Variável (VII). Esse índice avalia a relevância da variável usando a otimização da programação quadrática (QP) para reduzir informações redundantes entre as variáveis retidas e maximizar seu relacionamento em relação ao local de origem da amostra e informação mútua. Em seguida, por meio de SVM, classificam iterativamente as amostras do conjunto de treinamento seguindo a abordagem *forward* de acordo com a ordem sugerida pelo VII, mantendo o subconjunto que produz o melhor resultado. A categorização da amostra é também realizada usando técnicas KNN e DA (análise discriminante).

Holland, Kemsley e Wilson (1998)utilizam todo o espectro para classificar as amostras do Purê de Frutas usando regressão PLS baseado no algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*). Por sua vez, Tapp, Defernez e Kemsley (2003)empregam GA-LDA no qual um algoritmo genético GA é usado para procurar um pequeno subconjunto de variáveis para passar para o LDA com validação cruzada. O critério para o término da evolução da GA foi alcançar uma taxa de sucesso de classificação de 100% ou nenhuma melhoria adicional na taxa de sucesso por 10 gerações. Os autores também utilizam uma abordagem com o espectro em sua totalidade (PLS-LDA), sendo ao mesmo aplicadaa regressão dos mínimos quadrados parciais (PLS), seguido de análise discriminante linear (LDA) baseada na métrica de distância de Mahalanobis. Como resultado, a abordagem GA-LDA produziu melhores resultados que PLS-LDA.

Tabela 4.6– Comparação dos resultados entre a abordagem proposta e métodos propostos por outros autores.

Base de dados	Autor	Abordagem	COs retidos (%)	Acc Media (%)
Cialis®	**Abordagem proposta**	**χ^2 – RF**	**0,711**	98,8
	(KAHMANN et al., 2018b)	III baseado TSKS – SVM	12,5	**99,87**
Cocaina	**Abordagem proposta**	**χ^2 – RF**	**0,343**	99
	(ANZANELLO et al., 2015)	BD – PNN	1,536	**99,96**
Erva Mate	**Abordagem proposta**	**χ^2 – RF**	0,109	84,2
	(KAHMANN et al., 2017)	QP – SVM	3,7	94,29
Purê de Frutas	**Abordagem proposta**	**χ^2 – RF**	4,853	**94,9**
	(HOLLAND; KEMSLEY; WILSON, 1998)	Regressão PLS	100	94,3
Azeite de Oliva	**Abordagem proposta**	**χ^2 – RF**	**0,938**	87,2
	(TAPP; DEFERNEZ; KEMSLEY, 2003)	GA-LDA	6,67	**100**
Viagra®	**Abordagem proposta**	**χ^2 – RF**	**0,842**	90
	(KAHMANN et al., 2018b)	IIIbaseado TSKS– SVM	25	**98.91**

Observa-se que o método proposto possui acurácia superior apenas à regressão PLS aplicada por Holland, Kemsley e Wilson (1998) em Purê de Frutas. Porém, em todas as comparações o desempenho do χ^2–RF mostrou-se superior com relação ao número de comprimentos de onda retidos. Ao reduzir os subconjuntos de preditores importantes, o método garante ao usuário a entrega de informações que realmente contribuem para a análise do processo de aferição da qualidade, tendo como consequência positiva a simplificação da análise do espectro. Vale ressaltar que o método proposto no presente estudo busca se adaptar a bancos de dados de diferentes naturezas (binários e multiclasses) gerando bons resultados para seis diferentes bases de dados. Sendo assim, cabe ao usuário decidir se prefere penalizar o custo computacional em prol de uma acurácia mais alta, ou se o custo envolvido vale uma taxa de acerto menor.

4.5 Conclusão

Técnicas de espectrometria são consideradas uma poderosa ferramenta para detectar falsificações em produtos, porém estas usualmente dão origem a um grande número de

comprimentos de onda altamente ruidosos e correlacionados, o que tende a reduzir o desempenho preditivo de várias técnicas multivariadas. Por esta razão, o uso de métodos de seleção de variáveis em dados de origem espectral torna-se cada vez mais necessário. Tendo em vista a crescente demanda por métodos de redução de dimensionalidade para dados desta natureza, o presente artigo propôs um novo método para selecionar comprimentos de onda aplicados ao campo da ciência forense. O método identifica amostras autênticas e não autênticas de produtos. A questão da autenticidade inclui produtos que são fabricados em determinadas regiões e vendidos como se fossem de outra região, tipicamente de maior valor agregado. Isto confere ao infrator uma margem de lucro maior, uma vez queconsegue aplicar um preço mais elevado ao produto.

O método proposto neste estudo pode ser utilizado para construir modelos de precisão para o monitoramento da qualidade de produtos tanto em problemas binários quanto multiclasses. Este caracteriza-se como um *wrapper*, o qual integra a técnica estatística χ^2 com o algoritmo de aprendizado de máquina RF de acordo com a abordagem *foward inclusion stepwise*. Foram empregados seis bancos de dados espectrais para a avaliação de desempenho do método:drogas lícitas (cialis® e viagra®), ilícita (cocaína) eprodutos alimentícios (purê de frutas, erva mate e azeite de oliva). Os bancos referentes à erva mate e o azeite de oliva caracterizam problemas multiclasse, enquanto o restante configura problemas binários. As quatro etapas desse método são: (*i*) pré-processar os dados, ajustando uma escala única para todas as variáveis, (*ii*) dividir o conjunto de dados em conjuntos de conjuntos de calibração e validação, (*iii*) ranquear os comprimentos de onda e (*iv*) selecionar comprimentos de onda.

Para cialis® e cocaina o método proposto apresenta uma acurácia média acima de 98% com um percentual de variáveis retidas inferior a 0,8%. O desempenho na base purê de frutas foi de 95% de acurácia média e aproximadamente 5% de variáveis retidas. Na base viagra®, 90% de acurácia média e 0,85% de variáveis retidas. Quanto às bases multiclasses, Erva Mate e azeite de oliva, foi possível identificar que o método apresentou maior dificuldade uma vez que suas acurácias médias ficaram abaixo de 90%. Erva mate com 84% de acurácia e 0,11% de variáveis retidas e azeite de oliva com 87% e 0,94% de variáveis retidas. Estes resultados podem se justificar pela maior complexidade de problemas multiclasses. Porém, no geral, o método se mostra eficaz ao reduzir o número de variáveis além de apresentar acurácias satisfatórias.

Quando comparado a outros métodos, é possível notar que o método proposto não é soberano no que diz respeito à obtenção da máxima acurácia. Porém, ainda assim conduziu a resultados satisfatórios, além de reduzir substancialmente o número de variáveis selecionadas comparado aos outros métodos. Vale ressaltar que o método proposto apresentou boa robustez ao se adaptar a bancos com classes binárias e multiclasses (esse último especialmente complexo), gerando bons resultados para seis diferentes bases de dados. Sendo assim, cabe ao usuário decidir sobre o *trade-off* custo computacional versus acurácia de classificação.

Uma maneira de melhorar a robustez dos algoritmos de seleção de variáveis é utilizar a abordagem ensemble – técnica de combinar vários modelos para resolver o mesmo problema – (RIBEIRO; DOS SANTOS COELHO, 2020) baseando-se no pressuposto de que a combinação da saída de vários especialistas é melhor do que a saída de um único especialista. Embora a abordagem ensemble tenha provado sua eficácia nos últimos anos, a aplicação em outras disciplinas do aprendizado de máquina, como a seleção de variáveis, tem sido pouco explorada. Em geral, este tipo de técnica tem ganhado destaque em aplicações dentro do processo de aprendizado dos algoritmos (predição/classificação) (BOLÓN-CANEDO; ALONSO-BETANZOS, 2019).

Assim, para pesquisas futuras, aconselha-se um estudo da estratégia *ensemble* para unificar diferentes técnicas de ranqueamento e aprimorar o desempenho de métodos *wrappers*. A ideia é desenvolver uma nova abordagem de ranqueamento com base na estratégia *ensemble*, onde múltiplas técnicas de ranqueamento sejam combinadas.

4.6 Referências

ALMUALLIM, H.; KANEDA, S.; AKIBA, Y. Development and Applications of Decision Trees. **Expert Systems**. p.53–77, 2002.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97–105, 2012.

ANZANELLO, M. J.; FOGLIATTO, F. S. A review of recent variable selection methods in industrial and chemometrics applications. **European Journal of Industrial Engineering**, v. 8, n. 5, p. 619–645, 2014.

ANZANELLO, M. J.; KAHMANN, A.; MARCELO, M. C. A.; et al. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, v. 115, p. 562–569, 2015.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGERB, R. P.; MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209–214, 2013.

ASDAGHI, F.; SOLEIMANI, A. An effective feature selection method for web spam detection. **Knowledge-Based Systems**, v. 166, p. 198–206, 2019.

BERTOL, E.; MARI, F.; MILIA, M. G. DI; et al. Determination of aminorex in human urine samples by GC-MS after use of levamisole. **Journal of Pharmaceutical and Biomedical Analysis**, v. 55, n. 5, p. 1186–1189, 2011.

BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Ensembles for feature selection: A review and future trends. **Information Fusion**, v. 52, p. 1–12, 2019.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.

BRUNT, T. M.; RIGTER, S.; HOEK, J.; et al. An analysis of cocaine powder in the Netherlands: content and health hazards due to adulterants. **Addiction**, v. 104, n. 5, p. 798–805, 2009.

CÂMARA, A. B. F.; DE CARVALHO, L. S.; DE MORAIS, C. L. M.; et al. MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends. **Fuel**, v. 210, p. 497–506, 2017.

COCCHI, M.; BIANCOLILLO, A.; MARINI, F. Chemometric Methods for Classification and Feature Selection. **Comprehensive Analytical Chemistry**. v. 82, p.265–299, 2018.

DE SANTANA, F. B.; BORGES NETO, W.; POPPI, R. J. Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. **Food Chemistry**, v. 293, p. 323–332, 2019.

DOS SANTOS, M. K.; DE CASSIA MARIOTTI, K.; KAHMANN, A.; et al. Comparison between counterfeit and authentic medicines: A novel approach using differential scanning calorimetry and hierarchical cluster analysis. **Journal of Pharmaceutical and Biomedical Analysis**, v. 166, p. 304–309, 2019.

ELSALAMONY, H. A. Detecting distorted and benign blood cells using the Hough transform based on neural networks and decision trees. **Emerging Trends in Image Processing, Computer Vision and Pattern Recognition**. p.457–473, 2015.

FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, A. C. P. DE L. F. DE. **Inteligência artificial: uma abordagem de aprendizado de máquina**. LTC ed. Rio de Janeiro, 2011.

FERNANDEZ, F. M.; HOSTETLER, D.; POWELL, K.; et al. Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries. **The Royal Society of Chemistry**, v. 136, n. Analyst, p. 3073–3082, 2011.

FERRARI, D. G.; SILVA, L. N. D. C. **Introdução a mineração de dados**. Saraiva ed. São Paulo, 2016.

GARCIA, M. C. DE M. **Avaliação de métodos de data mining e regressão logística aplicados na análise de traumatismo cranioencefálico grave**, 2015. Universidade Federal de Santa Catarina.

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, p. 171–193, 2001.

GAYDOU, V.; KISTER, J.; DUPUY, N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. **Chemometrics and Intelligent Laboratory Systems**, v. 106, p. 190–197, 2011.

HE, X.; CAI, D.; NIYOGI, P. Laplacian Score for Feature Selection. **Advances in neural information processing systems**, p. 507- 514., 2006.

HINDAWI, M. **Feature selection for semi-supervised data analysis in decisional information systems.**, 2013.

HOLLAND, J. K.; KEMSLEY, E. K.; WILSON, R. H. Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. **Journal of the Science of Food and Agriculture**, v. 76, n. 2, p. 263–269, 1998.

JADHAV, S.; HE, H.; JENKINS, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. **Applied Soft Computing Journal**, v. 69, p. 541–553, 2018.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; MARCELO, M. C. A.; et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018a.

KAHMANN, A.; ANZANELLO, M. J.; FOGLIATTO, F. S.; CHAOVALITWONGSE, W. A.; et al. Interval importance index to select relevant ATR-FTIR wavenumber Intervals for falsified drug classification. **Journal of Pharmaceutical and Biomedical Analysis**, v. 158, p. 494–503, 2018b.

KAHMANN, A.; ANZANELLO, M. J.; MARCELO, M. C. A.; POZEBON, D. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in**

Agriculture, v. 140, p. 348–360, 2017.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1–2, p. 273–324, 1997.

LIAW, A.; WIENER, M. Breiman and Cutler's Random Forests for Classification and Regression. , 2018.

LIU, H.; MOTODA, H. **Computational Methods of Feature Selection**. Taylor and ed. 2007.

LOPES, M. V. R. **Tratamento de imprecisão na geração de árvores de decisão**, 2016. São Carlos: Universidade Federal De São Carlos.

MLADENIÉ, D. Feature Selection for Dimensionality Reduction. **International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection"**, p. 84–102, 2006.

NOVAKOVIĆ, J.; STRBAC, P.; BULATOVIĆ, D. Toward optimal feature selection using ranking methods and classification algorithms. **Yugoslav Journal of Operations Research**, v. 21, p. 119–135, 2011.

OLIVEIRA, E. F. T. DE; GRÁCIO, M. C. C. Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação. **Revista de Ciência da Informação**, v. 6, n. 3, p. 1–11, 2005.

ORTIZ, R. S.; MARIOTTI, K. DE C.; FANK, B.; et al. Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the same pharmaceutical powder mixture be used to falsify two medicines? **Forensic Science International**, v. 226, n. 1–3, p. 282–289, 2013.

PEREIRA, R. B. **Seleção Lazy de atributos para a tarefa de classificação**, 2009. Niterói: Universidade Federal Fluminense.

R CORE TEAM. R: A Language and Environment for Statistical Computing. , 2018. Vienna, Austria.

REID, L. M.; O'DONNELL, C. P.; DOWNEY, G. Recent technological advances for the determination of food authenticity. **Trends in Food Science and Technology**, v. 17, n. 7, p. 344–353, 2006.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in Biology and Medicine**, v. 112, p. 103375, 2019.

REYES, O.; MORELL, C.; VENTURA, S. Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. **Neurocomputing**, v. 161, p. 168–182, 2015.

RIBEIRO, M. H. D. M.; DOS SANTOS COELHO, L. Ensemble approach based on

bagging, boosting and stacking for short-term prediction in agribusiness time series. **Applied Soft Computing Journal**, v. 86, p. 105837, 2020.

ROBNIK-SIKONJA, M.; SAVICKY, P. Package 'CORElearn': Classification, Regression and Feature Evaluation. , 2018.

RODRIGUES, N. V. S.; CARDOSO, E. M.; ANDRADE, M. V. O.; DONNICI, C. L.; SENA, M. M. Analysis of Seized Cocaine Samples by using Chemometric Methods and FTIR Spectroscopy. **Article J. Braz. Chem. Soc**, v. 24, n. 3, p. 507–517, 2013.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics Review**, v. 23, p. 2507–2517, 2007.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 171–178, 2017.

SOMAN, K. P.; DIWAKAR, S.; AJAY, V. **Data mining: theory and practice [with CD]**. PHI learning Private Limited, 2006.

STAŃCZYK, U.; ZIELOSKO, B.; JAIN, L. C. **Intelligent Systems Reference Library 138 Advances in Feature Selection for Data and Pattern Recognition**. Springer I ed. 2018.

SUH, C.; YOU, K. Rdimtools: Dimension Reduction and Estimation Methods. , 2018.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining : mineração de dados**. Ciencia Moderna, 2009.

TAPP, H. S.; DEFERNEZ, M.; KEMSLEY, E. K. FTIR Spectroscopy and Multivariate Analysis Can Distinguish the Geographic Origin of Extra Virgin Olive Oils. , 2003.

WHO. **Guidelines for the development of measures to combat counterfeit drugs**. 1999.

ZGUROVSKY, M. Z.; ZAYCHENKO, Y. P. **Big Data: Conceptual Analysis and Applications**. Springer I ed. 2020.

5. Considerações finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

Esta dissertação teve como principal objetivo o desenvolvimento de novas abordagens para seleção de COs com vistas à classificação e predição de propriedades de produtos que, ao serem descritos por dados espectrais, acabam por se apoiar em um elevado número de COs ruidosos e altamente correlacionados.

Para alcançar o objetivo principal, foram delimitados os seguintes objetivos específicos: (*i*) Avaliaras principais técnicas de análise multivariada de dados, sua organização e contribuição para o processo de seleção de variáveis; (*ii*)Criar um índice de importância de variáveis com base na lei de *Lambert-Beer*com vistas a mensurar a relevância das variáveis (COs) para fins de predição de propriedades de amostras; (*iii*) Avaliar o desempenho da técnica floresta aleatória na calibração multivariada de dados espectroscópicos; (*iv*) Avaliar o desempenho dediferentes metodologias de ranqueamento de importância de variáveis; (*v*) Comparar os resultados dos métodos propostos a outras abordagens voltadas à seleção de variáveis; e (*vi*) Avaliar o desempenho das proposições em dados de NIR com diferentes dimensionalidades e tipos de variável resposta (discreta e contínua).

O objetivo (*i*) foi atingido no primeiro artigo, o qual apresentou os princípios e principais mecanismos dos métodos de seleção de variáveis, apresentando a estrutura geral para processos de seleção de variáveis, bem como a categorização de diversas abordagens, a fim de facilitar o entendimento das diversas abordagens existentes e oferecer subsídios para o desenvolvimento de novos algoritmos de seleção de variáveis.

O objetivo (*ii*) foi alcançado na primeira fase do segundo artigo, que propôs uma sistemática de seleção de COs para predição de propriedades químicas de combustível diesel apoiada em conceitos da lei de *Lambert-Beer* para o ordenamento dos COs que, posteriormente, são inseridos na ferramenta preditiva ANN de acordo com a abordagem de adição de variáveis (*forward*). Ao ser validado em três bancos distintos relativos a propriedades de diesel, o melhor

resultado para a base de dados ponto de ebulição reteve cerca de 2,74% dos comprimentos de onda, pouco menos de 1% para número de cetano, e 0,69% para ponto de fulgor. O RMSE médio para todos os conjuntos de dados ficou abaixo de 0,14.

Os objetivos (*iii*) e (*iv*) foram alcançados no terceiro artigo, onde foi estudada a utilização da ferramenta de aprendizado de máquina RF como alternativa para a calibração multivariada em seis bancos de dados de espectroscopia de natureza binária e multiclasse, com o objetivo de identificar amostras adulteradas em produtos alimentícios e em drogas (lícitas e ilícitas). Foram testadas seis diferentes técnicas de ranqueamento (χ^2, ReF, LS, GI, RG e IG) a fim de orientar a seleção das variáveis na ferramenta classificadora RF, de acordo com a abordagem *forward inclusion stepwise* seguindo o índice de importância de COs gerado por cada uma delas. A melhor combinação de *ranking* com RF, χ^2 – RF, conduziu a resultados satisfatórios em problemas binários e multiclasses. As bases caracterizadas por problemas binários obtiveram acurácia média acima de 90%, com um percentual de COs retidos inferior a 1% (com exceção de purê de frutas, que reteve uma porcentagem superior a 5%). As bases cialis® e cocaína se destacaram ao obter valores acima de 98% de acurácia média. Já as bases caracterizadas por problemas multiclasses obtiveram maior dificuldade na classificação, resultando em acurácias médias de 84% e 87% para erva mate e azeite de oliva, respectivamente, com cerca de apenas 1% de COs retidos.

Por fim, os objetivos (*v*) e (*vi*) foram alcançados no segundo e terceiro artigo, onde comparou-se os resultados das metodologias propostas a outras abordagens propostas pela literatura. O método proposto no segundo artigo mostrou-se superior a todos àqueles com os quais foi comparado, em termos de desempenho de predição e percentual de COs retidos. Já a abordagem proposta no terceiro artigo mostrou-se superior com relação à quantidade de COs retidos,porém conduziu a acurácias ligeiramente inferiores aos concorrentes. Ainda assim, o método proposto gerou bons resultados para seis diferentes bases de dados de diferentes naturezas (binária e multiclasse), tendo em vista que os métodos comparados foram desenvolvidos para resolver um problema específico.

Dentre os dois métodos abordados nesta dissertação, percebe-se que o método que combina o índice de importância baseado na lei de Lambert-Beer com ANN foi o que obteve maior destaque, conduzindo aos menores erros (RMSE) e retendo um percentual muito menor

de COs do que nos métodos concorrentes. Assim, os resultados obtidos neste artigo demonstram a robustez do método e sugerem que os métodos de seleção de variáveis podem ser simplificados usando uma estratégia de análise de banda de dados e, no entanto, apresentarão alto desempenho. Conclui-se que todos os objetivos específicos foram alcançados, permitindo afirmar que o objetivo principal desta dissertação foi cumprido.

5.2 Sugestões para trabalhos futuros

Como extensões das proposições apresentadas nessa dissertação, sugerem-se as seguintes pesquisas futuras:

- Propor um índice de consistência da posição de variáveis de forma a penalizar aqueles *rankings* que apresentem alto grau de dispersão de relevância, assim como baixa estabilidade;
- Utilizar a estratégia *ensemble* para unificar técnicas de ranqueamento e aprimorar o desempenho de métodos *wrappers*; e
- Estudo do grau de relação/iteração existente entre as variáveis independentes a fim de otimizar a geração dos subconjuntos.