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

Juliana de Abreu Fontes

ESTRATÉGIAS DE APRENDIZADO DE MÁQUINA PARA APERFEIÇOAMENTO DO CONTROLE DE QUALIDADE DE PRODUTOS

Porto Alegre
2024

Juliana de Abreu Fontes

Estratégias de aprendizado de máquina para aperfeiçoamento do controle de qualidade de produtos

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutor em Engenharia de Produção, na área de concentração em Sistemas de Qualidade.

Orientador: Professor Michel J. Anzanello, *Ph*.D.

Porto Alegre

2024

Juliana de Abreu Fontes

Estratégias de aprendizado de máquina para aperfeiçoamento do controle de qualidade de produtos

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia 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. Orientador 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, Ph.D (ITA)
Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

Dedicatória

Dedico esta conquista à memória do meu companheiro, cujo espírito e incentivo continuam a me guiar. Este trabalho é uma homenagem a ele e a todos que me ajudaram a transformar um momento de dor em uma realização significativa.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que contribuíram de alguma forma para que eu chegasse até aqui. Este trabalho não seria possível sem o apoio de muitas pessoas que estiveram ao meu lado ao longo desta jornada.

Em primeiro lugar, agradeço a Deus, por me conceder saúde e força para superar as dificuldades ao longo deste caminho.

Aos meus pais, pelo amor, incentivo e apoio incondicional, que sempre me sustentaram nos momentos mais desafiadores.

Ao meu irmão, por me ouvir pacientemente durante os treinos das minhas apresentações de qualificação e defesa, e por me dar feedback valioso, mesmo quando ambos sabíamos que ele preferiria estar assistindo ao YouTube. Sua presença e apoio foram inestimáveis.

À minha família do coração, família do meu companheiro, quero expressar minha profunda gratidão por fazerem parte da minha vida. A trágica e precoce perda dele foi um dos momentos mais difíceis que já enfrentei, mas o amor e o apoio de vocês foram essenciais para que eu pudesse seguir em frente. Vocês me lembraram constantemente do incentivo e da motivação que ele sempre me deu nos estudos. Sem vocês, essa jornada teria sido ainda mais desafiadora.

A cada parente que celebrou minhas vitórias como se fossem suas, meu sincero agradecimento por fazerem parte desta conquista, enriquecendo-a com seu amor e orgulho.

Aos meus amigos de longa data, que, mesmo à distância, enviaram palavras de incentivo e positividade, tornando-se pilares essenciais do meu sucesso.

A todos que entenderam minhas ausências, sabendo que cada momento estava sendo dedicado à batalha épica com minha tese.

Aos professores, que me acompanharam ao longo do curso e que, com empenho, se dedicam à arte de ensinar. Em especial, ao meu orientador, Michel Anzanello, pelo empenho dedicado à elaboração deste trabalho e pelo apoio inestimável durante o momento mais difícil da minha vida. Sua orientação e suporte foram essenciais para que eu pudesse continuar e

concluir esta jornada. Ao professor Marcelo Guterres, cujo apoio desde a graduação foi fundamental para me encorajar a alçar voos mais altos. Sem dúvida, ele é um dos grandes responsáveis por eu ter alcançado o título de doutora em Engenharia de Produção na UFRGS. À banca, pela disponibilidade e contribuições para a adequação dessa tese.

Ao meu colega de jornada, João Brito, pela parceria constante, socorro e apoio inestimável, além de me ensinar a arte da programação.

Ao Guilherme Bucco, pelas suas valiosas contribuições e apoio aos meus artigos.

Ao PPGEP/UFRGS que me acolheu, desafiou e transformou.

Ao CNPq, pela bolsa de financiamento sem a qual essa tese não teria sido realizada.

RESUMO

A gestão e análise de grandes volumes de dados são desafios complexos impulsionados pelo avanço tecnológico no monitoramento de processos e produtos. Neste contexto, destaca-se a relevância da seleção criteriosa de variáveis e de estratégias avançadas de pré-processamento, como fusão de dados e engenharia de variáveis, para aprimorar o desempenho dos modelos de aprendizado de máquina. A presente tese apresenta proposições inovadoras para otimizar o controle da qualidade e autenticidade de produtos, reduzindo custos operacionais e melhorando a performance dos modelos analíticos. O primeiro artigo visa identificar as técnicas analíticas e variáveis mais relevantes para avaliar a autenticidade de amostras de Cialis® e Viagra®. Para tanto, integra a estratégia de fusão de dados de baixo nível (LLDF) com o algoritmo de classificação XGBoost. Na sequência, é realizada uma análise descritiva detalhada dos achados, evidenciando como a combinação dessas técnicas não apenas proporciona resultados numéricos precisos, mas também direciona a análise para uma interpretação mais detalhada do problema. O segundo artigo propõe uma estrutura de duas fases que incorpora uma etapa inicial de pré-seleção de comprimentos de onda (COs) orientada por agrupamento de CO, integrada a uma abordagem baseada em wrapper. A proposta foi aplicada a 11 conjuntos de dados FTIR/NIR de diferentes domínios, com o objetivo de classificar amostras em níveis de qualidade e autenticidade. Por fim, o terceiro artigo aborda o desenvolvimento e implementação de um método que combina etapas de seleção e de engenharia de variáveis. O estudo avalia a contribuição de cada etapa do método proposto no aprimoramento da eficácia dos modelos de aprendizado de máquina. A proposta foi validada em 8 conjuntos de dados FTIR/NIR de diferentes domínios com o objetivo de classificar amostras em níveis de qualidade e autenticidade.

Palavras-chave: Seleção de variáveis. Seleção de técnicas analíticas. Engenharia de variáveis. Classificação. Cluster de variáveis.

ABSTRACT

The management and analysis of large data volumes are complex challenges driven by technological advancements in the collection and monitoring of processes and products. Emphasizing the importance of careful feature selection and advanced preprocessing strategies, such as data fusion and feature engineering, enhances machine learning model performance. This thesis presents innovative propositions to optimize product quality and authenticity control, reduce operational costs, and improve analytical model performance. The first article aims to identify the most relevant analytical techniques and variables for evaluating the authenticity of Cialis® and Viagra® samples. It integrates the low-level data fusion (LLDF) strategy with the XGBoost classification algorithm and provides a detailed descriptive analysis of the findings. This combination not only delivers precise numerical results but also guides a more detailed interpretation of the problem. The second article proposes a two-phase framework that incorporates an initial pre-selection stage of wavelengths (COs) guided by wavelength clustering, integrated with a wrapper-based approach. The proposal was applied to 11 FTIR/NIR datasets from different domains to classify samples into quality and authenticity levels. Finally, the third article addresses the development and implementation of a method combining feature selection and feature engineering stages. The study evaluates the impact of the proposed approach and the contribution of each stage to the efficacy of machine learning models. The proposal was validated on 8 FTIR/NIR datasets from different domains to classify samples into quality and authenticity levels.

Keywords: Feature selection. Analytical technique selection. Feature engineering. Classification. Variable clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

FE	*Feature Engineering*
ML	*Machine Learning*
XRF	*X-ray Fluorescence*
ESI-MS/MS	*Electrospray Ionization Tandem Mass Spectrometry*
UPLC-MS/MS	*Ultra-Performance Liquid Chromatography Mass Spectrometry*
ATR–FTIR	*Attenuated Total Reflectance Fourier-Transform Infrared*
NIR	*Near-Infrared Spectroscopy*
LLDF	*Low Level Data Fusion*
XGBoost	*Extreme Gradient Boosting*
SC	*Spectral Clustering*
kNN	*k-Nearest Neighbor*
SVM	*Support Vector Machine*
RF	*Random Forest*
LR	*Logistic Regression*
NB	*Naiive Bayes*
DT	*Decision Tree*
PCA	*Principal Component Analysis*
PLS	*Partial Least Squares*
NLPCA	*Nonlinear Principal Component Analysis*
LDA	*Linear Discriminant Analysis*
SVD	*Singular Value Decomposition*
SVR	*Support Vector regression*
SFFS	*Sequential Forward Floating Selection*
CoefVar	Coeficiente de Variação
BD	*Bhattacharyya Distance*
χ^2	*Qui-quadrado*
ReF	*ReliefF*
GI	*Gini*
PR–AUC	*Precision-Recall – Area Under the Curve*
CO	*Comprimento de onda*
IFA	Ingrediente Farmacêutico Ativo

SUMÁRIO

1 INTRODUÇÃO

Os avanços tecnológicos têm impulsionado notáveis progressos na gestão e no armazenamento de variadas e elevadas quantidades de dados em diversos segmentos e áreas (CORALLO et al., 2023; OUSSOUS et al., 2018). Embora os dados constituam a base essencial da informação, por si só não fornecem uma mensagem clara que permita compreender uma determinada situação. Assim, torna-se crucial recorrer a técnicas de análise de dados para extrair informações relevantes e aplicáveis. Nesse sentido, os algoritmos de aprendizado de máquina, também conhecidos como *machine learning (*ML), têm se destacado no meio acadêmico (ZHONG et al., 2021). Esses algoritmos referem-se a métodos de análise de dados que automatizam a construção de modelos analíticos, baseando-se na capacidade dos sistemas de aprender com os dados, identificar padrões e tomar decisões com mínima intervenção humana (HAMET; TREMBLAY, 2017).

Em paralelo, percebe-se um desafio crescente relacionado à alta dimensionalidade dos dados coletados de processos de controle e monitoramento de qualidade e autenticidade de produtos, no qual o número de variáveis preditivas pode ser significativamente maior do que o número de observações ou amostras (ABDULWAHAB; AJITHA; SAIF, 2022). Conforme apontado por Salimi et al. (2018), grandes conjuntos de dados frequentemente incluem variáveis irrelevantes ou redundantes, resultando no aumento desnecessário da complexidade da análise. Dados de alta dimensão, além de serem mais propensos à multicolinearidade entre os preditores (VASCONCELOS, 2017), impactam negativamente a eficiência e a precisão dos algoritmos de ML (BZDOK; KRZYWINSKI; ALTMAN, 2018; LEE; LOH; CHIN, 2017; URBANOWICZ et al., 2018).

Diante desse cenário desafiador, métodos de seleção de variáveis têm atraído grande interesse na área de ML, tanto em abordagens supervisionadas, quanto semi-supervisionadas e não supervisionadas. A seleção de variáveis, também conhecida como *feature selection* (FS), é uma das principais estratégias para reduzir a dimensionalidade dos dados (ABDULWAHAB; AJITHA; SAIF, 2022). Seu objetivo principal é identificar o conjunto de variáveis relevantes que mantêm as informações essenciais e a estrutura dos dados, eliminando as que são irrelevantes ou redundantes (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015;YAN et al., 2020). Assim, espera-se que as abordagens de FS melhorem a interpretabilidade dos modelos propostos, bem como a precisão e o desempenho dos métodos analíticos, resultando em modelos mais rápidos, confiáveis e

com menor custo computacional (COCCHI; BIANCOLILLO; MARINI, 2018; SAEYS; INZA; LARRAÑAGA, 2007; XIAOBO et al., 2010; YUN et al., 2019).

Existe uma vasta quantidade de métodos de seleção de variáveis na literatura. Tais métodos incluem três abordagens tradicionais: *filter*, *wrapper* e *embedded*. Métodos *filter* operam diretamente no banco de dados, e estão tipicamente apoiados em testes estatísticos que avaliam a significância das variáveis (SAEYS; INZA; LARRAÑAGA, 2007). Métodos *wrapper*, por sua vez, realizam uma busca entre os possíveis subconjuntos a serem avaliados e, conforme explicado por Kohavi e John (1997), em vez de usar um teste independente como abordagens *filter*, utilizam o próprio algoritmo de indução para avaliar os subconjuntos de variáveis de acordo com a sua capacidade preditiva. Já em abordagens do tipo *embedded*, a seleção do subconjunto é embutida ou integrada no próprio algoritmo de aprendizado (CHANDRASHEKAR; SAHIN, 2014).

Visando otimizar o desempenho dos algoritmos de ML, é imperativo que as variáveis preditivas capturem aspectos cruciais do problema para auxiliar o modelo na aprendizagem do resultado desejado. No entanto, se essas variáveis não descreverem adequadamente o problema subjacente ou não influenciarem de maneira significativa, podem distorcer os resultados (RAWAT; KHEMCHANDANI, 2019). Assim, além do desenvolvimento e aplicação de técnicas de FS, a combinação eficiente dessas técnicas com estratégias de pré-processamento de dados, como fusão de dados e *feature engineering* (FE, também conhecida como engenharia de variáveis), proporciona uma análise mais robusta do problema ao capturar padrões implícitos de forma mais precisa. A fusão de dados possibilita a integração de informações provenientes de diversas fontes, enriquecendo a análise e contribuindo para uma análise mais completa e eficiente dos dados (BORRÀS et al., 2016; DOS SANTOS et al., 2023; PIZARRO et al., 2013). Além disso, quando combinada com a FS, pode ser usada como ferramenta para a seleção da técnica analítica mais adequada para determinado contexto de aplicação, auxiliando na redução de custos operacionais e financeiros no que diz respeito ao processo de obtenção de dados. Por outro lado, a FE se concentra em estratégias direcionadas à criação e transformação de variáveis, com o objetivo de remapear os dados de maneira que facilite a discriminação das amostras. Essas abordagens visam fornecer *insights* adicionais, aprimorando a detecção de padrões e tendências ocultas nos dados, contribuindo para a melhoria do desempenho dos modelos analíticos (KHURANA et al., 2016).

Diante deste cenário, a presente tese visa desempenhar um papel significativo no avanço da redução da complexidade de modelos de ML, propondo o desenvolvimento e aplicação de métodos inovadores para tal finalidade. Nessa perspectiva, além da seleção de variáveis, esta tese também aborda a seleção das técnicas analíticas mais adequadas para obtenção de dados. O objetivo é classificar amostras e aprimorar o controle de qualidade de produtos em diversos setores industriais, simplificando a aplicação de testes para uma triagem eficaz, ao mesmo tempo em que custos operacionais e financeiros associados à geração de dados são reduzidos. As abordagens aqui apresentadas combinam técnicas de ML existentes, com ajustes em seus mecanismos, visando elevar a precisão e robustez dos modelos propostos. As técnicas de FS serão integradas a diferentes ferramentas de classificação e de medição de importância de variáveis, sendo posteriormente comparadas com métodos consolidados da literatura para avaliar seu desempenho. Por fim, também foram avaliadas estratégias de FE no desempenho dos algoritmos de ML.

1.1 Tema e objetivos

O tema desta tese é a proposição de novas abordagens para redução da complexidade de modelos de ML, permeando os contextos de seleção de técnica analítica, seleção de variáveis e transformação de variáveis, em ambientes de alta dimensionalidade com fins de categorização de produtos em níveis de qualidade e autenticidade.

Como objetivos específicos lista-se:

(i) Avaliar a eficácia da fusão de dados por meio de um método *embedded* na identificação das técnicas analíticas e variáveis mais eficientes para categorizar amostras de medicamentos como autênticas ou falsificadas;

(ii) Propor e validar uma sistemática de seleção de comprimentos de onda (COs) mais informativos com vistas à classificação binária de amostras em diferentes contextos baseado em agrupamento de variáveis;

(iii) Propor e validar uma sistemática que combina estratégias de FS e de FE com vistas ao aumento da eficiência das técnicas de classificação;

(iv) Avaliar o uso combinado de diferentes ferramentas de classificação com diferentes índices de importância de variáveis; e

(v) Avaliar a robustez dos métodos propostos em bancos de dados de processos diversos frente a outros métodos de seleção reportados pela literatura.

1.2 justificativa do tema e dos objetivos

De acordo com Salimi et al. (2018), elevados volumes de dados trazem consigo variáveis irrelevantes ou redundantes que aumentam a dimensionalidade e complexidade do espaço de atributos. Além disso, autores como Saeys et al. (2007), Urbanowicz et al. (2018) e Lee et al. (2017) afirmam que espaços de alta dimensão tendem a prejudicar o desempenho dos algoritmos de aprendizagem no que se refere à velocidade e taxa de acerto. Isso ocorre porque cada atributo *d* pode ser visto como uma coordenada do espaço *d*-dimensional (FACELI et al., 2011). Conforme o número de dimensões cresce, as instâncias tornam-se mais dispersas no espaço, havendo menos instâncias por região e tornando a tarefa de aprendizado mais difícil, uma vez que algoritmos de ML constroem preditores com base nas proporções de instâncias estimadas em cada classe por regiões do espaço. Consequentemente, o pré-processamento e modelagem dos dados passou a exigir abordagens mais complexas (ANZANELLO et al., 2013).

Além disso, com o avanço das tecnologias computacionais para coleta e monitoramento de processos e produtos, a aplicação de diferentes técnicas analíticas em um mesmo conjunto de amostras tem se tornado cada vez mais comum. Por exemplo, em ensaios laboratoriais voltados ao controle de qualidade, a mesma amostra pode ser analisada via técnicas de infravermelho e Raman, dentre outras. Contudo, essa prática pode resultar na escalada do volume de dados, o que eleva os custos operacionais e financeiros associados à gestão de dados. Nesse contexto, a seleção criteriosa da técnica analítica mais adequada para um determinado contexto não só otimiza a alocação de recursos e tempo, mas também contribui para a construção de modelos mais robustos e confiáveis. Portanto, destaca-se a importância crucial de identificar com precisão a técnica a ser utilizada, visando maximizar a eficiência, qualidade e confiabilidade dos resultados obtidos.

Outro fator que impacta na construção de modelos de ML mais eficientes e precisos é a otimização da representação dos dados. No âmbito dessa busca, técnicas de FE desempenham uma função crucial ao modificar e criar variáveis que se mostram relevantes para o problema em análise. Essa abordagem visa não apenas extrair informações significativas dos dados preexistentes, mas também gerar novas variáveis em um espaço remapeado com o intuito de aprimorar a capacidade discriminativa dos modelos.

Embora a literatura ofereça uma ampla gama de abordagens para a redução da complexidade de modelos de ML, a falta de consenso sobre a técnica mais eficaz em

diferentes contextos torna tanto a identificação das variáveis mais informativas quanto o aprimoramento dos dados para a construção do modelo uma questão complexa e aberta a novas abordagens (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015; MUÑOZ-ROMERO et al., 2020). Teoricamente, a ausência de uma abordagem consensual indica a necessidade de explorar novos métodos que possam preencher essa lacuna e contribuir para o avanço do conhecimento na área. Por outro lado, do ponto de vista prático, essa falta de consenso sugere que há espaço para a investigação de novas técnicas que possam ser mais eficazes em contextos específicos ou em face de desafios emergentes.

Assim, a presente tese está focada no desenvolvimento de novas abordagens para redução da complexidade de modelos de ML, permeando os contextos de seleção de técnica analítica, seleção de variáveis e transformação de variáveis. De tal forma, não apenas responde à necessidade teórica de avanço científico, mas também oferece soluções práticas para lidar com a complexidade e diversidade dos conjuntos de dados na prática da análise de dados.

1.3 Delineamento do estudo

Com os objetivos e justificativa desta tese definidos, esta seção apresenta o enquadramento da pesquisa do ponto de vista metodológico, descrevendo o método aplicado para alcançar os objetivos propostos, assim como um resumo das ferramentas utilizadas e contribuições científicas de cada artigo que compõe a tese.

1.3.1 Método de Pesquisa

Sob a perspectiva de sua natureza, esta tese é categorizada como pesquisa aplicada, uma vez que o conteúdo teórico é explorado e aplicado para abordar problemas genéricos (GIL, 2008). No âmbito metodológico, a tese se configura como pesquisa quantitativa, utilizando análises estatísticas e modelagem matemática para solução dos problemas apresentados (BERTO; NAKANO, 1999). Em relação aos objetivos, esta pesquisa é classificada como exploratória, pois busca compreender e obter uma visão geral do problema, abrindo caminho para a formulação de hipóteses com vistas à sua resolução (GIL, 2008).

1.3.2 Método de Trabalho

A tese é composta por três etapas, cada uma delas corresponde a um artigo com o intuito de atender os objetivos da tese. O primeiro artigo aborda a integração de conjuntos de dados provenientes de quatro técnicas analíticas diferentes (XRF, ESI-MS/MS, FTIR e UPLC-MS),

que descrevem as mesmas amostras de Viagra® e Cialis®. Além disso, o perfil físico das amostras de Viagra® também foi avaliado. Esses dados alimentam a ferramenta de classificação avançado, o *XGBoost* (CHEN; GUESTRIN, 2016; OMAR, 2018), com o objetivo principal de identificar as técnicas e variáveis mais relevantes para a detecção da autenticidade desses medicamentos. Além da análise quantitativa, o estudo também inclui uma análise descritiva para fornecer uma compreensão aprofundada dos resultados obtidos. Essa abordagem visa não apenas classificar as amostras, mas também destacar as técnicas analíticas mais informativas para futuras análises.

No segundo artigo, apresenta-se um método inovador para aprimorar a qualidade da classificação de amostras de diversos domínios descritas por dados FTIR/NIR. Esse método opera em duas fases, as quais integram a pré-seleção de comprimentos de onda (COs) orientada por agrupamento à estratégia baseada em *wrapper*. Na primeira fase, realiza-se a pré-seleção de COs agrupando-os através da técnica *Spectral Clustering* (SC). Os COs são agrupados em clusters, e aqueles que superam a acurácia de um modelo contendo todos os COs são pré-selecionados. Na segunda fase, refina-se a seleção dos COs usando o algoritmo *Sequential Forward Floating Selection* (SFFS), testando diferentes combinações de índices de importância (como distância de Bhattacharyya, Qui-quadrado, ReliefF e Gini) e técnicas de classificação (como Support Vector Machine, k-Nearest Neighbor e Random Forest). Seleciona-se então o subconjunto de COs responsável pela acurácia máxima. Essa abordagem permite, dentre outros, avaliar como as diferenças entre índices e classificadores influenciam as estruturas nos dados.

Por fim, o terceiro artigo propõe um método que combina abordagens simplificadas de seleção (FS) e engenharia de variáveis (FE) para obter desempenhos significativos sem a necessidade de otimização intensiva de hiperparâmetros ou técnicas de balanceamento de classes. Inicialmente, todas as variáveis são avaliadas quanto à sua importância utilizando o índice Gini. As variáveis com importância zero são removidas, e as restantes são divididas em quartis. Em seguida, são selecionadas aquelas que integram o quartil com as maiores importâncias (Q4). A segunda etapa do método remapeia as variáveis remanescentes por meio de ponderadores baseados no coeficiente de variação (CoefVar) em relação a cada classe resposta. O impacto dos processos acima foi avaliado em diferentes técnicas de classificação (como *Support Vector Machine*, *k-Nearest Neighbor*, *Decision Tree CART, Naiive Bayes* e Regressão Logística). A proposta visa capacitar os algoritmos, em suas configurações mais

básicas e padrão, a alcançarem altos níveis de desempenho em bases de dados de diversas naturezas e cenários.

Na Tabela 1.1 são apresentados os três artigos que compõem a tese, as ferramentas utilizadas e as contribuições científicas de cada artigo.

Tabela 1.1. Descrição dos artigos do projeto de tese.

Artigo	Título	Ferramentas	Contribuição
1	Fusão de dados para seleção eficiente de técnicas analíticas e variáveis na identificação de medicamentos	Fusão de dados de baixo nível (LLDF), *XGBoost*	Proposição de uma estratégia que integra fusão de dados com seleção de variáveis para identificar as técnicas analíticas e variáveis mais eficazes na autenticação de Cialis® e Viagra® Fornecimento de análise descritiva detalhada sobre os achados permitindo uma compreensão mais profunda dos padrões encontrados, auxiliando profissionais forenses a tomar decisões mais informadas
2	Uma nova abordagem composta por duas fases de seleção de comprimento de onda no infravermelho próximo e no infravermelho médio para classificação de amostras em categorias relativas à qualidade ou autenticidade	*Spectral Clustering* (SC), *k-Nearest Neighbor* (kNN), *Support Vector Machine* (SVM), *Random Forest* (RF), *Distância de Bhattacharyya* (DB), *Qui-quadrado* (χ^2), *ReliefF* (ReF), *Gini* (GI)	Proposição de um novo método duas fases de seleção de comprimentos de onda para categorização de amostras em categorias relativas à qualidade ou autenticidade
3	Seleção de variáveis integrada à engenharia de variáveis para aprimoramento da performance de modelos de aprendizado de máquina	Regressão Logística (RL), *k-Nearest Neighbor* (kNN), *Naïve Bayes* (NB), *Support Vector Machine* (SVM), *Decision Tree* (DT), *Gini* (GI)	Proposição de um novo método que combina estratégias simplificadas de FS e FE para aprimorar o desempenho de classificadores sem a necessidade de otimização intensiva de hiperparâmetros ou técnicas de balanceamento de classes.

(a) Artigo a ser submetido ao periódico Forensic Science International
(b) Artigo publicado no periódico Journal of Chemometrics: https://doi.org/10.1002/cem.3536
(c) Artigo a ser submetido ao periódico Chemometrics and Intelligent Laboratory Systems

1.4 Delimitações do Estudo

A presente pesquisa concentra-se no desenvolvimento e aplicação de métodos de seleção (FS) e engenharia de variáveis (FE), utilizando ferramentas e conceitos existentes na

literatura. O trabalho não propõe novas técnicas de classificação ou regressão, restringindo-se a combinar tais técnicas de forma a gerar novas abordagens para FS. O enfoque da abordagem de fusão de dados nesta tese é do tipo LLDF (fusão de dados de baixo nível), excluindo abordagens de fusão de dados de médio (MLDF) e alto nível (HLDF). Para a classificação das amostras, foram construídos modelos kNN usando apenas a distância euclidiana e SVM com kernel linear. Quanto ao agrupamento de variáveis, foram aplicadas apenas técnicas de clusterização não hierárquica baseadas na distância euclidiana, sem considerar técnicas de clusterização hierárquica ou baseadas em densidade.

Em relação à abrangência, a pesquisa concentra-se nas áreas alimentícia, petroquímica, farmacêutica e forense. Quanto aos bancos de dados, foram analisados apenas bancos supervisionados, com fins de classificação. Com isso as avaliações dos modelos propostos se restringem principalmente a métricas de avaliação de qualidade já conhecidas, como acurácia, sensibilidade, especificidade, área sob a curva precision-recall (AUC-PR). Não foram avaliados aspectos de redução de custos decorrentes da simplificação do processo de coleta de dados.

1.5 Referências

ABDULWAHAB, H. M.; AJITHA, S.; SAIF, M. A. N. Feature selection techniques in the context of big data: taxonomy and analysis. **Applied Intelligence 2022 52:12**, v. 52, n. 12, p. 13568–13613, 27 jan. 2022.

ANZANELLO, M. J. et al. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209–214, set. 2013.

BERTO, R. M. V. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Production**, v. 9, n. 2, p. 65–75, dez. 1999.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. **Feature Selection for High-Dimensional Data.** Springer ed. [s.l: s.n.].

BORRÀS, E. et al. Prediction of olive oil sensory descriptors using instrumental data fusion and partial least squares (PLS) regression. **Talanta**, v. 155, p. 116–123, 1 ago. 2016.

BZDOK, D.; KRZYWINSKI, M.; ALTMAN, N. Machine learning: supervised methods. **Nature methods**, v. 15, n. 1, p. 5, 3 jan. 2018.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, v. 40, n. 1, p. 16–28, 1 jan. 2014.

CHEN, T.; GUESTRIN, C. **XGBoost: A scalable tree boosting system**. Proceedings of the

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: Association for Computing Machinery, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 14 jan. 2021

COCCHI, M.; BIANCOLILLO, A.; MARINI, F. Chemometric Methods for Classification and Feature Selection. In: JAUMOT, J.; BEDIA, C.; TAULER, R. (Eds.). **Data Analysis for Omic Sciences: Methods and Applications**. Comprehensive Analytical Chemistry. [s.l.] Elsevier, 2018. v. 82p. 265–299.

CORALLO, A. et al. Evaluating maturity level of big data management and analytics in industrial companies. **Technological Forecasting and Social Change**, v. 196, p. 122826, 1 nov. 2023.

DOS SANTOS, F. R. et al. Data fusion of XRF and vis-NIR using p-ComDim to predict some fertility attributes in tropical soils derived from basalt. **Microchemical Journal**, v. 191, p. 108813, 1 ago. 2023.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. LTC ed. Rio de Janeiro: [s.n.].

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6ª ed. São Paulo: Atlas S. A., 2008.

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

KHURANA, U. et al. Cognito: Automated Feature Engineering for Supervised Learning. **IEEE International Conference on Data Mining Workshops, ICDMW**, v. 0, p. 1304–1307, 2 jul. 2016.

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

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.

MUÑOZ-ROMERO, S. et al. Informative variable identifier: Expanding interpretability in feature selection. **Pattern Recognition**, v. 98, p. 107077, fev. 2020.

OMAR, K. B. A. **XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison Semester Project**. **Semester Project**, 2018.

OUSSOUS, A. et al. Big Data technologies: A survey. **Journal of King Saud University - Computer and Information Sciences**, v. 30, n. 4, p. 431–448, 1 out. 2018.

PIZARRO, C. et al. Classification of Spanish extra virgin olive oils by data fusion of visible spectroscopic fingerprints and chemical descriptors. **Food Chemistry**, v. 138, n. 2–3, p. 915–922, 1 jun. 2013.

RAWAT, T.; KHEMCHANDANI, V. Feature Engineering (FE) Tools and Techniques for Better Classification Performance. 2019.

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

SALIMI, 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.

URBANOWICZ, R. J. et al. 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.** Brasília: Universidade de Brasília (UnB), 2017.

XIAOBO, Z. et al. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, v. 667, n. 1–2, p. 14–32, 2010.

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

YUN, Y. H. et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra. **TrAC - Trends in Analytical Chemistry**, v. 113, p. 102–115, 2019.

ZHONG, S. et al. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. **Environmental Science and Technology**, v. 55, n. 19, p. 12741–12754, 5 out. 2021.

5 CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as conclusões da pesquisa realizada neste projeto de tese, além de sugestões para a próxima etapa da tese que está em desenvolvimento.

5.1 Conclusões

A presente tese tem como objetivo principal a proposição de novas abordagens para a redução da complexidade de modelos de ML, permeando os contextos de seleção de técnica analítica, seleção de variáveis e transformação de variáveis, com fins de identificação de qualidade de produtos em dados de alta dimensionalidade oriundos de diferentes segmentos. Neste trabalho foram apresentados três artigos finalizados de modo a alcançar os objetivos específicos propostos. São eles: (*i*) combinar fusão de dados com um algoritmo de ML de alto desempenho (*XGBoost*) para maximizar a eficiência, qualidade e confiabilidade dos resultados desejados; *(ii)* propor e validar uma sistemática de seleção de variáveis apoiada em clusterização de variáveis; *(iii)* propor e validar uma sistemática que combina estratégias de FS e de FE vistas ao aumento da eficiência das técnicas de classificação; *(iv)* avaliar o uso combinado de diferentes ferramentas de classificação com diferentes índices de importância de variáveis; *(v)* avaliar a robustez dos métodos propostos comparando-os a outros métodos encontrados na literatura.

Os objetivos *(i)* e *(v)* foram alcançados no primeiro artigo, que empregou uma abordagem de fusão de dados combinada com o algoritmo *XGBoost*. Essa estratégia foi utilizada para selecionar as técnicas analíticas mais promissoras na categorização de amostras de Cialis® e Viagra® como autênticas ou falsificadas. Além disso, o estudo realizou uma análise descritiva das relações e interações das variáveis identificadas como mais relevantes, comparando os resultados com outros estudos encontrados na literatura. Como resultado, a estrutura proposta rendeu uma acurácia média de classificação superior a 95%, mantendo menos de 3% das variáveis em cada replicação para ambos os medicamentos. Para a categorização das amostras de Viagra®, os resultados indicaram uma falta de padrão definido nas variáveis retidas da técnica ESI-MS/MS, o que levanta questionamentos sobre a necessidade de obter esse tipo de dado para este propósito em comparação com outras técnicas mais eficazes. Quanto ao Cialis®, as variáveis do ATR–FTIR, UPLC–MS/MS e ESI–MS/MS não contribuíram de maneira significativa para caracterizar as amostras quanto à autenticidade, sugerindo a viabilidade de técnicas alternativas mais eficazes para este fim e

possibilitando uma redução na coleta de dados. A abordagem proposta obteve o melhor resultado quando comparado com outro método encontrado na literatura.

Os objetivos (*ii*), (*iv*) e (*v*) foram alcançados no segundo artigo, o qual propôs uma estrutura de duas fases que integra uma etapa de pré-seleção de CO orientada por agrupamento de CO a uma estratégia baseada em *wrapper*. A primeira fase realiza um processo de poda nos dados que remove os COs menos informativos contando com o *Spectral Clustering* (SC). Na segunda etapa é realizado o processo iterativo de inserção ordenada dos COs pré-selecionados usando o algoritmo *Sequential Forward Floating Selection* (SFFS) que testa a combinação de diferentes índices de importância de CO (ou seja, distância de Bhattacharyya, χ^2, ReF e GI) e técnicas de classificação (ou seja, SVM, kNN e RF). O método foi aplicado a onze bancos de dados de diferentes contextos (indústria alimentícia, petroquímica, farmacêutica e do âmbito forense) e a combinação recomendada foi SC-RF-GI. Também demonstrou-se os benefícios do emprego da estratégia de duas fases com outras abordagens que dependem de procedimentos mais complexos em termos de desempenho de classificação e tempo de processamento computacional.

Por fim, os objetivos *(iii)* e *(v)* foram atingidos no terceiro artigo, dedicado ao desenvolvimento e implementação de um novo método (FS_{GI}+FE) que combina estratégias de: seleção de variáveis (*feature selection* – FS) e engenharia de variáveis (*feature engineering* – FE). A etapa de FS é baseada na abordagem *filter*, utilizando o índice de importância de variáveis gini e aplicando o conceito de quartis para selecionar as variáveis com as maiores importâncias. Já a etapa de FE é fundamentada no coeficiente de variação (CoefVar) de cada variável em relação a cada classe resposta. O método proposto é testado em oito bancos de dados, abrangendo tanto naturezas binárias quanto multiclasse, e contemplando diferentes contextos de aplicação, incluindo indústria alimentícia, farmacêutica e âmbito forense. O impacto da abordagem foi avaliado em diferentes técnicas de classificação (SVM, k-NN, DT, NB e LR). Os resultados destacam a eficácia do método proposto (FS_{GI}+FE) em diversos cenários, como evidenciado pelas diferentes bases de dados analisadas neste estudo. A etapa de FS do método proposto contribui na melhoria do desempenho do modelo ao priorizar as variáveis que melhor distinguem entre as diferentes classes dos dados. Além disso, os resultados obtidos evidenciam a eficácia da etapa de FE do

método proposto quando aplicada de forma isolada, mesmo em cenários de dados altamente desbalanceados.

Com base no que foi exposto acima, conclui-se que esta tese cumpriu todos os objetivos específicos propostos e contribuiu para o avanço dos estudos na área de seleção de variáveis com fins de classificação e predição de propriedades das amostras.

5.2 Sugestões para trabalhos futuros

Como possíveis extensões da pesquisa apresentada nesta tese, sugerem-se as seguintes ações para pesquisas futuras:

(i) Empregar estratégias de médio e alto nível para fusão de dados;
(ii) Melhorar a fase WLPS proposta no Artigo 2, incorporando técnicas de otimização ao processo de clusterização, com o intuito de produzir clusters de maior qualidade;
(iii) Refinar a etapa de FS do método proposto pelo Artigo 3, transformando-a em um processo de duas fases. Primeiramente, realizar uma pré-seleção de variáveis utilizando a abordagem *filter* proposta e testar diferentes quartis como limiar de corte para identificar um subconjunto inicial promissor e eliminar variáveis ruidosas que podem prejudicar as predições. Em seguida, refinar o processo de seleção, considerando a correlação e as redundâncias entre as variáveis remanescentes.